Learning Topic Models: Identifiability and Finite-Sample Analysis

Yinyin Chen*[1], Shishuang He*[2], Yun Yang[2], and Feng Liang[2]

[1]Meta Platforms, Inc.

[2]Department of Statistics, University of Illinois at Urbana-Champaign

Abstract

Topic models provide a useful text-mining tool for learning, extracting, and discovering latent structures in large text corpora. Although a plethora of methods have been proposed for topic modeling, lacking in the literature is a formal theoretical investigation of the statistical identifiability and accuracy of latent topic estimation. In this paper, we propose a maximum likelihood estimator (MLE) of latent topics based on a specific integrated likelihood that is naturally connected to the concept, in computational geometry, of *volume minimization*. Our theory introduces a new set of geometric conditions for topic model identifiability, conditions that are weaker than conventional separability conditions, which typically rely on the existence of pure topic documents or of anchor words. Weaker conditions allow a wider and thus potentially more fruitful investigation. We conduct finite-sample error analysis for the proposed estimator and discuss connections between our results and those of previous investigations. We conclude with empirical studies employing both simulated and real datasets.

Keywords: Topic models, Identifiability, Sufficiently scattered, Volume minimization, Maximum likelihood, Finite-sample analysis.

1 Introduction

Topic models, such as Latent Dirichlet Allocation (Blei et al., 2003) models and probabilistic Latent Semantic Analysis (Hofmann, 1999), have been widely used in natural language processing, text mining,

*Yinyin Chen and Shishuang He contributed equally to this work.

information retrieval, etc. The purpose of those models is to learn a lower-dimensional representation of the data, in which each document can be expressed as a convex combination of a set of latent topics.

Consider a corpus of d documents with vocabulary size V. A topic model with k latent topics can be summarized as the following matrix factorization:

$$\mathbf{U}_{V\times d} = \mathbf{C}_{V\times k}\mathbf{W}_{k\times d}, \tag{1}$$

where all matrices are column-stochastic[1]. In particular, $\mathbf{U}_{V\times d}$ is the true term-document matrix whose columns are the true underlying word frequencies for the d documents; $\mathbf{C}_{V\times k}$ is the *topic matrix* whose columns are the multinomial parameters (i.e., word frequencies) for the k topics; and $\mathbf{W}_{k\times d}$ is the *mixing matrix* whose columns present the mixing weights over k topics for d documents.

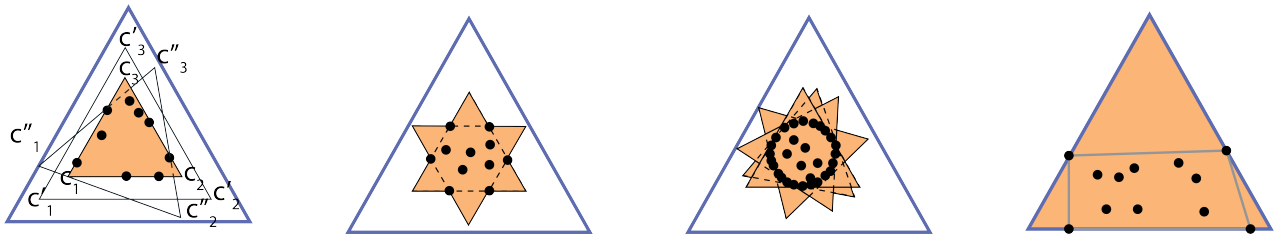
The primary interest here is to reveal the latent structure of a collection of documents, i.e., to estimate the collection's topic matrix $\mathbf{C}$. Despite the popularity and success of topic models, work on the estimation accuracy of $\mathbf{C}$ is scarce. An obstacle to rigorous analysis of that important question is that the factorization (1) may not be unique up to permutation (throughout we ignore any non-uniqueness due to permutations of the k topics). The non-uniqueness issue can be easily understood via the following geometric interpretation of Equation (1): recovering $\mathbf{C}$ based on $\mathbf{U}$ is equivalent to finding a k-vertex convex polytope that encloses all columns of $\mathbf{U}$; the vertices of this k-vertex convex polytope form the columns of $\mathbf{C}$. Apparently, such a convex polytope may not be unique; see Figure 1(a). In statistical language, topic models parameterized by $(\mathbf{C}, \mathbf{W})$ without any further constraints are not identifiable (modulo column permutations).

This leads to the following two questions that we aim to address in this paper.

1. *Identifiability.* Under what conditions is a topic model parameterized by $(\mathbf{C}, \mathbf{W})$ identifiable up to permutation? It is easy to achieve identifiability by imposing stringent conditions that significantly limit the usefulness of the result. Our goal is to develop a set of identifiability conditions that are weaker than ones proposed in prior studies but whose accuracy may nevertheless be well estimated.

2. *Finite-sample error.* For an identifiable topic model, can we provide an estimator of $\mathbf{C}$ whose finite-sample error leads to the desired rate of convergence? The rate will depend on the number of documents d and/or the number of words per document n (which, without loss of generality, is assumed to be the same for all documents). Throughout, we assume the vocabulary size V and the number of topics k to be known and fixed.

[1] We say a matrix is column-stochastic if its entries are non-negative and columns sum to one.



(a) (b) (c) (d)

Figure 1: Geometric view of the simplex $\Delta^{V-1}(k = V = 3)$. Black dots are columns of $\mathbf{U}$. Black-lined triangles are k-vertex convex polygons; the shaded triangles are those with minimum volume.

1.1 Related Work

Topic models have been studied under two settings: one in which the mixing weights, columns of $\mathbf{W}$, are assumed to be stochastically generated from some distribution; the other in which they are assumed to be fixed but unknown. The Bayesian approach, for example, focuses on the former.

1.1.1 The Bayesian Approach

In the Bayesian setting, the mixing weights are often assumed to be stochastically generated from a *known* distribution with a full support on the simplex Δ^{k-1}. Therefore, identifiability can be guaranteed under very mild conditions; for example, one such condition is just that $\mathbf{C}$ be of full rank (Anandkumar et al., 2012). Under such Bayesian settings, Nguyen (2015) and Tang et al. (2014) established posterior concentration rates; Anandkumar et al. (2012, 2014) and Wang (2019) established convergence rates for the maximum likelihood estimator (MLE).

In this paper, we focus on a more general setting, in which the mixing weights may not be stochastically generated; if they are, moreover, we do not assume any knowledge of the corresponding distribution. Identifiability and estimation accuracy turn out to be much more challenging under this general setting.

1.1.2 The Separability Condition

Several earlier investigations have addressed identifiability by imposing the so-called *separability condition* or its generalization (Donoho and Stodden, 2004; Arora et al., 2012; Azar et al., 2001; Kleinberg and Sandler, 2008, 2003; Recht et al., 2012; Ge and Zou, 2015; Ke and Wang, 2017; Papadimitriou et al., 2000; McSherry, 2001; Anandkumar et al., 2012). The separability condition can be imposed either on rows of $\mathbf{C}$ or on columns of $\mathbf{W}$, due to the symmetry between these two matrices

in the factorization (1).

When imposed on the topic matrix $\mathbf{C}_{V\times k}$, this condition assumes that, after the rows of $\mathbf{C}$ have been re-arranged, its top k rows will form a diagonal matrix. Words associated with those rows are called *anchor words*; anchor words can be used to identify topics since they appear only in one particular topic.

When imposed on the mixing matrix $\mathbf{W}_{k\times d}$, this condition again assumes that, after the columns of $\mathbf{W}$ have been re-arranged, the first k columns will form a diagonal matrix. We can further conclude that that diagonal matrix must be an identity matrix since $\mathbf{W}$ is column-stochastic; therefore, there are k documents that belong to one and only one topic (Nascimento and Dias, 2005; Javadi and Montanari, 2020). A geometric interpretation of this condition is that we can use the convex hull of k columns of $\mathbf{U}$ to form the k-vertex polytope that contains all other columns of $\mathbf{U}$. In other words, the topic matrix $\mathbf{C}_{V\times k}$ can be recovered by identifying the corresponding subset of k documents.

The separability condition can be easily violated, however, in real applications. In practice it is commonly the case that topics are correlated, tend to share keywords, and therefore are not separable.

Nevertheless, several algorithms have been proposed to estimate $\mathbf{C}$ with a convergence rate of the order $1/\sqrt{nd}$ (Arora et al., 2012; Ke and Wang, 2017), but they assume separability. This rate of convergence would indicate that such algorithms can pool information in the d documents, each with n words, to estimate $\mathbf{C}$; therefore they have an effective sample size of nd, instead of n or d. However, as discussed in Section 4.3, such a fast convergence rate is achievable only under the stringent separability assumption. This is because the strong separability condition greatly simplifies the statistical and computational hardness of the topic matrix estimation problem and turns it into a searching problem. As a consequence, such separability-condition-based methods circumvent the hidden non-regular statistical problem of boundary estimation (c.f. Section 4.3), which often leads to an extremely slow rate of convergence. See Section 3.2 for a review of separability-condition-based methods and how they relate to ours, from a two-stage estimation perspective.

1.1.3 Beyond the Separability Condition

To relax the separability assumption, the aforementioned connection between estimating a topic model and finding a k-vertex convex polytope that encloses all columns of $\mathbf{U}$ has led researchers to start looking at geometric conditions.

When there are multiple k-vertex convex polytopes enclosing columns of $\mathbf{U}$, it is natural to restrict our attention to the ones with minimum volume, that is, convex polytopes that circumscribe the data

as compactly as possible. Many *volume minimization* algorithms have been proposed (Craig, 1994; Nascimento and Dias, 2005; Miao and Qi, 2007; Fu et al., 2015) for nonnegative matrix factorization similar to (1). However, most of these methods consider the noiseless setting. Blindly applying them to topic model estimation fails to respect the error structure in the counting data and may lead to a loss of statistical efficiency. Moreover, little theoretical work has been conducted on model identifiability and estimation accuracy beyond the limited context of topic modeling that assumes the separability condition. In particular, it is important to acknowledge that the minimum volume constraint alone does not guarantee uniqueness; see examples in Figure 1(b)(c).

Recently, a set of geometric conditions known as the *sufficiently scattered* (SS) condition, which is weaker than the separability condition, has been introduced to study identifiability of topic models (Huang et al., 2016; Jang and Hero, 2019). Huang et al. (2016) ensure identifiability under the SS condition by adding the constraint that the determinant of $\mathbf{W}\mathbf{W}^T$ is minimized. Jang and Hero (2019) have proved that the SS condition, along with volume minimization on the convex hull of $\mathbf{C}$, ensures identifiability when $V = k$ (vocabulary size is the same as topic size); their analysis is valid only for $V = k$ since it is built on the assumption that the volume of the convex hull of $\mathbf{C}$ is equal to the determinant of $\mathbf{C}$ (or to a monotonic function of the determinant of $\mathbf{C}^T\mathbf{C}$) which holds true only when $V = k$. In addition, neither Huang et al. (2016) nor Jang and Hero (2019) provided a theoretical analysis of estimation errors for their proposed estimators, which are based on minimizing a squared loss based objective rather than on maximizing the multinomial likelihood associated with counting data.

Javadi and Montanari (2020) is the only study we are aware of that provides a theoretical analysis of estimation errors without assuming the separability condition. They proposed to estimate the k columns of $\mathbf{C}$ by minimizing their distance to the convex hull of the data points, and established a convergence rate for their estimator. In their setting, model identifiability is equivalent to the uniqueness of the minimizer in the noiseless setting; that is, they assume that a unique set of k columns (of $\mathbf{C}$) is closest to the convex hull formed by the columns of $\mathbf{U}$. They show that the minimizer is indeed unique when the separability condition is imposed on $\mathbf{W}$; other than that, they do not provide any checkable conditions for identifiability.

1.2 Summary of Our Contribution

First, we resolve the non-identifiability issue by focusing on convex hulls (of $\mathbf{C}$) of the smallest volume, and show that under volume minimization, the SS condition ensures identifiability regardless of the

values of V and k (Section 2).

Although volume minimization helps to ensure model identifiability, since the volume of a low-dimensional simplex in a high-dimensional space does not take a simple form (Miao and Qi, 2007), it is difficult to incorporate volume minimization into an estimation procedure. This difficulty explains why many prior investigations have either assumed $V = k$ or used an approximation formula.

Our second contribution is to establish the connection between volume minimization and maximization of a particular integrated likelihood (Section 3.1). Specifically, we propose an estimator as the MLE of the topic matrix $\mathbf{C}$, based on an integrated likelihood, in which the mixing weights (i.e., columns of $\mathbf{W}$) are profiled out by integrating with respect to a uniform distribution over $(k-1)$-simplex. A geometric consequence of the use of uniform distribution is that, while maximizing the integrated likelihood, we implicitly minimize the volume of the convex hull of $\mathbf{C}$ without explicitly evaluating its volume. Here we emphasize that the uniform distribution is used only to integrate over nuisance parameters (i.e., the mixing weights), and that our theoretical analysis does not require the mixing weights to be generated stochastically from a uniform distribution.

Our third contribution is to establish a finite-sample error bound of the proposed estimator of $\mathbf{C}$, of the order $\sqrt{\log(n \vee d)/n}$ under the fixed design setting where the mixing weights $\mathbf{W}$ can be arbitrarily allocated—as long as the SS condition pertains (Section 4.2). As a consequence, our result implies asymptotic consistency as the number of documents d and/or the number of words n (in each document) increases to infinity. In the stochastic setting, where the mixing weights $\mathbf{W}$ are independently generated according to some unknown underlying distribution over the simplex, we show that, for sufficiently large d, $\mathbf{W}$ still satisfies a perturbed version of the SS condition with high probability—as long as the support of the weight generating distribution satisfies the SS condition. Based on this observation, we also provide a finite-sample error bound in the stochastic (or random design) setting (Section B in the supplementary material). Furthermore, by drawing a connection between our estimating approach and some representative existing methods, through a two-stage perspective (Section 3.2), we illustrate that the separability condition greatly simplifies the topic matrix estimation problem by circumventing the highly nontrivial and non-regular statistical problem of boundary estimation (Section 4.3). This explains why our finite-sample error bound is similar to that of Javadi and Montanari (2020) which is based on an archetypal analysis that, like ours, does not assume the separability condition; however, our error bound is (not surprisingly) worse (in terms of the dependence on d) than those (Ke and Wang, 2017; Arora et al., 2012) arrived at under the separability condition.

As a byproduct, our work provides a theoretical justification for the empirical success of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) models, since the proposed estimator is essentially the maximum likelihood estimator of $\mathbf{C}$ from the LDA model, with a particular choice of prior on $\mathbf{W}$. More generally, the LDA model with other prior choices on $\mathbf{W}$ can be interpreted as maximizing the data likelihood while minimizing a weighted volume in which a non-uniform volume element is integrated over the convex hull of $\mathbf{C}$ when defining the volume (see Section 5.1.2 for some numerical comparisons).

Although presented in the context of topic modeling, our results can be adapted to many other applications by using the data-specific likelihood. For example, the decomposition $\mathbf{U} = \mathbf{CW}$ plays an important role in *hyperspectral imaging analysis*, in which each column of $\mathbf{U}$ represents the intensity levels over V channels at a pixel. Due to the low spatial resolution of hyperspectral images, pixel spectra are usually mixtures of spectra from several pure materials, known as endmembers. So a key step in hyperspectral imaging analysis is to separate (or unmix) the pixel spectra into convex combinations of endmember spectra; endmember spectra are essentially columns of $\mathbf{C}$ (Winter, 1999). Similar models also arise in *reinforcement learning* (Singh et al., 1995; Duan et al., 2019) as a way to compress the transition matrix of an underlying Markov decision process; a detailed discussion is given in Section 5.2.2.

1.3 Notation and Organization

Let $\mathbf{1}_k$ denote the all-ones vector of length k, and $\mathbf{e}_f$ the f-th column of the $k \times k$ identity matrix $\mathbf{I}_k$. Let $\Delta^{k-1} = \{\mathbf{x} \in \mathbb{R}^k : 0 \leqslant x_i \leqslant 1, \sum_{i=1}^k x_i = 1\}$ denote the $(k-1)$-dimensional probability simplex. For a matrix $\mathbf{A}_{p\times q} = (\mathbf{A}_1, \cdots, \mathbf{A}_q)$, let

$$\begin{aligned} \mathrm{Conv}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda} \in \Delta^{q-1}\}, \\ \textit{cone}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda} \geqslant 0\}, \\ \text{and} \quad \mathrm{aff}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda}^T\mathbf{1}_q = 1, \boldsymbol{\lambda} \in \mathbb{R}^q\}, \end{aligned}$$

denote the *convex polytope*, *simplicial cone* and *affine space* generated by (the q columns of) $\mathbf{A}$, respectively. For $\mathbf{A} \in \mathbb{R}^{p\times q} (p \geqslant q)$, we define $|\mathrm{Conv}(\mathbf{A})|$ as the $(q-1)$-dimensional volume of $\mathrm{Conv}(\mathbf{A})$ on $\mathrm{aff}(\mathbf{A})$, which can be computed by the Cayley–Menger determinant or Lemma D.1 in Appendix D. For any vector $\mathbf{x}$, $\mathbf{x} \geqslant a$ means $\mathbf{x}$ is element-wisely greater than or equal to a. Denote $a \vee b$ and $a \wedge b$ as the larger and smaller number between a and b, respectively. For any cone $\mathcal{C}$, let $\mathcal{C}^* = \{\mathbf{x} : \mathbf{x}^T\mathbf{y} \geqslant 0, \forall \mathbf{y} \in \mathcal{C}\}$ denote its *dual cone*. Recall some useful facts of dual cones (Donoho and

Stodden, 2004): (i) $cone(\mathbf{A})^* = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^T\mathbf{A} \geqslant 0\}$; (ii) if $\mathcal{A}$ and $\bar{\mathcal{A}}$ are convex cones, and $\mathcal{A} \subseteq \bar{\mathcal{A}}$, then $\bar{\mathcal{A}}^* \subseteq \mathcal{A}^*$. Unless stated otherwise, all the constants in the paper are independent of number of words per document n and number of documents d.

The rest of the paper is organized as follows. In Section 2, we discuss identifiability under volume minimization as well as a set of sufficient conditions. In Section 3, we propose the MLE based on an integrated likelihood, establish its connection with volume minimization, and describe its computation. Theoretical analysis of the proposed estimator is presented in Section 4. Finally, empirical evidence is reported in Section 5. Proofs and technical results are included in the supplementary material.

2 Identifiability of Topic Models

In this section we start with a formal definition of topic model identifiability under the minimum volume constraint. After that, we describe two sufficient conditions that lead to the identifiability, namely the separability condition and the sufficiently scattered condition. Finally, for the latter condition, which is weaker and less stringent than conventional separability, we provide a geometric interpretation.

2.1 Identifiability under Volume Minimization

We have observed (see Figure 1(a)) that without any constraint, a topic model is almost always non-identifiable. We thus focus on identifiability under the minimum volume volume minimization constraint, due to its natural interpretation as finding the most parsimonious topic model that explains the documents in the corpus data, or equivalently, the most compact k-vertex convex polytope in which the documents reside.

We begin by defining the following distance metric between two topic matrices $\mathbf{C}$ and $\bar{\mathbf{C}}$:

$$\mathcal{D}(\mathbf{C}, \bar{\mathbf{C}}) = \min_{\mathbf{\Pi}} \|\bar{\mathbf{C}} - \mathbf{C}\mathbf{\Pi}\|_2, \tag{2}$$

where $\|\cdot\|_2$ denotes the spectral norm and $\mathbf{\Pi}$ is a permutation matrix. Note that $\mathcal{D}(\mathbf{C}, \bar{\mathbf{C}}) = 0$ if and only if $\bar{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}$, that is, $\mathbf{C}$ and $\bar{\mathbf{C}}$ are identical up to a permutation of columns. Since k and V are fixed, the spectral norm in (2) is not important because all matrix norms are equivalent. In particular, if the Frobenius norm is employed instead of the spectral norm, then the distance metric $\mathcal{D}$ coincides with the 2-Wasserstein distance between column vectors of $\mathbf{C}$ and $\bar{\mathbf{C}}$.

Next, we state the definition of identifiability under the minimum volume constraint:

Definition 1 (Identifiability)**.** *A topic model associated with parameters* $(\mathbf{C}, \mathbf{W})$ *is identifiable, if for any other set of parameters* $(\bar{\mathbf{C}}, \bar{\mathbf{W}})$*, the following conditions hold,*

$$\mathbf{CW} = \bar{\mathbf{C}}\bar{\mathbf{W}} \ \textit{ and } \ |\operatorname{Conv}(\bar{\mathbf{C}})| \leqslant |\operatorname{Conv}(\mathbf{C})|, \tag{3}$$

if and only if $\mathcal{D}(\mathbf{C}, \bar{\mathbf{C}}) = 0$*.*

It is easy to check that model identifiability is achieved under the separability condition on columns of $\mathbf{W}$, as it implies that $\mathbf{W}$ contains a $k \times k$ identity matrix after a proper column permutation; that is, there exist k columns in $\mathbf{U}$ that are the k corners of $\operatorname{Conv}(\mathbf{C})$. Therefore, no other k-vertex convex polytope of smaller or equal volume can still enclose all columns in $\mathbf{U}$.

Proposition 1. *If the separability condition is satisfied on* $\mathbf{W}$*, then* $(\mathbf{C}, \mathbf{W})$ *is identifiable.*

Since the separability condition can be overly stringent in practice, we next show that a condition weaker than the separability condition can also achieve model identifiability. Our analysis is related to the following geometric condition, known as *sufficiently scattered* (SS). Its definition relies on the second order cone $\mathcal{K}$, its boundary $bd\mathcal{K}$, and its dual cone $\mathcal{K}^*$, which are defined below:

$$\begin{aligned} \mathcal{K} &= \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T \mathbf{1}_k\}, \\ bd\mathcal{K} &= \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = \mathbf{x}^T \mathbf{1}_k\}, \\ \text{and} \quad \mathcal{K}^* &= \{\mathbf{x} \in \mathbb{R}^k : \mathbf{x}^T \mathbf{1}_k \geqslant \sqrt{k-1}\|\mathbf{x}\|_2\}. \end{aligned}$$

Definition 2 (SS Condition)**.** *A matrix* $\mathbf{W}$ *is* **sufficiently scattered***, if it satisfies:*

(S1). $cone(\mathbf{W})^* \subseteq \mathcal{K}$*, or equivalently,* $cone(\mathbf{W}) \supseteq \mathcal{K}^*$*;*

(S2). $cone(\mathbf{W})^* \bigcap bd\mathcal{K} \subseteq \{\lambda \mathbf{e}_f, f = 1, \cdots, k, \lambda \geqslant 0\}$*.*

It is easy to verify that the separability condition on $\mathbf{W}$ implies $\mathbf{W}$ to be sufficiently scattered. In fact, the separability condition on $\mathbf{W}$ means that $\operatorname{Conv}(\mathbf{W}) = \Delta^{k-1}$ fills up the entire simplex, and that $cone(\mathbf{W})^* = cone(\Delta^{k-1})$ is the most extreme cone (smallest possible cone, corresponding to the solid triangle in Figure 2; see the following section for details) that satisfies (S1) - (S2) in the SS condition.

Theorem 2. *If* $\mathbf{W}$ *is sufficiently scattered and* $\mathbf{C}$ *is of rank* k *(full column rank), then* $(\mathbf{C}, \mathbf{W})$ *is identifiable.*

Proof of Theorem 2 is given in the supplementary material (Section D.1). Here we give a sketch of the proof. Suppose $\mathbf{CW} = \bar{\mathbf{C}}\bar{\mathbf{W}}$. We have $\mathbf{C} = \bar{\mathbf{C}}\mathbf{B}$, where $\mathbf{B} = \bar{\mathbf{W}}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$. It suffices to show $\mathbf{B}$ is a permutation matrix, which we prove by verifying that any row of $\mathbf{B}$ is in $cone(\mathbf{W})^* \bigcap bd\mathcal{K} = \{\lambda \mathbf{e}_f, \lambda \geqslant 0\}$ and is also of unit length.

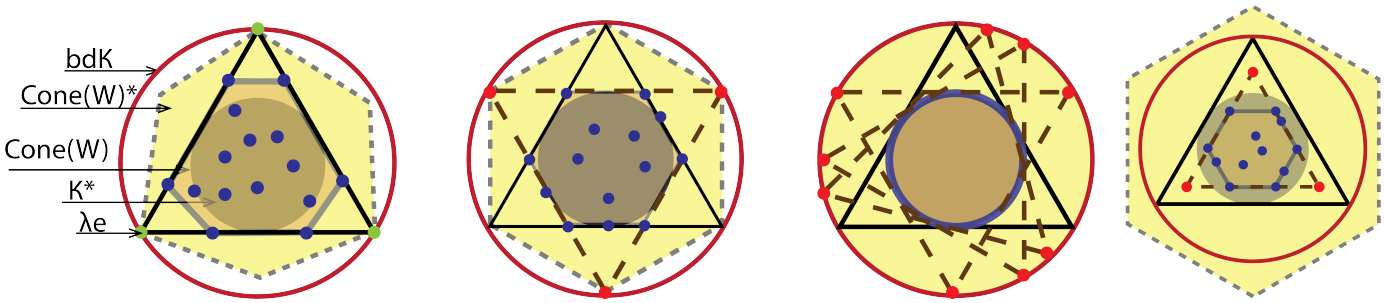
Remark 2.1 (Comparison with definition in Javadi and Montanari (2020))**.** The model identifiability defined in Javadi and Montanari (2020) is different from ours. They define a model to be identifiable if there is a unique convex polytope that minimizes the sum of distances from vertices of Conv($\mathbf{C}$) (i.e., columns of $\mathbf{C}$) to the convex hull of $\mathbf{U}$. Their notion of identifiability is easier than ours to be formulated into a statistical estimator that minimizes an empirical evaluation of the distance sum from data. In our approach, the volume of our low-dimensional polytope does not take a simple form, which greatly complicates the estimator construction. Fortunately, we find that maximizing a particular integrated likelihood leads to an estimator that implicitly minimizes the volume. (See Appendix A for further discussion of this topic.)

Remark 2.2 (SS condition is not a necessary condition)**.** The SS condition is not necessary for identifiability — one reason is that it does not take into account additional parameter constraints (e.g., in the topic model, each column of topic matrix $\mathbf{C}$ should be a probability weight vector belonging to the simplex). See Figure 1(d) for an example ($V = k = 3$ and $\mathbf{C} = \mathbf{I}_3$) where the SS condition does not hold but the model is identifiable. Since any alternative topic matrix $\mathbf{C}$ as a convex polytope with three vertices must be inside Δ^2, due to the parameter constraint, $\mathbf{I}_3$ is the only topic matrix enclosing all columns of $\mathbf{U}$ and is within simplex Δ^2. However, the SS condition does not hold since, apparently, $cone(\mathbf{W}) \supseteq \mathcal{K}^*$ is not true.

2.2 Geometrical Interpretation of Sufficiently Scattered Condition

We provide a geometric interpretation of the SS condition in Figure 2 with $k = 3$. Since the mixing weights are all on Δ^2, what is shown in Figure 2 is the intersection of the cones with the hyperplane $\mathbf{x}^T\mathbf{1}_3 = 1$. The mixing weights, $\mathbf{w}_1, \ldots, \mathbf{w}_d$, are represented as blue dots. Other items related to Definition 2 are: $bd\mathcal{K}$ is the red circle, $\mathcal{K}^*$ is the dark brown ball inscribed in the triangle, and $cone(\mathbf{W})^*$ is the yellow convex region with dashed boundary.

We illustrate three different scenarios: "SS" means that the SS condition is satisfied, "not SS" means that the SS condition is violated, and "sub-SS" means that (S1) is satisfied but (S2) is not.



(a) SS (b) sub-SS (c) sub-SS (d) not SS

Figure 2: Geometric views of the SS condition shown on the hyperplane $\mathbf{x}^T\mathbf{1}_k = 1$ ($k = 3$). Mixing weights $\mathbf{w}$ are represented as blue dots; blue dots in (c) are all on the boundary of the inner circle. Any dashed triangle in (b)(c)(d) is an alternative 3-vertex convex polytope that contains all $\mathbf{w}$'s and is of a volume no larger than Δ^{k-1}.

An equivalent form of Condition (S1) is $cone(\mathbf{W}) \supseteq \mathcal{K}^*$. So (S1) has a simple and intuitive interpretation: the mixing weights (blue dots) should form a convex polytope that contains the dual cone $\mathcal{K}^*$, the inner ball inscribed in the triangle. See Figure 2(d) for a violation of (S1). In particular, the separability condition on $\mathbf{W}$ implies that the three vertices (blue circles) of the triangle are included in $\mathbf{W}$. As a consequence, $cone(\mathbf{W})^* = cone(\mathbf{W})$ is the entire triangle, which is the most extreme/superfluous instance that satisfies the SS condition.

Condition (S1) ensures that $\text{Conv}(\mathbf{C})$ has the smallest possible volume, but such minimum volume convex polytopes may not be unique. The purpose of condition (S2) is to determine the "orientation" of the convex polytope and consequently to ensure that it is unique. When (S2) is violated, it is possible to rotate the convex polytope to produce different feasible convex polytopes of the same volume; see Figure 2(b)(c).

The SS condition was first introduced by Huang et al. (2016) to study the identifiability of topic models, where identifiability is ensured under the SS condition along with a minimal determinant on $\mathbf{W}\mathbf{W}^T$. This condition is used differently in their work and ours: Huang et al. (2016) impose the SS condition on rows of $\mathbf{C}$; we impose this condition on columns of $\mathbf{W}$. Although volume is not discussed in Huang et al. (2016), imposing the SS condition on rows of $\mathbf{C}$ in fact leads to a convex polytope of maximum volume; in contrast, we seek a convex polytope of the smallest volume.

Remark 2.3 (Algorithm for checking SS condition)**.** Checking $cone(\mathbf{W}) \supseteq \mathcal{K}^*$ in the SS condition is equivalent to verifying whether a convex polytope contains a ball (after being projected to Δ^{k-1}), which is in general an NP-complete problem in computational geometry (Freund and Orlin, 1985;

Huang et al., 2014). Consequently, it can be computationally difficult to provide a definitive conclusion as to whether or not the SS condition holds in high dimensions. However, if making a small probability mistake is allowed, then we propose that the following randomized algorithm to check the SS condition will give the correct answer with acceptable high probability. Since it suffices to verify that $\text{Conv}(\mathbf{W}) \supseteq bd\mathcal{K}^* \bigcap \Delta^{k-1}$, we can independently choose M sample points uniformly from $bd\mathcal{K}^* \bigcap \Delta^{k-1}$ and check whether all of them are in $\text{Conv}(\mathbf{W})$. If $\mathbf{W}$ satisfies the SS condition, then the M sampled points should belong to $\text{Conv}(\mathbf{W})$; if $\mathbf{W}$ does not satisfy the SS condition, then, since the probability of each sampled point falling in $\text{Conv}(\mathbf{W})$ is a fixed number, the probability of making a mistake decays exponentially in M. For real datasets where $\text{Conv}(\mathbf{W})$ is not observed, we can use an estimator of it to empirically check the SS condition by reporting the frequency of sampled points not falling into the estimated $\text{Conv}(\mathbf{W})$.

3 Maximum Integrated Likelihood Estimation

Before introducing the proposed estimator for topic matrix $\mathbf{C}$, let us describe some more notations and the data generating process. Let $\mathbf{X} = (\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(d)})$ denote the observed data as a collection of word sequences. Without loss of generality, we assume each document has the same number of words, denoted by n. Given parameters $(\mathbf{C}, \mathbf{W})$, word sequences from different documents are independent, with the word sequence from the i-th document, $\mathbf{x}^{(i)} = (x_{i,1}, \ldots, x_{i,n})$, being n i.i.d. samples from the categorical distribution $\text{Cat}(\mathbf{u}_i)$, where $\mathbf{u}_i = \mathbf{C}\mathbf{w}_i$ is the V-dimensional probability vector in Δ^{V-1}, and $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,k})$ denotes the i-th column of matrix $\mathbf{W}$. We use $f_n(\cdot \mid \mathbf{u}_i)$ to denote the multinomial likelihood function of the i-th document. Let $\mathbf{c}_j$ denote the j-th topic vector, i.e., the j-th column of matrix $\mathbf{C}$, for $j = 1, 2, \ldots, k$. Under this notation, we can express the word frequency vector $\mathbf{u}_i = \sum_{j=1}^{k} w_{i,j}\mathbf{c}_j$ associated with the i-th document as a convex combination of the topic vectors, where $\mathbf{w}_i$ serves as the mixing weight vector.

3.1 Implicit Volume Minimization

Since our primary interest is on the topic matrix $\mathbf{C}$, we can profile out the nuisance parameters $\mathbf{w}_i$'s by integrating them with respect to some distribution, resulting an integrated likelihood function of $\mathbf{C}$. After that, we can estimate $\mathbf{C}$ by maximizing the integrated likelihood (Berger et al., 1999). We propose to integrate out $\mathbf{w}_i$'s with respect to the *uniform distribution* over simplex Δ^{k-1}, which induces a uniform distribution on $\mathbf{u}_i = \mathbf{C}\mathbf{w}_i$ over $\text{Conv}(\mathbf{C})$. This is because the linear transformation

$\mathbf{w} \mapsto \mathbf{C}\mathbf{w}$ has a constant Jacobian. The integrated likelihood can be formally written as follows:

$$F_{n\times d}(\mathbf{C};\mathbf{X}) = \prod_{i=1}^{d} \int_{\mathrm{Conv}(\mathbf{C})} \frac{f_n(\mathbf{x}^{(i)} \,|\, \mathbf{u})}{|\,\mathrm{Conv}(\mathbf{C})|}\, d\mathbf{u}, \tag{4}$$

where $|\,\mathrm{Conv}(\mathbf{C})|$ denotes the $(k-1)$-dimensional volume of the set $\mathrm{Conv}(\mathbf{C})$. The corresponding maximum likelihood estimator (MLE) is defined to be

$$\hat{\mathbf{C}}_n = \arg\max_{\mathbf{C}} F_{n\times d}(\mathbf{C};\mathbf{X}), \tag{5}$$

where the maximum is over all V-by-k column-stochastic matrices.

Although the integrated likelihood (4) is equivalent to the marginal likelihood from an LDA model after integrating out the mixing weight $\mathbf{w}$ with respect to a Dirichlet($\mathbf{1}_k$) prior, we emphasize again that the uniform prior is just used to profile out the nuisance parameters so that we can derive an MLE for the topic matrix. In our theoretical analysis below, we do not assume data to be generated from the LDA model with a uniform prior on $\mathbf{w}$.

Why uniform distribution? To understand the motivation behind the use of a uniform distribution in (4), let us consider the noiseless case (corresponding to the limiting case as $n \to \infty$), in which we "observe" the true word-frequency vectors for the d documents: $\mathbf{u}_1^0, \cdots, \mathbf{u}_d^0$. In this ideal setting, from a standard Laplace approximation argument, the i-th integral inside the product in (4) after rescaling by a factor of order $n^{(V-1)/2}$ converges to $\mathbb{1}(\mathbf{u}_i^0 \in \mathrm{Conv}(\mathbf{C}))$, and the MLE $\hat{\mathbf{C}}$ becomes:

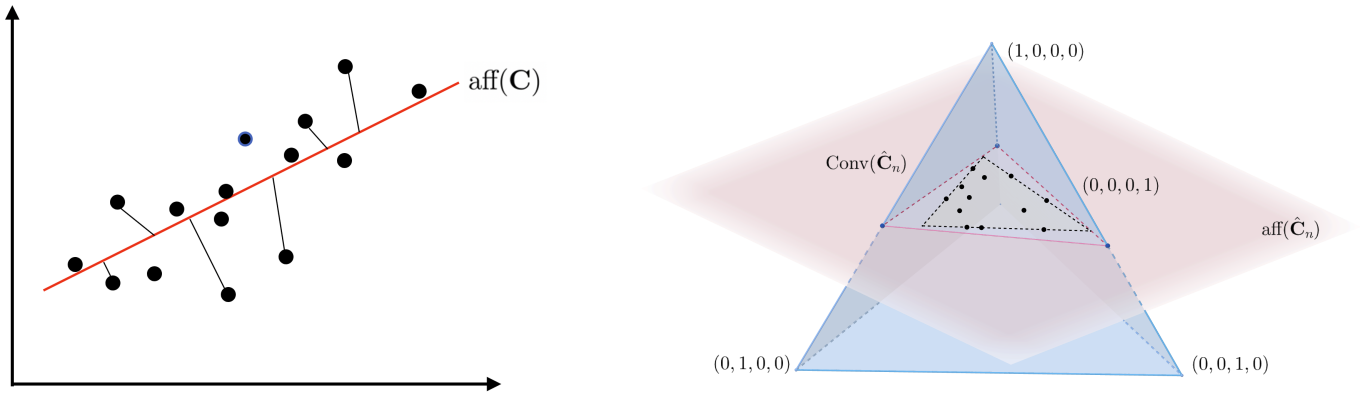
$$\arg\max_{\mathbf{C}} \prod_{i=1}^{d} \frac{\mathbb{1}(\mathbf{u}_i^0 \in \mathrm{Conv}(\mathbf{C}))}{|\,\mathrm{Conv}(\mathbf{C})|} = \arg\max_{\mathbf{C}} \frac{\mathbb{1}(\mathbf{u}_1^0, \cdots, \mathbf{u}_d^0 \in \mathrm{Conv}(\mathbf{C}))}{|\,\mathrm{Conv}(\mathbf{C})|}, \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Therefore, maximizing the integrated likelihood function (4) is asymptotically equivalent to minimizing the volume of $\mathrm{Conv}(\mathbf{C})$ subject to the constraint that $\mathrm{Conv}(\mathbf{C})$ contains all true word-frequency vectors.

In the rest of this section we first provide an alternative interpretation of our approach as a two-stage estimation procedure. We compare it with some representative topic learning methods designed under the separability condition that can also be cast as two-stage procedures. After that, we describe an MCMC-EM algorithm designed for implementing the optimization problem of maximizing the integrated likelihood.

3.2 Interpretation as Two-Stage Optimization

Our method of estimating $\mathbf{C}$ can be viewed as a two-stage procedure: in the first stage, we estimate the $(k-1)$-dimensional hyperplane $\mathrm{aff}(\mathbf{C})$ in which the convex polytope of $\mathbf{C}$ lies; then in the second stage,



(a) First stage

(b) Second stage

Figure 3: Illustration of the two-stage perspective of maximizing the integrated likelihood (4). The left figure illustrates the first stage when $V = k = 2$. Black dots are the sample word frequency vectors $\hat{\mathbf{u}}^{(i)}$'s. aff($\mathbf{C}$) is the red line. We target to minimize the sum of the squared distances, where each distance is induced from its own local norm $\|\cdot\|_i$ (see main text for details about the norm). That is why the black lines correspond to the projection directions are not necessarily parallel to each other. The right figure illustrates of the second stage when $V = 4, k = 3$. The blue tetrahedron is the simplex Δ^{V-1} and the red hyperplane is the estimated aff($\hat{\mathbf{C}}_n$) from the first stage. The black dots are the projections of the sample word frequency vectors $\hat{\mathbf{u}}^{(i)}$'s on aff($\hat{\mathbf{C}}_n$). The black dashed triangle is our estimator $\hat{\mathbf{C}}_n$, whose convex hull is roughly the 3-vertex convex polytope that encloses all the black dots and has the minimal volume.

we determine the boundary of Conv($\mathbf{C}$) by estimating its k vertices within the estimated hyperplane obtained in the first stage. See Figure 3 for an illustration, and the following for a heuristic derivation.

It is worth mentioning that many recent separability condition based topic modeling methods in the literature (such as Arora et al. (2012); Azar et al. (2001); Kleinberg and Sandler (2008, 2003); Ke and Wang (2017); Papadimitriou et al. (2000); McSherry (2001); Anandkumar et al. (2012)) can be explained under this general two-stage framework. For example, some papers (Azar et al., 2001; Kleinberg and Sandler, 2008, 2003) aim only at recovering the column span of topic matrix $\mathbf{C}$ using singular value decomposition (SVD), which suffices for their applications. This corresponds to solving the hyperplane estimation problem in our first stage. Some papers (Arora et al., 2012; Papadimitriou et al., 2000; McSherry, 2001; Anandkumar et al., 2012) directly search for a subset of words (separability condition on anchor words, Arora et al. (2012)) or documents (separability condition on pure topic documents, Papadimitriou et al. (2000); McSherry (2001); Anandkumar et al.

(2012)) in their first stage, and then in their second stage recover the population-level term-document matrix (or the hyperplane aff($\mathbf{C}$)) based on the estimated anchor words/pure topic documents. This corresponds to our two-stage procedure, in reverse order. Others such as Ke and Wang (2017) also use a two-stage procedure based, first, on projecting a certain transformation of the sample term-document matrix onto a lower-dimensional hyperplane via SVD, and then searching for the anchor words over that hyperplane. Notice that all aforementioned methods reply crucially on the separability condition, which greatly simplifies the statistical and computational hardness of the problem and turns it into a searching problem; thus they are able to circumvent the hidden non-regular statistical problem of boundary estimation (c.f. Section 4.3).

To illustrate the two-stage interpretation of our method, we observe that the integrated likelihood (4) is equivalent to the following expression:

$$\frac{1}{|\operatorname{Conv}(\mathbf{C})|^d} \prod_{i=1}^{d} \int_{\operatorname{Conv}(\mathbf{C})} \exp\big\{ -n\, D_{\mathrm{KL}}(\hat{\mathbf{u}}^{(i)} \,||\, \mathbf{u})\big\}\, d\mathbf{u}, \tag{7}$$

where $\hat{\mathbf{u}}^{(i)}$ denotes the sample word frequency vector for document i. Here, we use $D_{\mathrm{KL}}(\mathbf{p}\,||\,\mathbf{q}) = \sum_{v=1}^{V} p_v \log(p_v/q_v)$ to denote the Kullback-Leibler divergence between two categorical distributions with parameters $\mathbf{p} = (p_1, \ldots, p_V)$ and $\mathbf{q} = (q_1, \ldots, q_V)$. When n is large, the classical Laplace approximation to the integral in (7) uses a nonnegative quadratic form $\|\mathbf{u} - \hat{\mathbf{u}}^{(i)}\|_i^2 := (\mathbf{u} - \hat{\mathbf{u}}^{(i)})^T \mathbf{H}_i (\mathbf{u} - \hat{\mathbf{u}}^{(i)})$ to approximate the exponent $D_{\mathrm{KL}}(\hat{\mathbf{u}}^{(i)} \,||\, \mathbf{u})$ in a local neighborhood of $\hat{\mathbf{u}}^{(i)}$. Since such a quadratic form defines the norm $\|\cdot\|_i$, we can decompose it into $\|\mathbf{u} - \hat{\mathbf{u}}^{(i)}\|_i^2 = \|\mathbf{u} - \mathbb{P}_{\mathbf{C}}^{(i)}\, \hat{\mathbf{u}}^{(i)}\|_i^2 + \|\big(\mathbf{I}_V - \mathbb{P}_{\mathbf{C}}^{(i)}\big)\, \hat{\mathbf{u}}^{(i)}\|_i^2$, where $\mathbb{P}_{\mathbf{C}}^{(i)}$ denotes the projection operator onto the $(k-1)$-dimensional hyperplane aff($\mathbf{C}$) with respect to the distance induced from $\|\cdot\|_i$. Finally, we can approximate the integrated likelihood in the preceding display as

$$\exp\Big\{ -n \underbrace{\sum_{i=1}^{d} \|\big(\mathbf{I}_V - \mathbb{P}_{\mathbf{C}}^{(i)}\big)\, \hat{\mathbf{u}}^{(i)}\|_i^2}_{\text{residual sum of squares}} \Big\} \cdot \frac{1}{|\operatorname{Conv}(\mathbf{C})|^d} \prod_{i=1}^{d} \underbrace{\int_{\operatorname{Conv}(\mathbf{C})} \exp\big\{ -n\, \|\mathbf{u} - \mathbb{P}_{\mathbf{C}}^{(i)}\, \hat{\mathbf{u}}^{(i)}\|_i^2 \big\}\, d\mathbf{u}}_{\approx C_i\, n^{-(k-1)/2}\, \mathbb{1}(\mathbb{P}_{\mathbf{C}}^{(i)}\, \hat{\mathbf{u}}^{(i)} \in \operatorname{Conv}(\mathbf{C}))}, \tag{8}$$

where the display underneath the second curly bracket is due to the Laplace approximation to the $(k-1)$-dimensional integral, and the constants C_i depends only on $\hat{\mathbf{u}}^{(i)}$.

We see from this approximation that the maximization of integrated likelihood (7) can be approximately cast into a two-stage sequential optimization problem. In the first stage, we find an optimal $(k-1)$-dimensional hyperplane spanned by $\mathbf{C}$ that is closest to $\hat{\mathbf{u}}^{(i)}$'s by minimizing the residual sum of squares in (8) (see Figure 3(a)). This corresponds to the SVD approach for estimating the true topic supporting hyperplane adopted by Azar et al. (2001); Kleinberg and Sandler (2008, 2003); Ke

and Wang (2017), and several others under the separability condition. In the second stage, we find the most compact (i.e., minimal volume) k-vertex convex polytope $\mathrm{Conv}(\mathbf{C})$ that encloses the projections of $\hat{\mathbf{u}}^{(i)}$'s onto the hyperplane $\mathrm{aff}(\mathbf{C})$, so that the second term in (8) is maximized.

With the separability condition on $\mathbf{C}$ or $\mathbf{W}$, the vertex search in the second stage can be greatly simplified and restricted to a small number of choices. For example, the anchor-word assumption implies that each column of $\mathbf{C}$ has at least $(k-1)$ zeros; consequently, columns of $\mathbf{C}$ should be chosen from the intersection of $\mathrm{aff}(\mathbf{C})$ and the simplex Δ^{V-1} in the second stage (as shown in Figure 3(b)).

Our second stage, in the absence of a separability condition, is essentially the much more challenging non-regular statistical problem of boundary estimation. To see this, consider the same toy example of $(V, k) = (4, 3)$ as illustrated in Figure 3(b). The separability condition on $\mathbf{C}$ implies that once the hyperplane $\mathrm{aff}(\mathbf{C})$ (red hyperplane) is determined, the only candidate topic matrix $\mathbf{C}$ is the one whose columns are the intersections (blue circles) of this hyperplane and the three 1-dimensional edges of the simplex Δ^3 (blue tetrahedron), making the second stage trivial. On the contrary, the statistical problem in our setting is to estimate the minimal volume k-vertex convex polytope (black dashed triangle as our estimator) that encloses all true underlying word probability vectors of the documents, which is highly nontrivial (see Section 4.3 for a more detailed comparison). Fortunately, our computational algorithm described in the following subsection circumvents this difficulty directly maximizing the integrated likelihood via a variant of the expectation maximization (EM) algorithm, which implicitly constructs such an estimator.

3.3 Computing Maximum Integrated Likelihood Estimator

For computation, we employ an MCMC-EM algorithm to find the maximizer $\hat{\mathbf{C}}_n$ of the integrated likelihood objective (4) by augmenting the model with a set of latent variables $\mathbf{Z} = \{Z_{ij} : i = 1, 2, \ldots, d,\ j = 1, 2, \ldots, n\}$, where, given the mixing weights $\mathbf{w}_i$, $Z_{ij} \in \{1, 2, \ldots, k\}$ follows $\mathrm{Cat}(\mathbf{w}_i)$ and is interpreted as the topic indicating variable for the j-th word $\mathbf{x}_j^{(i)}$ in the i-th document. Our MCMC-EM algorithm proceeds in a manner similar to that of the classical EM algorithm with, first, an **E**-step of computing the expected log-likelihood function $\log p(\mathbf{X}, \mathbf{Z} \,|\, \mathbf{C})$, where the expectation is with respect to the distribution of latent variable $\mathbf{Z}$ after marginalizing out $\mathbf{W}$, and then an **M**-step of maximizing the expected log-likelihood function over topic matrix $\mathbf{C}$. An MCMC scheme is introduced in the **E**-step for sampling $(\mathbf{Z}, \mathbf{W})$ pairs from the joint conditional distribution of $p(\mathbf{Z}, \mathbf{W} \,|\, \mathbf{X}, \mathbf{C})$ in order to compute the expected log-likelihood function via Monte-Carlo approximation.

As discussed before, our proposed estimator is essentially the MLE estimator from the LDA model

(Blei et al., 2003) with a particular choice of priors on $\mathbf{W}$. Many algorithms have been proposed for the LDA model, such as the Gibbs sampler (Griffiths and Steyvers, 2004), partially collapsed Gibbs samplers (Magnusson et al., 2018; Terenin et al., 2018), and various variational algorithms (Blei et al., 2003). The use of MCMC-EM here is a personal preference. Our MCMC-EM algorithm is a stochastic EM algorithm similar to the Gibbs sampler in Griffiths and Steyvers (2004), and to the partially collapsed Gibbs samplers in Magnusson et al. (2018); Terenin et al. (2018). According to the asymptotic results of stochastic EM algorithms in Nielsen et al. (2000), the estimation of the topic matrix produced by our algorithm is guaranteed to converge to the proposed MLE, provided that $\mathbf{W}^0$ is sufficiently scattered. In Section 5.2, we compare our algorithm with the algorithms mentioned above and find all very similar in performance. Since computation is not the main focus of this paper, we confine the details, including derivations for the full algorithm, to the supplementary material.

4 Finite-Sample Error Analysis

In this section, we study the finite-sample error bound and its implied asymptotic consistency of the proposed estimator $\hat{\mathbf{C}}_n$. We consider the fixed design setting where columns of $\mathbf{W}$ can take arbitrary positions in Δ^{k-1} as long as a perturbed version of the SS condition described in the following is satisfied. For the stochastic setting where columns of $\mathbf{W}$ are generated from some distribution, the error analysis and consistency can be found from Section B in the supplementary material. To avoid ambiguity, we use $\mathbf{C}^0$, $\mathbf{W}^0$, $\mathbf{U}^0$ to denote the ground truth, and leave $\mathbf{C}$, $\mathbf{W}$, $\mathbf{U}$ as generic notations for parameters.

4.1 Noise Perturbed SS Condition

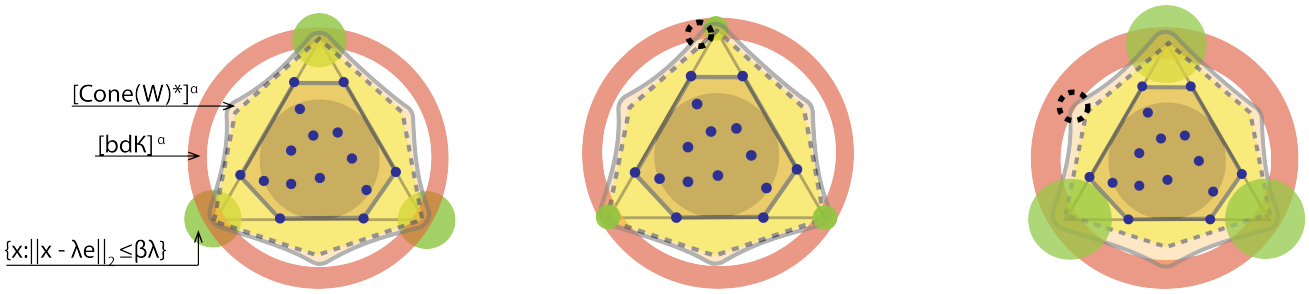
Before introducing our results from the error analysis, it is helpful to introduce a perturbed version of the SS condition, called (α, β)-SS condition, which characterizes the robustness/stability of the (population level) SS condition against random noise perturbation due to the finite sample size.

Definition 3 ((α, β)-SS Condition)**.** *A matrix* $\mathbf{W}$ *is* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$**-sufficiently scattered** *for some* $\alpha, \beta \geqslant 0$*, if it satisfies (S1) and*

(S3). $[cone(\mathbf{W})^*]^\alpha \bigcap [bd\mathcal{K}]^\alpha \subseteq \{\mathbf{x} : \|\mathbf{x} - \lambda \mathbf{e}_f\|_2 \leqslant \beta\lambda, \lambda \geqslant 0\}$*, where*

$[cone(\mathbf{W})^*]^\alpha = \{\mathbf{x} : \mathbf{x}^T\mathbf{W} \geqslant -\alpha\|\mathbf{x}\|_2\}$ *and* $[bd\mathcal{K}]^\alpha = \{\mathbf{x} : |\|\mathbf{x}\|_2 - \mathbf{x}^T\mathbf{1}_k| \leqslant \alpha\|\mathbf{x}\|_2\}$ *are the* α*-enlargements of* $cone(\mathbf{W})^*$ *and* $bd\mathcal{K}$*, respectively.*

We provide a geometric view of the (α, β)-SS condition in Figure 4. Similar to the setting of Figure 2, everything is projected onto the hyperplane $\mathbf{x}^T\mathbf{1}_k = 1$: blue dots denote columns of $\mathbf{W}$, the inner brown ball inscribed in the triangle denotes $\mathcal{K}^*$, and the shaded yellow region denotes $cone(\mathbf{W})^*$ along with the dashed gray line as its boundary. The boundary of the enlarged cone of $cone(\mathbf{W})^*$, $[cone(\mathbf{W})^*]^\alpha$, is marked by the solid gray line, and the thickened boundary of $\mathcal{K}$, $[bd\mathcal{K}]^\alpha$, is the outside ring in red. The set $\{\mathbf{x} : \|\mathbf{x} - \lambda\mathbf{e}_f\|_2 \leqslant \beta\lambda, \lambda \geqslant 0, f \in [k]\}$, when being projected to the hyperplane $\mathbf{x}^T\mathbf{1}_k = 1$, corresponds to the green balls centered at the vertices of Δ^{k-1} with radius β.



(a) (α, β)-SS

(b) not (α_2, β_2)-SS
$(\alpha_2 = \alpha, \beta_2 < \beta)$

(c) not (α_3, β_3)-SS
$(\alpha_3 > \alpha, \beta_3 > \beta)$

Figure 4: Geometric view of (α, β)-SS sliced at the hyperplane $\mathbf{x}^T\mathbf{1}_k = 1$ $(k = 3)$. $\mathbf{W}$ is the same in (a)(b)(c) while the values of α and β are different. In (b) and (c), we highlight the region (the dashed circle) that are in $[cone(\mathbf{W})^*]^\alpha \bigcap [bd\mathcal{K}]^\alpha$ but not in $\{\mathbf{x} : \|\mathbf{x} - \lambda\mathbf{e}_f\|_2 \leqslant \beta\lambda\}$.

For a matrix $\mathbf{W}$ to satisfy the (α, β)-SS condition, the corresponding convex hull of the blue dots need to contain $\mathcal{K}^*$, the inner brown ball. In addition, the intersection of the red ring, $[bd\mathcal{K}]^\alpha$, and the region enclosed by the solid gray line, $[cone(\mathbf{W})^*]^\alpha$, must be inside the green balls; see Figure 4(a). In other words, $[cone(\mathbf{W})^*]^\alpha$ only touches $[bd\mathcal{K}]^\alpha$ near the k vertices of the simplex Δ^{k-1}.

The (α, β)-SS condition can be viewed as a generalization of the SS condition with the two parameters (α, β) quantifying the robustness of $cone(\mathbf{W})^*$ under noise perturbation. In particular, α characterizes the *tolerable noise level*, and β, which we refer to as the vertices *sensitivity coefficient*, represents the maximum estimation error induced by noises below level α. Due to this interpretation, the (α, β)-SS condition becomes stronger as α increases and β decreases (c.f. Proposition 3). In particular, the minimal allowable β under (S3) should increase as α increase. In most examples, β should be proportional to α up to some constant depending on the geometric structure of $cone(\mathcal{K})$ (for a concrete example, c.f. Proposition 5).

While the SS condition requires $cone(\mathbf{W})^*$ and $bd\mathcal{K}$ to intersect exactly at the positive semi-axis rays $\{\lambda\mathbf{e}_f, \lambda \geqslant 0\}$, the (α, β)-SS condition requires the intersection of $[cone(\mathbf{W})^*]^\alpha$ and $[bd\mathcal{K}]^\alpha$—the

perturbed versions of $cone(\mathbf{W})^*$ and $bd\mathcal{K}$, respectively, with noise level α—to be within distance β away from the semi-axis rays. Note that (α, β)-SS degenerates to the SS condition when $\alpha = \beta = 0$.

Intuitively, if a matrix $\mathbf{W}$ has vertices sensitivity coefficient β under noise level α, then condition (S3) remains valid at the same sensitivity coefficient as we decrease the noise level and at the same tolerable noise level as we increase the sensitivity coefficient. The following proposition provides a more general picture about the relation of the (α, β)-SS conditions under different combinations of (α, β).

Proposition 3. *The followings are some properties of (α, β)-SS condition and SS condition.*

(i) If $\alpha \geqslant \alpha'$ and $\beta \leqslant \beta'$, then (α, β)-SS implies (α', β')-SS.

(ii) If $\mathbf{W}$ is (α, β)-SS and $\mathrm{Conv}(\mathbf{W}) \subseteq \mathrm{Conv}(\bar{\mathbf{W}})$, then $\bar{\mathbf{W}}$ is also (α, β)-SS.

(iii) If $\mathbf{W}$ is SS and $cone(\mathbf{W}) \subseteq cone(\bar{\mathbf{W}})$, then $\bar{\mathbf{W}}$ is also SS.

By Proposition 3(i), the (α, β)-SS condition gets more stringent if we increase the tolerable noise level α and/or reduce the vertices sensitivity coefficient β. This is because when α gets larger, the intersection $[cone(\mathbf{W})^*]^\alpha \bigcap [bd\mathcal{K}]^\alpha$ gets larger and consequently may not be packed inside the green ball with radius β. Similarly, when β gets smaller, the green balls may not be large enough to contain the intersection. See Figure 4(b)(c) for illustration. Since $\mathrm{Conv}(\mathbf{W}) \subseteq \mathrm{Conv}(\bar{\mathbf{W}})$ implies $cone(\mathbf{W}) \subseteq cone(\bar{\mathbf{W}})$, we provide a more general sufficient condition for SS in Proposition 3(iii) compared to that in Proposition 3(ii), where SS is a special case of (α, β)-SS. However, in this paper, the columns of $\mathbf{W}$ we consider are all on the hyperplane $\mathbf{x}^T \mathbf{1}_k = 1$, so $\mathrm{Conv}(\mathbf{W}) \subseteq \mathrm{Conv}(\bar{\mathbf{W}})$ is equivalent to $cone(\mathbf{W}) \subseteq cone(\bar{\mathbf{W}})$. As a direct consequence of Proposition 3(ii), if some columns of $\mathbf{W}$ is (α, β)-SS, then $\mathbf{W}$ is (α, β)-SS.

The maximal allowable tolerable noise level α is determined by the geometric structure of $cone(\mathbf{W})$. Given α, the (α, β)-SS condition can be satisfied by almost any $\mathbf{W}$ when β is large enough. However, such a condition is meaningless since β will appear as one of the error terms later in Theorem 4. So we would like to set β as small as possible in order to derive a tight error bound. For example, we need β to have an order of $\sqrt{\frac{\log(n \vee d)}{n}}$ in Theorem 4 to ensure a desired error rate that matches the order of our α choice reflecting the effective noise level in the data.

4.2 Error Analysis and Consistency

In this subsection, we consider the setting where columns of $\mathbf{W}$ are fixed, and satisfy a set of conditions related to the noise perturbed SS condition discussed in the previous subsection. Note that the results

in this subsection also apply to randomly generated mixing weights, as long as we can verify that the set of conditions below holds for the random mixing weights with high probability (c.f. Section B in the supplementary material). Before presenting our main results on the finite-sample error bound of the estimator $\hat{\mathbf{C}}_n$, let us first state our assumptions.

Assumptions. *Assume the following:*

(A1) $\mathbf{C}^0$ *is of rank* k *and its columns are bounded away from the boundary of* Δ^{V-1}.

(A2) Eigenvalues of $\frac{1}{d}\mathbf{W}_c{\mathbf{W}_c}^T$ *are lower bounded by a positive constant, where* $\mathbf{W}_c = \mathbf{W}^0 - \frac{1}{d}\mathbf{W}^0\mathbf{1}_d\mathbf{1}_d^T$ *is the centered version of* $\mathbf{W}^0$. *In addition, there exist* k *affinely independent columns of* $\mathbf{W}^0$ *with minimum positive singular value larger than a positive constant.*

(A3) There exist s *columns of* $\mathbf{W}^0$ *which are* (α, β)*-SS with* $\alpha \geqslant C_1\sqrt{\frac{s\log(n\vee d)}{n}}$, *where* s *and* C_1 *are constants.*

Now we are ready to present our main result on the estimation accuracy.

Theorem 4. *Under Assumptions (A1)-(A3), with probability at least* $(1 - 3/(n \vee d)^c)^d$,

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant D_1\sqrt{\frac{s\log(n\vee d)}{n}} + D_2\sqrt{s}\beta, \tag{9}$$

where c, D_1 *and* D_2 *are positive constants. In particular, if* $\beta \leqslant C_2\sqrt{\frac{\log(n\vee d)}{n}}$ *where* C_2 *is a constant, then*

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant D_1'\sqrt{\frac{s\log(n\vee d)}{n}}. \tag{10}$$

In the theorem, constants D_1 and c have the relation that $D_1 = C_3 \cdot \sqrt{c} + C_4$ where C_3 and C_4 are constants independent of (n, d). Some remarks about the assumptions are in order.

(A1) is commonly imposed for technical reasons in other related work, such as Nguyen (2015) and Wang (2019), to avoid singularity issues. The geometric interpretation of the assumption in (A2) on $\mathbf{W}_c$ is that $\mathrm{Conv}(\mathbf{U}_0)$ should contain a ball of a constant radius, which is again imposed to avoid singularity issues when a large proportion of the mixing weight vectors are too concentrated. Similar assumptions are also made in Ke and Wang (2017); Javadi and Montanari (2020).

Next, we discuss Assumption (A3) in detail. First, note that a subset of columns of $\mathbf{W}^0$ satisfying the (α, β)-SS condition immediately implies the full matrix $\mathbf{W}^0$ itself to satisfy the same condition, due to Proposition 3(ii). Second, note that to attain the error bound (10) we need the existence of a sub-matrix $\mathbf{W}^0$ to satisfy condition (A3) with β of the same order as α. The following proposition

provides a sufficient condition for fulfilling this requirement. For example, when $k = 3$ as illustrated in Figure 4(a), all we need are two data points on each of the three line segments connecting $\mathbf{e}_i$ and $\mathbf{e}_j$ $(i \neq j)$ (i.e., totally six points) with the distance from each data point to the nearest vertex is less than 1/3.

Proposition 5. *Suppose for all $1 \leqslant i \neq j \leqslant k$, there exists a column of $\mathbf{W}^0$ that can be represented as $(1 - x_{ij})\mathbf{e}_i + x_{ij}\mathbf{e}_j$ where $0 \leqslant x_{ij} < 1/k$, then $\mathbf{W}^0$ is $(\epsilon, C\epsilon)$-SS for all $\epsilon > 0$, where C is constant only depending on the geometry of $\mathbf{W}^0$.*

Third, we discuss the parameter s, the smallest number of columns in $\mathbf{W}^0$ that are (α, β)-SS, in Assumption (A3). The following proposition shows that when the columns of $\mathbf{W}^0$ are stochastically generated according to some underlying distribution over Δ^{k-1} with appropriate properties, then s can be chosen as a constant with high probability. Note that even if s is not a constant, the error bound in (10) still goes to zero as long as s is of a smaller order of $\frac{n}{\log(n \vee d)}$ in the asymptotic setting where $(n, d) \to \infty$.

Proposition 6. *Suppose the columns of $\mathbf{W}^0$ are i.i.d. samples from a probability density function that is uniformly larger than a positive constant on neighborhoods of the vertices of Δ^{k-1}. If $C \cdot n^{\frac{k-1}{2}} \leqslant d \leqslant e^{n^c}$, then with probability at least $1 - C_0 \cdot k/d$, there exist k columns in $\mathbf{W}^0$ that are $\left(C_1\sqrt{\frac{\log(n \vee d)}{n}}, C_2\sqrt{\frac{\log(n \vee d)}{n}}\right)$-SS, where $c \in (0, 1)$, C, C_0, C_1 and C_2 are positive constants.*

Next, we show the asymptotic consistency of $\hat{\mathbf{C}}_n$, that is, $\hat{\mathbf{C}}_n \to \mathbf{C}^0$ in probability as $(n, d) \to \infty$. In particular, we assume the existence of a sequence of α and β values along which the (α, β)-SS conditions are satisfied, which is summarized in the following.

Assumptions. *Assume the following:*

(A3') *For any sufficiently small $\epsilon > 0$, there exists some β_ϵ such that $\beta_\epsilon \to 0$ when $\epsilon \to 0$, and there are s columns of $\mathbf{W}^0$ satisfying the $(\epsilon, \beta_\epsilon)$-SS condition, where s is a bounded constant.*

(A4) $\log d/n \to 0$ *as* $(n, d) \to \infty$.

Theorem 7 (Estimation Consistency)**.** *Under Assumptions (A1), (A2) and (A3') with a fixed d, we have*

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \to 0 \quad \text{in probability as } n \to \infty. \tag{11}$$

If d is also increasing in n in a way such that Assumption (A4) holds, then

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \to 0 \quad \text{in probability as } (n, d) \to \infty. \tag{12}$$

Note that Proposition 5 again provides a set of sufficient conditions for Assumption (A3'). However, our current condition on $\mathbf{W}^0$ in Proposition 5 is stronger than the SS condition on $\mathbf{W}^0$. We conjecture that Assumption (A3') is equivalent to the SS condition on $\mathbf{W}^0$, and leave a formal proof to future work.

4.3 Comparison with Existing Theoretical Results

Our error bound in Theorem 4 does not decay as the number of documents d increases, which is seemingly weaker than some existing results, such as Arora et al. (2012), Bansal et al. (2014), Anandkumar et al. (2014), Ke and Wang (2017), and Wang (2019). In particular, under the anchor word assumption, Arora et al. (2012) and Ke and Wang (2017) showed an error upper bound as $1/\sqrt{nd}$.

As discussed in Section 3.2, many algorithms for estimating the topic matrix can be explained through a two-stage optimization, corresponding to either a single stage or both. Under this perspective, each stage will incur an error. With the anchor word assumption, the main source of errors comes from the first stage of applying an SVD approach (Azar et al., 2001; Kleinberg and Sandler, 2008, 2003; Ke and Wang, 2017) to find a $(k-1)$-dimensional hyperplane best approximating the data whose error bound is $1/\sqrt{nd}$. In fact, the anchor word assumption greatly reduces the search space in the second stage of identifying columns of $\mathbf{C}$ as either a subset of anchor words or a subset of pure topic documents, yielding negligible estimation error. For example, the vertex hunting algorithm adopted in Ke and Wang (2017) directly focuses on all the k combinations of the noisy data points in the $(k-1)$-dimensional hyperplane obtained in the first stage, and chooses the combination that minimizes the predetermined criterion. With the separability condition, they show that the estimated vertices are all close to their corresponding true vertices in a $(k-1)$-dimensional hyperplane, from which they draw the conclusion that the estimation error of the second stage is no larger than that of the first stage (see Lemma A.3, Ke and Wang (2017)).

Without the anchor word (or separability) assumption, errors incurred in the second stage become dominant. Consider the toy examples illustrated in Figures 1 and 2 with $K = V = 3$. The first stage is trivial since the data are already in $(k-1)$-dimension and projection to a hyperplane is not needed. In the second stage, we need to estimate a k-vertex convex polytope enclosing all true word probability vectors of the documents that generates the data, which can be formulated as the non-regular statistical problem of boundary estimation. As pointed out by Goldenshluger and Tsybakov (2004); Brunel et al. (2021), estimation of convex supports from noisy measurements as in our second stage is an extremely difficult problem. For example, in the one-dimensional case, even with the knowledge that the noises

are homogeneous and follow a known Gaussian distribution, the minimax rate of boundary estimation based on d observations is as slow as $1/\sqrt{\log d}$, let alone the more complex situation where the noise distribution is heterogeneous and only partly known. For example, in our case the projection $\mathbb{P}_{\mathbf{C}}^{(i)}\,\hat{\mathbf{u}}^{(i)}$ onto aff($\mathbf{C}$) of the sample word frequency vector $\hat{\mathbf{u}}^{(i)}$ for document i, for $i = 1, \ldots, d$, plays the role of a noisy measurement from the convex polytope Conv($\mathbf{C}$). Note that a typical noise level in our second stage is of order $1/\sqrt{n}$ due to n number of words within each document; however, the error distribution depends on both the position of the hyperplane aff($\mathbf{C}$) obtained in the first stage as well as the location of $\mathbb{P}_{\mathbf{C}}^{(i)}\hat{\mathbf{u}}^{(i)}$ on the data simplex Δ^{V-1}. Therefore, we cannot expect to achieve the $1/\sqrt{nd}$ error bound as those separability condition based methods. It is an interesting open problem of determining the precise minimax-optimal rate in topic models without separability condition and whether our error bound is optimal, which we leave as a future direction.

5 Empirical Studies

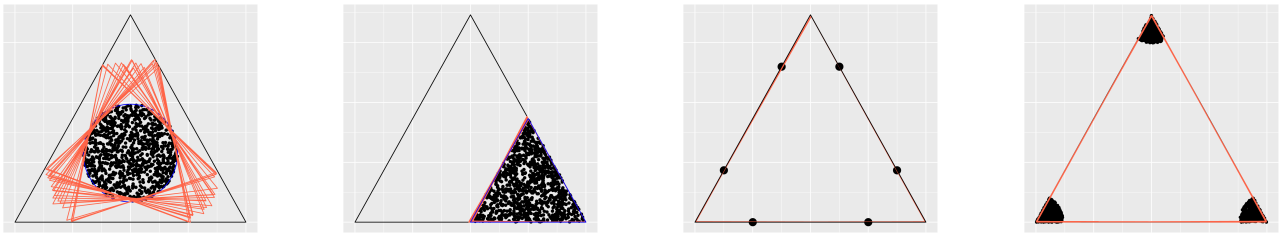
In this section, we describe numerical studies we have performed to test our theoretical results. We report the performance of our model on two real datasets.

5.1 Simulation Studies

We have conducted three simulation studies to verify our theoretical results and to test the performance of our proposed algorithms. In Section 5.1.1, we apply the MCMC-EM algorithm to the data generated by non-identifiable and identifiable models, and compare the recovered convex polytopes with the truth, to show the importance of the SS condition. In Section 5.1.2, we compare the proposed uniform prior $\boldsymbol{\beta}_0 = \mathbf{1}_k$ with other priors, using data generated from different distributions, to demonstrate empirically the robust performance of our estimator. In Section 5.1.3, we apply Monte Carlo simulation to visualize the convergence of the proposed MLE.

5.1.1 Effect of the SS Condition

Data are generated from a simple setup: $k = V = 3$, $\mathbf{C}^0 = \mathbf{I}_3$, and the number of words for each document is sampled from Poisson(2000). For the true matrix $\mathbf{W}^0$, we consider four different configurations for $\mathbf{w}_i^0$: (a) concentrated in the center of Δ^2; (b) concentrated in the bottom right; (c) satisfying the SS condition; (d) spread around three vertices. The four configurations are displayed in



(a) non-identifiable (b) non-identifiable (c) identifiable (d) identifiable

Figure 5: Results of the simulation in Section 5.1.1. Black dots are columns of $\mathbf{W}^0$; the black triangle is the ground truth $\text{Conv}(\mathbf{C}^0) = \Delta^2$; red triangles are estimates of $\text{Conv}(\hat{\mathbf{C}}_n)$.

Figure 5, where the black dots denote $\mathbf{w}_i^0$ and the large black triangle represents $\text{Conv}(\mathbf{C}^0) = \Delta^2$. In cases (a)(b)(d), we set the number of documents $d = 1000$, while in case (c) we set $d = 6$.

We run our MCMC-EM algorithm 20 times with different initialization; Figure 5 displays the estimates of $\text{Conv}(\hat{\mathbf{C}}_n)$ as red triangles. Our simulation results demonstrate that if the SS condition is not satisfied, even when the sample size d is fairly large ($d = 1000$ in (a) and (b)), $\text{Conv}(\mathbf{C}^0)$ cannot be correctly recovered. However, when SS is satisfied, even with just a few samples ($d = 6$ in (c)), our algorithm can accurately recover the ground truth. Identifiability is thus determined primarily by the scatteredness of $\mathbf{w}_i^0$ rather than by the number of documents d.

5.1.2 Performance under Prior Misspecification

When deriving our estimator, we choose to integrate over the mixing weights with respect to the uniform prior. A natural question is how our estimator would perform when the true mixing weight $\mathbf{W}^0$ is stochastically generated from a distribution other than uniform.

In this simulation study we consider the following setup: $k = 3$, $V = 1000$, $d = 200$, $\mathbf{C}^0 \sim \text{Dirichlet}_V(\mathbf{1})$, and the number of words for each document is generated from Poisson(20000). The true mixing weights $\mathbf{W}^0$ are stochastically generated from the following distributions: (a) $\text{Dirichlet}_3(\mathbf{1})$; (b) uniformly from 10 Euclidean balls whose centers satisfy the SS condition; (c) a mixture of Dirichlet distributions: $0.2 \times \text{Dir}_3(10, 1, 1) + 0.2 \times \text{Dir}_3(0.1, 1, 1) + 0.2 \times \text{Dir}_3(10, 10, 1) + 0.2 \times \text{Dir}_3(0.1, 0.1, 1) + 0.2 \times \text{Dir}_3(1, 2, 3)$.

We compare our estimator and estimators based on other Dirichlet priors using the averaged Relative RMSE (i.e., RMSE divided by the average of RMSE of random guesses) of $\hat{\mathbf{C}}_n$ over 100 replications. The results are reported in Table 1. We can see that in all three cases, our proposed estimator outperforms other estimators.

Table 1: Relative RMSE (Simulation 2).

priors	(1, 1, 1)	(0.1, 0.1, 0.1)	(10, 1, 1)	(0.1, 1, 1)	(0.1, 0.1, 1)	(10, 1, 0.1)	(1, 2, 3)	(3, 3, 3)
case (a)	**0.048**	0.064	0.058	0.059	0.061	0.068	**0.048**	0.049
case (b)	**0.053**	0.065	0.062	0.060	0.061	0.075	0.056	0.057
case (c)	**0.040**	0.042	0.048	**0.040**	0.041	0.049	0.042	0.044

5.1.3 Convergence of the Estimation

We use the Monte Carlo simulation to show the convergence of the integrated likelihood $F_{n\times d}(\mathbf{C})$ and the MLE $\hat{\mathbf{C}}_n$.

In the first experiment, we consider the setup where $V = 9$, $k = 3$, and the sample size n and number of documents d increase simultaneously. The sample size n varies as $n = 50, 200, 400, 1600$ and $d = n/5$. Let

$$\mathbf{C}^0 = \begin{bmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 2/3 & 1/6 \\ 1/6 & 1/6 & 2/3 \end{bmatrix}, \quad \mathbf{W}^0 = \begin{bmatrix} 5/6 & 0 & 1/6 & 5/6 & 1/6 & 0 \\ 1/6 & 5/6 & 0 & 0 & 5/6 & 1/6 \\ 0 & 1/6 & 5/6 & 1/6 & 0 & 5/6 \end{bmatrix}.$$

We generate the "noiseless" data, i.e., $\mathbf{X} = n\mathbf{C}^1\mathbf{W}^1$, where $\mathbf{C}^1 = \frac{1}{3}\left(\mathbf{C}^{0T}, \mathbf{C}^{0T}, \mathbf{C}^{0T}\right)^T$, the first six columns of $\mathbf{W}^1$ are $\mathbf{W}^0$, and the rest of the columns are randomly generated from $\mathrm{Dir}_k(\mathbf{1})$. We compare the integrated likelihood among candidate topic matrices of the form $\mathbf{C} = \frac{1}{3}\left(\mathbf{A}^T, \mathbf{A}^T, \mathbf{A}^T\right)^T$, where $\mathbf{A}$ is

$$\begin{bmatrix} c & (1-c)/2 & (1-c)/2 \\ (1-c)/2 & c & (1-c)/2 \\ (1-c)/2 & (1-c)/2 & c \end{bmatrix}, \tag{13}$$

with c taking values from $[0.5, 1]$. We use the Monte Carlo method to evaluate the integrated likelihood (4):

$$\hat{F}_{n\times d,T}(\mathbf{C}) \approx \prod_{i=1}^{d}\left[\frac{1}{T}\sum_{t=1}^{T} f_n(\mathbf{x}^{(i)}|\mathbf{u} = \mathbf{C}\mathbf{w}_t)\right],$$

where $\mathbf{w}_1, \cdots, \mathbf{w}_T$ are i.i.d. random samples from $\mathrm{Dir}_k(\mathbf{1})$ and $T = 100,000$.

Figure 6 shows $\hat{F}_{n\times d,T}(\mathbf{C})/\max_{\mathbf{C}} \hat{F}_{n\times d,T}(\mathbf{C})$, the relative value of the estimated integrated likelihood. From the plot we can see that the integrated likelihood converges quickly to the truth as both n and d increase. That is because n is the sample size, and the integrated likelihood is the product of d terms. As d increases, the product is more concentrated.

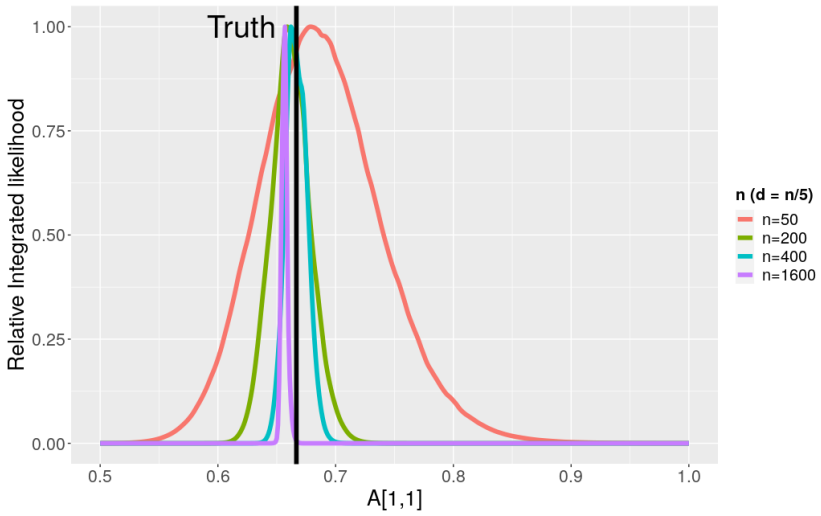
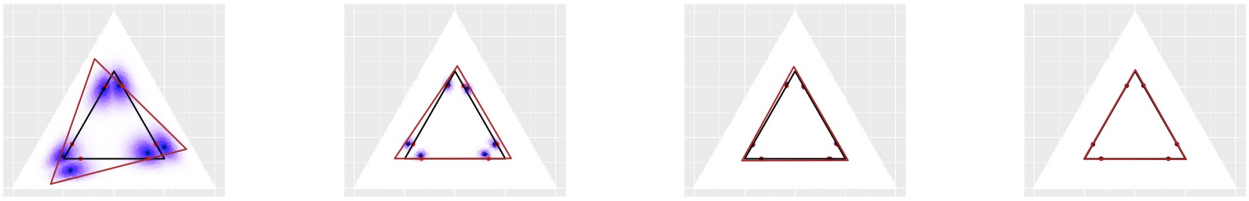


Figure 6: Results of the first experiment in Section 5.1.3. The curves show the relative integrated likelihood of "noiseless" data when n and d increase simultaneously.

In the second experiment, we consider the case where $V = k = 3$ and $d = 6$. We add some noise to the data, i.e., $\mathbf{x}^{(i)} \sim \text{Multi}(n, \mathbf{C}^0\mathbf{w}^{0(i)})$. In Figure 7 we plot the multinomial likelihood density function $f_n(\mathbf{u}; \mathbf{x}^{(i)})$ (represented by the purple clusters) for the d documents and the estimated $\text{Conv}(\hat{\mathbf{C}}_n)$ (represented by the red triangle).



(a) $n = 60$ (b) $n = 600$ (c) $n = 6000$ (d) $n = 60000$

Figure 7: Results of the second experiment in Section 5.1.3. The likelihood density $f_n(\mathbf{u}; \mathbf{x}^{(i)})$ over Δ^2 for different n. The colored circles represent the values of $f_n(\mathbf{u}; \mathbf{x}^{(i)})$: the darker the color is, the higher the likelihood is. The black triangle is $\text{Conv}(\mathbf{C}^0)$; the dark red triangle is $\text{Conv}(\hat{\mathbf{C}}_n)$ produced by MCMC-EM. The red dots are the true means $\mathbf{u}^{0(i)}$, and the black dots are the sample means $\hat{\mathbf{u}}^{(i)}$.

We observe that $\text{Conv}(\hat{\mathbf{C}}_n)$ tends to cover these density balls while maintaining its volume small. Recall that $\hat{\mathbf{C}}_n = \arg\max_{\mathbf{C}} \prod_{i=1}^{d} \int_{\text{Conv}(\mathbf{C})} \frac{f_n(\mathbf{x}^{(i)}|\mathbf{u})}{|\text{Conv}(\mathbf{C})|} d\mathbf{u}$. $\text{Conv}(\hat{\mathbf{C}}_n)$ can be considered to be the convex polytope that has the highest value of the averaged likelihood density, as well as the smallest convex polytope containing the sample means $\hat{\mathbf{u}}^{(i)}$. Therefore, $\text{Conv}(\hat{\mathbf{C}}_n)$ tends to trade off its volume for a larger coverage of the density balls. In this case, the true means $\mathbf{u}^{0(i)}$ are all located on the boundary of $\text{Conv}(\mathbf{C}^0)$; to fulfill the SS condition, a fraction of each circle thus lies outside $\text{Conv}(\mathbf{C}^0)$.

Consequently, the averaged likelihood density over $\mathrm{Conv}(\hat{\mathbf{C}}_n)$ is larger than that of $\mathrm{Conv}(\mathbf{C}^0)$, though $|\mathrm{Conv}(\hat{\mathbf{C}}_n)| > |\mathrm{Conv}(\mathbf{C}^0)|$. As proved in Theorem 4, the convergence rate of $\hat{\mathbf{C}}_n$, in the order of $\sqrt{\log(n \vee d)/n}$, is slightly slower than that of $\hat{\mathbf{u}}^{(i)}$, which is in the order of $\sqrt{1/n}$.

5.2 Real Applications

We next apply our algorithm to some real-world datasets. In Section 5.2.1 we compare the quantitative performance of our algorithms, and of several baseline methods, on two text datasets: an NIPS dataset that contains long academic documents, and the Daily Kos dataset that contains short news documents. In Section 5.2.2 we analyze a taxi-trip dataset that contains New York City (NYC) taxi trip records, including pick-up and drop-off locations.

5.2.1 Text Data sets

The NIPS dataset[2] contains $V = 11463$ unique words and $d = 5811$ NIPS conference papers, with an average document length of 1902 words. The Daily Kos dataset[3] contains $V = 6906$ unique words and $d = 3430$ Daily Kos blog entries, with an average document length of 136 words. As the two datasets are formatted in document-term matrices without stop words or rarely occurring words, we do not apply any pre-processing procedures.

We compare the performance of our algorithm (MC^2-EM) with the following baseline algorithms: Anchor Free (AnchorF) (Huang et al., 2016), Geometric Dirichlet Means (GDM) (Yurochkin and Nguyen, 2016), and two MCMC algorithms—one based on Gibbs sampler (Gibbs) (Griffiths and Steyvers, 2004), and the other based on a partially collapsed Gibbs sampler (pcLDA) (Magnusson et al., 2018; Terenin et al., 2018). The hyper-parameters of the baselines are set as their default, except that the prior of the mixing weights in Gibbs and pcLDA is set as uniform as ours. For our algorithm, the number of MCMC samples is 100 without burn-in; the stopping criterion is that the relative change of likelihood goes below 10^{-9} or that 200 EM iterations are completed, whichever comes first.

To evaluate the results, we employ the following three metrics. *Topic Coherence* is used to measure the single-topic quality, defined as $\sum_{l=1}^{k} \sum_{v_1,v_2 \in \mathcal{V}_l} \log\left({}^{\mathrm{freq}(v_1,v_2)+\epsilon}/_{\mathrm{freq}(v_2)}\right)$, where $\mathcal{V}_l$ is the leading 20 words for topic l, $\mathrm{freq}(\cdot)$ is the occurrence count, and ϵ is a small constant added to avoid numerical issues. *Similarity Count* is used to measure similarity between topics (Arora et al., 2013; Huang et al., 2016); it is obtained simply by adding up the overlapped words across $\mathcal{V}_l$. *Perplexity Score* is used to measure

[2] https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

[3] https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/

goodness of fit, which is the multiplicative inverse of the likelihood, normalized by the number of words. For the first metric, the larger the better; for the latter two, the smaller the better. (Detailed definition of these three metrics can be found in Appendix F.)

In practice, the number of topics k is unknown. We propose a procedure to select k based on the *effective rank* of the sample document-term matrix $\hat{\mathbf{U}}$. Since the topic matrix $\mathbf{C}$ is assumed to have full rank (Theorem 2), the true term-document matrix $\mathbf{U}$ has rank k. By Weyl's inequality (Weyl, 1912), the singular values of $\hat{\mathbf{U}}$ are expected to be close to those of $\mathbf{U}$. Therefore we can plot the ordered singular values of $\hat{\mathbf{U}}$ versus its index, and then select k by detecting the location of a significant drop of the curve. See Appendix F for a simulation illustrating this approach.

Table 2: Experiment results on the NIPS and the Daily Kos Datasets.

	NIPS					Daily Kos				
	AnchorF	GDM	Gibbs	pcLDA	MC^2-EM	AnchorF	GDM	Gibbs	pcLDA	MC^2-EM
Topic Coherence										
$k = 5$	-904	-501	-365	-355	**-342**	-699	**-643**	-752	-709	-723
$k = 10$	-1954	-1083	-960	**-942**	-975	-1659	**-1551**	-1708	-1609	-1614
$k = 15$	-2935	-1770	-1648	-1599	**-1573**	-2727	**-2307**	-2465	-2380	-2411
$k = 20$	-3664	-2409	-2314	-2373	**-2254**	-3942	**-3182**	-3840	-3115	-3299
Similarity Counts										
$k = 5$	24	**10**	25	26	24	24	**14**	23	25	25
$k = 10$	69	**44**	63	67	63	85	**55**	**55**	66	57
$k = 15$	102	**98**	99	99	102	151	111	**78**	103	90
$k = 20$	154	161	**134**	155	147	224	175	**116**	153	143
Perplexity Score										
$k = 5$	4431	2955	2256	2183	**2182**	2252	2252	1755	1758	**1724**
$k = 10$	4317	2479	2067	**1973**	**1973**	2124	2004	1546	1532	**1507**
$k = 15$	4176	2273	1975	**1870**	1874	2061	1912	1452	1438	**1404**
$k = 20$	3877	2166	1918	1801	**1800**	2012	1791	1405	1384	**1342**

The results are summarized in Table 2 and Table 3, where $k = 5$ and $k = 7$, respectively, are the recommended number of topics for NIPS and Daily Kos dataset, chosen by the procedure mentioned above (the singular values plots can be found in Appendix F). The best score in each case is highlighted in boldface. Overall, our estimator (MC^2-EM) gives promising results. For all three metrics in both

Table 3: Results on the Daily Kos dataset based on $k = 7$ chosen by the singular values plot.

	Daily Kos ($k = 7$)				
	AnchorF	GDM	Gibbs	pcLDA	MC^2-EM
Topic Coherence	**-998**	-1007	-1095	-1090	-1053
Similarity Counts	47	**36**	40	40	40
Perplexity Score	2190	2147	1649	1643	**1607**

datasets, it gives the highest score or a score close to the highest. For topic coherence, it is the best for $k = 5, 15$, and 20 in NIPS. For similarity counts, it performs similarly to Gibbs and pcLDA in both datasets, and in Daily Kos largely outperforms AnchorF and GDM for $k = 10, 15$, and 20. For perplexity score, it is consistently the best in Daily Kos, and in NIPS except for $k = 15$; its scores are very close to the best one given by pcLDA.

The leading 10 topic words given by MC^2-EM can be found in the supplementary material.

5.2.2 New York Taxi-trip Dataset

Reinforcement learning algorithms have been widely used in solving real-world Markov decision problems. Use of a compact representation of the underlying states, known as state aggregation, is crucial for those algorithms to scale with large datasets. As shown below, learning a soft state aggregation (Singh et al., 1995) is equivalent to estimating a topic model.

We say that a Markov chain $X_0, X_1, \cdots, X_T$ admits a *soft state aggregation* with k meta-states, if there exist random variables $Z_0, Z_1, \cdots, Z_{n-1} \in \{1, \cdots, k\}$ such that

$$\mathbb{P}(X_{t+1}|X_t) = \sum_{l=1}^{k} \mathbb{P}(Z_t = l|X_t) \cdot \mathbb{P}(X_{t+1}|Z_t = l), \tag{14}$$

for all t with probability 1 (Singh et al., 1995). Here, $\mathbb{P}(Z_t = l|X_t)$ and $\mathbb{P}(X_{t+1}|Z_t = l)$ are independent of t and are referred to as the *aggregation distributions* and *disaggregation distributions*. Let $\mathbf{U} \in \mathbb{R}^{V\times V}$ denote the transition matrix with $U_{ji} = \mathbb{P}(X_{t+1} = j|X_t = i)$. Let $\mathbf{C} \in \mathbb{R}^{V\times k}$ and $\mathbf{W} \in \mathbb{R}^{k\times V}$ denote the disaggregation and aggregation distribution matrices, respectively, with $C_{jl} = \mathbb{P}(X_{t+1} = j|Z_t = l)$ and $W_{li} = \mathbb{P}(Z_t = l|X_t = i)$. Then (14) can be written as $\mathbf{U} = \mathbf{CW}$, the same as the matrix form for topic modelling.

In this section, we consider a New York taxi-trip[4] dataset. This dataset contains $\sum_{i=1}^{d} n_i = 7,667,792$ New York City yellow cab trips in January 2019. The location information is discretized into $V = 263$

[4] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

taxi zones with 69 in Manhattan, 69 in Queens, 61 in Brooklyn, 43 in Bronx, 20 in Staten Island, and 1 in EWR. For each trip, we are given its pick-up and drop-off zones. On the left of Figure 8, we plot 30 example trips from the data. Following a similar analysis of this dataset from Duan et al. (2019), we aim to merge the $V = 263$ taxi zones into meta-states via soft state aggregation.

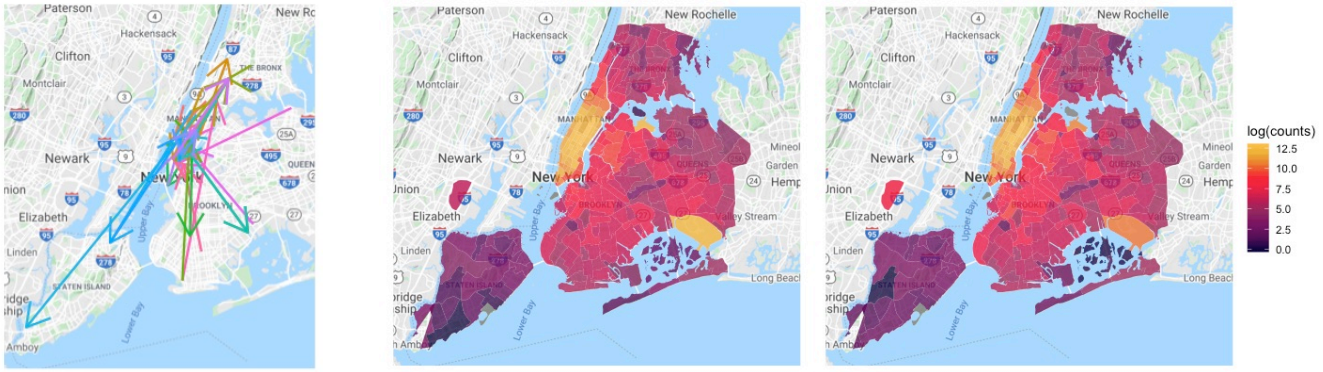


Figure 8: NYC taxi-trip data glance. Left: 30 example trips with arrows pointing from pick-up zones to drop-off zones. Middle: the pick-up distribution. Right: the drop-off distribution.

In the middle and the right of Figure 8, we use heat maps to visualize the distributions of the trip counts for pick-up and drop-off over $V = 263$ zones. Most of the traffic concentrates in midtown and downtown Manhattan, as well as at the JFK airport on the southeast side of Queens, for both pick-up and drop-off.

At the top of Figure 9 we plot the estimation results for the drop-off distributions conditioned on the meta-state, $\mathbb{P}(X_{t+1}|Z_t = l)$. We observe that the drop-off traffic is decomposed into three clusters, (1) downtown Manhattan, (2) west midtown Manhattan, and (3) east midtown Manhattan, for each of the three meta states (topics); this implies that people dropped off in downtown Manhattan may come from the first meta state, and that people dropped off in midtown east and west may come from the second and the third meta states, respectively. The JFK airport has a relatively high probability mass in all three states but is not on the top list for any of them, which implies that people arriving at JFK may come from anywhere in NYC.

At the bottom of Figure 9 we plot the conditional probability over the meta-state (topics), given the pick-up zone, $\mathbb{P}(Z_t = l|X_t)$. The three meta states consist of (1) Staten Island, Brooklyn, Queens, and downtown Manhattan; (2) uptown Manhattan and Bronx; and (3) east midtown Manhattan. Note that the scales of the estimates for $\mathbf{C}$ and $\mathbf{W}$ are quite different. In specific, the sum of values over each map of the top three is 1 since $\sum_{v=1}^{V} \mathbb{P}(X_{t+1} = v|Z_t = l) = 1,\ l = 1, 2, 3$, while the sum of values

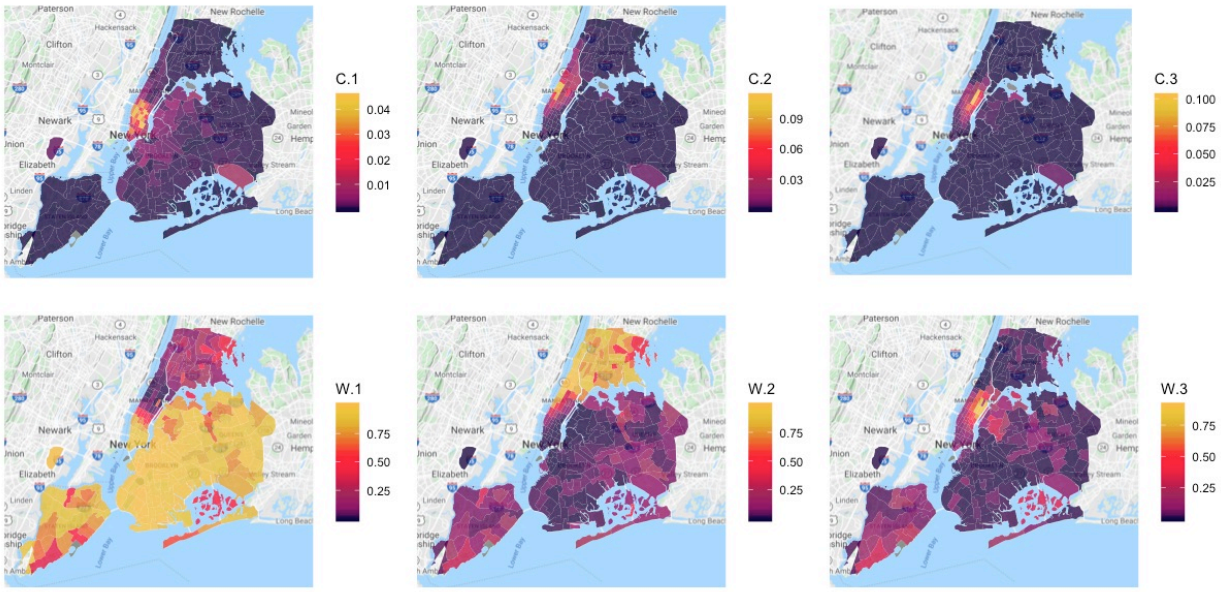


Figure 9: Estimation results for NYC taxi-trip data for $k = 3$. The top three plots represent the estimated disaggregation distributions (topic vectors) $\hat{\mathbf{C}}_1, \hat{\mathbf{C}}_2, \hat{\mathbf{C}}_3 \in \mathbb{R}^V$, where $\hat{\mathbf{C}}_l = \mathbb{P}(X_{t+1}|Z_t = l)$. The bottom three plots represent the estimated aggregation distributions $\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \hat{\mathbf{W}}_3 \in \mathbb{R}^V$, where $\hat{\mathbf{W}}_l = \mathbb{P}(Z_t = l|X_t)$.

for each zone over the bottom three maps is 1 since $\sum_{l=1}^{3} \mathbb{P}(Z_t = l|X_t = v) = 1,\ v = 1, \cdots, V$. The interpretation of, say, the second meta state, is that the destinations of trips starting from uptown Manhattan and Bronx are likely to be in midtown Manhattan. We observe that the pick-up and the drop-off locations in the same meta state are generally close regionally; this result is reasonable, as people tend to take a taxi for short trips, preferring less expensive public transportation for longer trips.

The estimated disaggregation and aggregation distributions plots for $k = 9$ can be found in the supplementary material. They reveal that the traffic in the first eight meta states is within Manhattan, which is the most heavy-traffic place in NYC, and that the partition is more fine-grained compared with the results for $k = 3$. Similar to the results for $k = 3$, the pick-up and drop-off locations for each meta state are regionally close at this time. It is interesting that such a strong regional relationship emerges, since the data fed into our algorithm do not contain any regional information.

6 Discussion

In this paper, we introduce a new set of geometric conditions for topic model identifiability under

volume minimization, a weaker set than the commonly used separability conditions. For computation, we propose a maximum likelihood estimator of the latent topics matrix, based on an integrated likelihood. Our approach implicitly promotes volume minimization. We conduct finite-sample error analysis for the estimator and discuss the connection of our results to existing ones. Experiments on simulated and real datasets demonstrate the strength of our method. Our work makes an important contribution to the general theory of estimation of latent structures arising for topic models. Some interesting future work might include: (1) exploring a sufficient and necessary condition for model identifiability, as the SS condition is not necessary; (2) providing explicit verifiable sufficient conditions for the (α, β)-SS condition — we conjecture that the (α, β)-SS condition can be implied by the SS condition; (3) establishing the minimax rate of convergence of topic matrix estimation, and verifying whether the proposed estimator is (nearly) optimal. Although presented in the context of topic models, results from our work are immediately applicable to a wide range of mixed membership models arising from various machine learning applications. In addition, we may incorporate additional low-dimensional structures into the model, such as (group) sparsity, to enhance the estimation accuracy.

References

Anandkumar, A., D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu (2012). A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research 15*(1), 2773–2832.

Anandkumar, A., D. Hsu, and S. M. Kakade (2012). A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pp. 33–1. JMLR Workshop and Conference Proceedings.

Arora, S., R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pp. 280–288.

Arora, S., R. Ge, and A. Moitra (2012). Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 1–10.

Azar, Y., A. Fiat, A. Karlin, F. McSherry, and J. Saia (2001). Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pp. 619–626.

Bansal, T., C. Bhattacharyya, and R. Kannan (2014). A provable svd-based algorithm for learning topics in dominant admixture corpus. *Advances in Neural Information Processing Systems 27*, 1997–2005.

Berger, J. O., B. Liseo, and R. L. Wolpert (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical science 14*(1), 1–28.

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

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge University Press.

Brunel, V.-E., J. M. Klusowski, and D. Yang (2021). Estimation of convex supports from noisy measurements. *Bernoulli 27*(2), 772–793.

Chen, Y. M., X. S. Chen, and W. Li (2016). On perturbation bounds for orthogonal projections. *Numerical Algorithms 73*(2), 433–444.

Craig, M. D. (1994). Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing 32*(3), 542–552.

Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis 7*(1), 1–46.

Devroye, L. et al. (1983). The equivalence of weak, strong and complete convergence in l_1 for kernel density estimates. *The Annals of Statistics 11*(3), 896–904.

Donoho, D. and V. Stodden (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, pp. 1141–1148.

Duan, Y., T. Ke, and M. Wang (2019). State aggregation learning from markov transition data. In *Advances in Neural Information Processing Systems*, pp. 4488–4497.

Freund, R. M. and J. B. Orlin (1985). On the complexity of four polyhedral set containment problems. *Mathematical programming 33*(2), 139–145.

Fu, X., W.-K. Ma, K. Huang, and N. D. Sidiropoulos (2015). Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing 63*(9), 2306–2320.

Ge, R. and J. Zou (2015). Intersecting faces: Non-negative matrix factorization with new guarantees. In *International Conference on Machine Learning*, pp. 2295–2303.

Goldenshluger, A. and A. Tsybakov (2004). Estimating the endpoint of a distribution in the presence of additive observation errors. *Statistics & probability letters 68*(1), 39–49.

Götze, F., H. Sambale, and A. Sinulis (2019). Higher order concentration for functions of weakly dependent random variables. *Electronic Journal of Probability 24*, 1–19.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(suppl 1), 5228–5235.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, pp. 289–296.

Huang, K., X. Fu, and N. D. Sidiropoulos (2016). Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pp. 1786–1794.

Huang, K., N. D. Sidiropoulos, and A. Swami (2014). Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing 62*(1), 211–224.

Jang, B. and A. Hero (2019). Minimum volume topic modeling. In *International Conference on Artificial Intelligence and Statistics*, pp. 3013–3021.

Javadi, H. and A. Montanari (2020). Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association 115*(530), 896–907.

Ke, Z. T. and M. Wang (2017). A new svd approach to optimal topic estimation. *arXiv preprint arXiv:1704.07016*.

Kleinberg, J. and M. Sandler (2003). Convergent algorithms for collaborative filtering. In *Proceedings of the 4th ACM conference on Electronic commerce*, pp. 1–10.

Kleinberg, J. and M. Sandler (2008). Using mixture models for collaborative filtering. *Journal of Computer and System Sciences 74*(1), 49–69.

Magnusson, M., L. Jonsson, M. Villani, and D. Broman (2018). Sparse partially collapsed mcmc for parallel inference in topic models. *Journal of Computational and Graphical Statistics 27*(2), 449–463.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE.

Miao, L. and H. Qi (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing 45*(3), 765–777.

Nascimento, J. M. and J. M. Dias (2005). Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing 43*(4), 898–910.

Nguyen, X. (2015). Posterior contraction of the population polytope in finite admixture models. *Bernoulli 21*(1), 618–646.

Nielsen, S. F. et al. (2000). The stochastic em algorithm: Estimation and asymptotic results. *Bernoulli 6*(3), 457–489.

Papadimitriou, C. H., P. Raghavan, H. Tamaki, and S. Vempala (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences 61*(2), 217–235.

Perrone, V., P. A. Jenkins, D. Spano, and Y. W. Teh (2016). Poisson random fields for dynamic feature models. *arXiv preprint arXiv:1611.07460*.

Recht, B., C. Re, J. Tropp, and V. Bittorf (2012). Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pp. 1214–1222.

Singh, S. P., T. Jaakkola, and M. I. Jordan (1995). Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, 361–368.

Tang, J., Z. Meng, X. Nguyen, Q. Mei, and M. Zhang (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning*, pp. 190–198.

Terenin, A., M. Magnusson, L. Jonsson, and D. Draper (2018). Polya urn latent dirichlet allocation: A doubly sparse massively parallel sampler. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*(7), 1709–1719.

Wang, Y. (2019). Convergence rates of latent topic models under relaxed identifiability conditions. *Electronic Journal of Statistics 13*(1), 37–66.

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen 71*(4), 441–479.

Winter, M. E. (1999). N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Imaging Spectrometry V*, Volume 3753, pp. 266–275. International Society for Optics and Photonics.

Yurochkin, M. and X. Nguyen (2016). Geometric Dirichlet means algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513.

Supplementary Material: Learning Topic Models: Identifiability and Finite-Sample Analysis

The supplementary material is organized as follows.

- Section A: Discussion on identifiability related to Remark 2.1.
- Section B: Error analysis and consistency under stochastic mixing weights.
- Section C: Derivation of the MCMC-EM algorithm.
- Section D: Proofs of main theorems.
- Section E: Proofs of technical lemmas and propositions.
- Section F: Additional simulations and experiments.
- Sections G & H: Top 10 words of the latent topics returned by our algorithm for the two real applications.
- Section I: Mined meta states for the taxi-trip dataset.

A Discussion on Identifiability Related to Remark 2.1

Javadi and Montanari (2020) and we both follow the same principle to address the non-identifiability issue — among all equivalent parameters that lead to the same statistical model, the one that minimizes a chosen criterion function is used to represent the equivalence class (therefore the most parsimonious representation). However, the adopted criterion functions are different: ours is the volume of Conv($\mathbf{C}$), while theirs is the sum of distances from the vertices of Conv($\mathbf{C}$) (i.e., columns of $\mathbf{C}$) to the convex hull of $\mathbf{U}$. The criterion function adopted Javadi and Montanari (2020) is easier to be formulated into a statistical estimator that minimizes an empirical evaluation of it. However, as we discussed in Section 1.2, our criterion function as the volume of a low-dimensional polytope in a high-dimensional space does not take a simple form, which greatly complicates the estimator construction. Fortunately, we find that maximizing a particular integrated likelihood leads to an estimator that implicitly minimizes the volume.

Regarding the two notions of identifiability, minimizers of the two criterion functions are usually different — except for some special cases, such as when the pure topic documents condition hold so that vertices of Conv($\mathbf{C}$) are data points. Therefore, the two notions of identifiability are not directly comparable. Figure S1 helps to illustrate this point. In Figure S1, the grey region is Conv($\mathbf{U}$) and the black triangle ABC is the unique volume minimizer among all three-vertex convex polytopes enclosing Conv($\mathbf{U}$). However, triangle ABC is not the minimizer of the criterion function in Javadi and Montanari (2020) with the Euclidean distance as the distance function: it is easy to verify that when the ratio of the height to the base of triangle ABC is larger than 6, the red triangle FGH has a smaller summation of distances to the gray region than ABC (see the caption that describes how we construct the red triangle FGH).

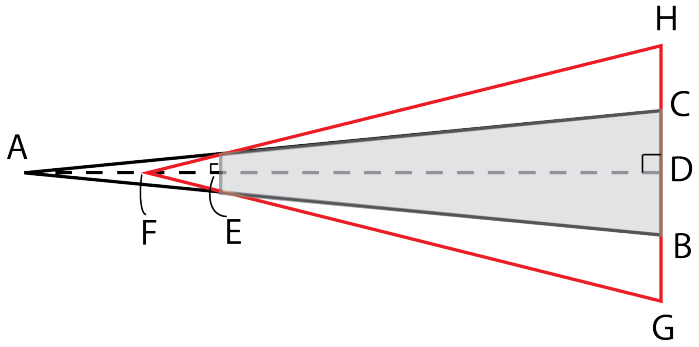


Figure S1: An example ($V = k = 3$), in which both ABC and FGH are isosceles triangles enclosing Conv($\mathbf{U}$) (grey region). In addition, $BC = b$, $AD = h$, $AE = h/4$, and F is the midpoint of AE.

Under the principle of using the minimizer to represent the whole equivalence class, a trivial identifiability condition is to assume the uniqueness of the minimizer, which is exactly the identifiability condition given in Javadi and Montanari (2020). The drawback, however, is that it is often not trivial, if not impossible, to check whether the minimizer of a criterion function is unique. In Javadi and Montanari (2020), uniqueness is checked only for a simple case when the vertices of Conv($\mathbf{C}$) are data points (their Remark 3.1). In contrast, our identifiability condition, the SS condition, is a set of explicit, verifiable conditions. Consider the example given in Figure S1. By our Theorem 2, the model is identifiable with respect to our volume minimization. But, it is difficult to verify whether the model is identifiable in Javadi and Montanari (2020): We do not know whether the triangle FGH, although shown to be a better choice than triangle ABC, indeed minimizes the criterion function; even if it does, we do not know whether it is unique.

In summary, neither definition of identifiability is more general than the other. Since the identification condition in Javadi and Montanari (2020) is difficult to check, we are not able to provide an example where the model is identifiable under one notion but not under the other. Due to the same reason, it is unclear whether our SS condition implies their definition of identifiability. Although the two notions of identifiability are not comparable, we would like to highlight that an advantage of our volume minimization criterion is that it helps to justify the empirical success of the Latent Dirichlet Allocation (LDA) model, because the proposed estimator is essentially the maximum likelihood estimator of $\mathbf{C}$ from the LDA model with the prior of $\mathbf{W}$ being the uniform distribution. LDA models with general priors can be interpreted as maximizing the data likelihood while minimizing a weighted volume where a non-uniform volume element is integrated over the convex hull of $\mathbf{C}$ when defining the volume.

B Error Analysis and Consistency under Stochastic Mixing Weights

In this appendix, we explore cases in which $\mathbf{w}_1^0, \cdots \mathbf{w}_d^0$ are random i.i.d. samples from some unknown distribution $\mathcal{P}$ over Δ^{k-1} (the theoretical result in the main manuscript considers the fixed mixing weights setting). In such cases, we will apply Theorem 4 to this set of stochastic mixing weights by showing that under a suitable set of conditions to be described below, Assumptions (A1)-(A3) hold with high probability.

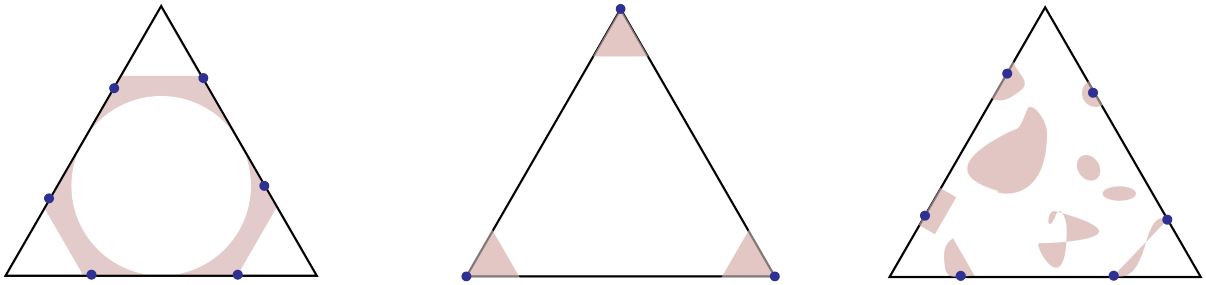


Figure S2: Examples of (α, β)-SS distributions for $k = 3$: $\mathbf{w}^\sharp$'s (blue dots) from $supp(\mathcal{P})$ (pink area) are (α, β)-SS on Δ^{k-1}(the triangle).

Formally, we introduce a "stochastic" version of the SS condition on $\mathcal{P}$, called (α, β)-*SS distribution*, to ensure the (α, β)-SS condition to hold for $\mathbf{W}$ with high probability as long as the number of documents d is sufficiently large.

Definition 4 ((α, β)-SS distribution)**.** *A distribution $\mathcal{P}$ is an* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$**-SS distribution**, *if there exist s distinct points in its support, $\mathbf{w}_1^\sharp, \cdots, \mathbf{w}_s^\sharp \in supp(\mathcal{P})$, and some positive constants r_0, c_0, such that $\mathbf{W}^\sharp = \{\mathbf{w}_i^\sharp\}_{i=1}^s$ is (α, β)-SS, and for each $i \in [s]$,*

$$\mathcal{P}(\|\mathbf{w} - \mathbf{w}_i^\sharp\|_2 \leqslant r) \geqslant (k-1)! \cdot c_0 \cdot r^{k-1}, \quad \forall\, 0 < r \leqslant r_0.$$

The condition in Definition 4 is mild and can be satisfied by many commonly encountered distributions over the simplex Δ^{k-1}. For example, any distribution whose density function does not vanish on Δ^{k-1}, such as the uniform distribution and Dirichlet distributions, is $(\epsilon, C\epsilon)$-SS for any sufficiently small $\epsilon > 0$, where C is some constant depends on the distribution. In addition, an (α, β)-SS distribution does not need to have a full support over Δ^{k-1}—as long as a distribution has positive density values around a set of (α, β)-SS points, then it is (α, β)-SS. See Figure S2 for some examples of SS distributions whose supports are sparsely scattered over the simplex.

Next, we state our assumption on the true underlying distribution $\mathcal{P}^0$ that generates the stochastic mixing weights.

Assumptions. *Assume the following:*

(A5) $\mathbf{w}_1^0, \cdots, \mathbf{w}_d^0$ *are i.i.d. random samples from an* (α, β)*-SS distribution* $\mathcal{P}^0$*, with* $\alpha \geqslant C_1' \sqrt{\frac{\log(n \vee d)}{n}} + \left(\frac{\log d}{d}\right)^{\frac{1}{k-1}}$*, where* C_1' *is a constant.*

The following Theorem 8 establishes the finite-sample error bound when $\mathbf{W}$ is stochastically generated.

Theorem 8. *Under Assumptions (A1), (A2) and (A5), it holds with probability at least* $1 - D_1' s/d - D_2' d/(n \vee d)^c$ *that*

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant D_3' \sqrt{\frac{\log(n \vee d)}{n}} + D_4' \beta, \tag{B.1}$$

where $c, D_1', D_2', D_3', D_4'$ *are positive constants. In particular, if* $\beta \leqslant C_2' \left(\sqrt{\log(n \vee d)/n} + (\log d/d)^{1/(k-1)}\right)$ *for some constant* C_2'*, then*

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant D_3'' \sqrt{\frac{\log(n \vee d)}{n}} + D_4'' \left(\frac{\log d}{d}\right)^{\frac{1}{k-1}}. \tag{B.2}$$

Similar to the remark of Assumption (A3), in most cases the parameter β can be chosen as the same order as α in the (α, β)-SS condition in Theorem 8. For example, according to Proposition 5, if the support of $\mathcal{P}^0$ contains the point $(1 - x_{ij})\mathbf{e}_i + x_{ij}\mathbf{e}_j$, where $0 \leqslant x_{ij} < 1/k$, for all $1 \leqslant i \neq j \leqslant k$, and $\mathcal{P}^0$ has positive density values around these points, then $\mathcal{P}$ is $(\epsilon, C\epsilon)$-SS for all $\epsilon > 0$.

It is important to emphasize that our method does not require any prior knowledge about the distribution $\mathcal{P}^0$ (albeit our theory requires it to be SS). In comparison, in most Bayesian latent variable mixture model literature such as Tang et al. (2014), Nguyen (2015) and Wang (2019), $\mathcal{P}^0$ is assumed to be known and have a full support over the simplex Δ^{k-1}.

Similar to Theorem 7, we provide conditions for the estimator $\hat{\mathbf{C}}_n$ to have the estimation consistency under the double asymptotic setting by letting $(n, d) \to \infty$ in a suitable manner in Theorem 8.

Assumptions. *Assume the following:*

(A5') *For all sufficiently small* $\epsilon > 0$*, there exist some* $\beta_\epsilon > 0$*, such that* $\beta_\epsilon \to 0$ *as* $\epsilon \to 0$*, and* $\mathbf{w}_1^0, \cdots, \mathbf{w}_d^0$ *are i.i.d. random samples from distribution* $\mathcal{P}^0$ *that is* $(\epsilon, \beta_\epsilon)$*-SS.*

Theorem 9 (Estimation consistency)**.** *Under Assumptions (A1), (A2), (A4) and (A5'), we have*

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \to 0 \quad \textit{in probability as } (n, d) \to \infty.$$

C Derivation of the MCMC-EM Algorithm

We use an MCMC-EM algorithm to compute the MLE of the integrated likelihood function (4). First we introduce a set of latent variables $\mathbf{Z} = \{Z_{ij}\}$, where Z_{ij} is the topic label for $x_{i,j}$. Then express the LDA model as follows:

$$\begin{aligned} x_{i,j}|\mathbf{C}, Z_{ij} = l &\sim \text{Multi}_V(\mathbf{C}_l) \\ Z_{ij}|\mathbf{w}_i &\sim \text{Multi}_k(\mathbf{w}_i) \\ \mathbf{w}_i|\boldsymbol{\beta}_0 &\sim \text{Dir}_k(\boldsymbol{\beta}_0), \end{aligned}$$

where

$$i = 1, \ldots, d; \quad j = 1, \ldots, n; \quad l = 1, \ldots, k.$$

We fix $\boldsymbol{\beta}_0 = \mathbf{1}_k$ throughout, since we consider a uniform "prior" on $\mathbf{W}$. The integrated likelihood (4) can be written as

$$\begin{aligned} F_{n\times d}(\mathbf{C}; \mathbf{X}) = p(\mathbf{X} \mid \mathbf{C}) &= \int p(\mathbf{X}, \mathbf{Z} \mid \mathbf{C}) d\mathbf{Z} \\ &\propto \prod_{i=1}^{d} \int \left[\int \prod_{j=1}^{n} p(x_{i,j}|\mathbf{C}, Z_{ij}) p(Z_{ij}|\mathbf{w}) p(\mathbf{w}|\boldsymbol{\beta}_0) d\mathbf{w} \right] d\mathbf{Z}_{i\cdot} \\ &\propto \prod_{i=1}^{d} \int \prod_{j=1}^{n} p(x_{i,j}|\mathbf{C}, Z_{ij}) p(\mathbf{Z}_{i\cdot} = \mathbf{z}|\boldsymbol{\beta}_0) d\mathbf{z} \end{aligned}$$

where $\mathbf{Z}_{i\cdot} = (Z_{i1}, \cdots, Z_{in})$.

E-step Define $Q(\mathbf{C}|\mathbf{C}^{(0)})$ as the expected value of the log likelihood function of $\mathbf{C}$, with respect to $\mathbf{Z}$ given $\mathbf{X}$ and $\mathbf{C}^{(0)}$, where $\mathbf{C}^{(0)}$ is the estimated topic matrix obtained from the last EM iteration.

$$\begin{aligned} Q(\mathbf{C}|\mathbf{C}^{(0)}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{C}^{(0)}} \log[F_{n\times d}(\mathbf{C}; \mathbf{X}, \mathbf{Z})] \\ &= \mathbb{E}_{\mathbf{Z}|\mathbf{C}^{(0)}} \sum_{i=1}^{d} \sum_{j=1}^{n} \log p(x_{i,j}|\mathbf{C}, Z_{ij}) + Const \end{aligned}$$

We ignore the constant term in the following derivation. Since the marginal probability $p(\mathbf{Z}_{i\cdot} = \mathbf{z}|\boldsymbol{\beta}_0)$ is infeasible, we apply MCMC to and iteratively sample $\mathbf{Z} = \{Z_{ij}\}_{i,j}$ and $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^{d}$ as follows:

$$\begin{aligned} Z_{ij}|\mathbf{C}, x_{i,j} = v &\sim \text{Multi}_k \left(\frac{c_{vl} w_{li}}{\sum_{l=1}^{k} c_{vl} w_{li}} \right)_{l=1,\ldots,k} \\ \mathbf{w}_i|\mathbf{Z}_{i\cdot} &\sim \text{Dir}_k \left(\beta_{0l} + \sum_{j=1}^{n} \mathbb{1}(Z_{ij} = l) \right)_{l=1,\ldots,k}. \end{aligned}$$

We approximate $Q(\mathbf{C}|\mathbf{C}^{(0)})$ function by the samples of $\mathbf{Z}$,

$$
\begin{aligned}
Q(\mathbf{C}|\mathbf{C}^{(0)}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{C}^{(0)}} \sum_{v=1}^{V}\sum_{l=1}^{k} \log c_{vl} \left[\sum_{i=1}^{d}\sum_{j=1}^{n} \mathbb{1}(Z_{ij}=l, x_{i,j}=v)\right] \\
&\approx \frac{1}{T}\sum_{t=b+1}^{b+T}\sum_{v=1}^{V}\sum_{l=1}^{k}\left[\log c_{vl} \sum_{i=1}^{d}\sum_{j=1}^{n} \mathbb{1}(Z_{ij}^{(t)}=l, x_{i,j}=v)\right].
\end{aligned}
$$

where c_{vl} is the (v,l)-th element of $\mathbf{C}$. Here the $Z_{ij}^{(t)}$ denotes the sample of Z_{ij} at t-th MCMC iteration, b denotes the burn-in period and T denotes the number of the samples after burn-in.

M-step We maximize the approximated $Q(\mathbf{C}|\mathbf{C}^{(0)})$ with respect to $\mathbf{C}$ by the following closed-form solution:

$$
c_{vl} = \frac{\sum_{i,j,t} \mathbb{1}(Z_{ij}^{(t)}=l, x_{i,j}=v)}{\sum_{i,j,t} \mathbb{1}(Z_{ij}^{(t)}=l)}.
$$

The algorithm of the E-step is given in Algorithm 1. Here we use, $\mathscr{Z}, \mathscr{Z} \in \mathbb{R}^{d\times V\times k}$, to denote the counts of the samples of $\mathbf{Z}$. Specifically, $\mathscr{Z}[i,v,l] = \sum_j \mathbb{1}(Z_{ij}^{(t)}=l, x_{i,j}=v)$ is the count of $\mathbf{Z}$ at t-th MCMC iteration, and $\mathcal{Z}[i,v,l] = \sum_{t=b+1}^{b+T}\sum_j \mathbb{1}(Z_{ij}^{(t)}=l, x_{i,j}=v)$ is the sum of count of $\mathbf{Z}$ over T iterations.

Algorithm 1: The E-step of the MCMC-EM Algorithm

Input: $\mathbf{C}$;

$\mathcal{Z}[:,:,:] \leftarrow \mathbf{0}_{d\times V\times k}$; ▷ Initialize $\mathcal{Z}$

$\mathbf{W}[:,i] \leftarrow \mathrm{Dir}_k(\mathbf{1})$, $i=1,\cdots,d$; ▷ $\mathbf{W}[:,i]$ is the i-th column of $\mathbf{W}$

for $t = 1,\cdots,b,b+1,\cdots,b+T$ **do**

  $\mathscr{Z}[:,:,:] \leftarrow \mathbf{0}_{d\times V\times k}$; ▷ Initialize $\mathscr{Z}$

  for $i = 1,\cdots,d$ **do**

    for $v = 1,\cdots,V$ **do**

      $\boldsymbol{p} \leftarrow \mathbf{C}[v,:] \odot \mathbf{W}[:,i]$ ▷ $\mathbf{C}[v,:]$ is the v-th row of $\mathbf{C}$

      ▷ $\odot$ denotes an element-wise multiplication

      $\boldsymbol{p} \leftarrow \boldsymbol{p}/\sum_{l=1}^{k}\boldsymbol{p}[l]$ ▷ $\boldsymbol{p}[l]$ is the l-th element of $\boldsymbol{p}$

      $\mathscr{Z}[i,v,:] \leftarrow \mathrm{Multi}(n=x_v^{(i)}, p=p)$; ▷ $x_v^{(i)}$ is the count of v-th word in the i-th doc

    $\mathbf{W}[:,i] \leftarrow \mathrm{Dir}_k(\sum_v \mathscr{Z}[i,v,:] + \boldsymbol{\beta}_0)$;

  if $t > b$ **then**

    $\mathcal{Z} \leftarrow \mathcal{Z} + \mathscr{Z}$;

$\mathcal{Z} \leftarrow \frac{1}{T}\mathcal{Z}$;

Output: $\mathcal{Z}$.

Empirically, since $\mathscr{Z}$ and $\mathscr{X}$ are sparse, to save the computation space, we recommend to use two 2-dim arrays instead, namely $\mathscr{C} = \sum_{i=1}^{d} \mathscr{Z}[i,:,:]$ and $\mathscr{W} = \sum_{v=1}^{V} \mathscr{Z}[:,v,:]$, and $\mathscr{C}, \mathscr{W}$ can be used efficiently in updating $\mathbf{C}$ and $\mathbf{W}$, respectively. In addition, the operations in the two nested for-loops over i and v in Algorithm 1 can be paralleled, as they are independent with each other.

The full algorithm is given in Algorithm 2.

Algorithm 2: The MCMC-EM Algorithm

Input: Data $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{d}$; number of topics k;

$\mathbf{C}[l,:] \leftarrow \mathrm{Dir}_V(\mathbf{1})$, $l = 1, \cdots, k$; ▷ Initialize $\mathbf{C}$

repeat

  Obtain $\mathscr{Z}$ using Algorithm 1; ▷ E-step

  $\mathbf{C}[v,l] \leftarrow \sum_{i=1}^{d} \mathscr{Z}[i,v,l] / \sum_{v=1}^{V} \sum_{i=1}^{d} \mathscr{Z}[i,v,l]$,

  $v = 1, \cdots, V$, $l = 1, \cdots k$; ▷ M-step

until *convergence*;

$\mathbf{W}[l,i] \leftarrow \sum_{v=1}^{V} \mathscr{Z}[i,v,l] / \sum_{l=1}^{k} \sum_{v=1}^{V} \mathscr{Z}[i,v,l]$,

$l = 1, \cdots k$, $i = 1, \cdots, d$; ▷ Estimate $\mathbf{W}$

Output: $\mathbf{C}$; $\mathbf{W}$.

D Proofs of Main Theorems

D.0 Notation

For a vector $\mathbf{x}$, we denote by $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ its L_2 norm and $\|\mathbf{x}\|_1 = \sum_i |x_i|$ its L_1 norm. Write $\mathbf{x} \geqslant a$ to indicate that $\mathbf{x}$ is element-wisely no smaller than a. In particular, $\mathbf{1}_k$ denotes the all-ones vector of length k, and $\mathbf{e}_f$ the f-th column of the $k \times k$ identity matrix $\mathbf{I}_k$.

For a matrix $\mathbf{A}_{p\times q}$, $\mathbf{A}(i,:)$ and $\mathbf{A}(:,j)$ are i-th row and j-th column vectors, respectively. We use $\sigma_{\max}(\mathbf{A})$ to denote the square root of the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$, and $\sigma^+_{\min}(\mathbf{A})$ the square root of the smallest nonzero eigenvalue of $\mathbf{A}^T\mathbf{A}$. We denote by $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$ the spectral norm and $\|\mathbf{A}\|_1 = \max_{j=1}^q \sum_{i=1}^p |\mathbf{A}_{ij}|$ the L_1 matrix norm. Some useful facts we will use in the proof: (i) $\sigma_{\max}(\mathbf{AB}) \leqslant \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{B})$; (ii) $\sigma^+_{\min}(\mathbf{AB}) \geqslant \sigma^+_{\min}(\mathbf{A})\sigma^+_{\min}(\mathbf{B})$; (iii) $\|\mathbf{A}\|_2 \leqslant \sqrt{q}\|\mathbf{A}\|_1$; (iv) if $p \geqslant q$ and $\mathbf{A}^T\mathbf{A}$ is invertible, then $\sigma_{\max}((\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T) = 1/\sigma^+_{\min}(\mathbf{A})$.

We denote by $\Delta^{k-1} = \{\mathbf{x} \in \mathbb{R}^k : 0 \leqslant x_i \leqslant 1, \sum_{i=1}^k x_i = 1\}$ the standard $(k-1)$-dimensional simplex. For a matrix $\mathbf{A}_{p\times q}$, let

$$\begin{aligned} \mathrm{Conv}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda} \in \Delta^{q-1}\} \\ \mathrm{cone}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda} \geqslant 0\} \\ \mathrm{aff}(\mathbf{A}) &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x} = \mathbf{A}\boldsymbol{\lambda}, \boldsymbol{\lambda}^T\mathbf{1}_q = 1\} \end{aligned}$$

denote the *convex polytope*, the *simplicial cone* and the *affine space* generated by the q columns of $\mathbf{A}$, respectively.

For any cone $\mathcal{C}$, let $\mathcal{C}^* = \{\mathbf{x} : \mathbf{x}^T\mathbf{y} \geqslant 0, \forall \mathbf{y} \in \mathcal{C}\}$ denote its *dual cone*. In particular, let $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T\mathbf{1}_k\}$. The boundary of $\mathcal{K}$ is denoted by $bd\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = \mathbf{x}^T\mathbf{1}_k\}$, and its dual cone takes the form as $\mathcal{K}^* = \{\mathbf{x} \in \mathbb{R}^k : \mathbf{x}^T\mathbf{1}_k \geqslant \sqrt{k-1}\|\mathbf{x}\|_2\}$. Some useful facts of dual cones from Donoho and Stodden (2004): (i) $cone(\mathbf{A})^* = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^T\mathbf{A} \geqslant 0\}$; (ii) if $\mathcal{A}$ and $\bar{\mathcal{A}}$ are convex cones, and $\mathcal{A} \subseteq \bar{\mathcal{A}}$, then $\bar{\mathcal{A}}^* \subseteq \mathcal{A}^*$.

The true $\mathbf{C}$, $\mathbf{W}$, and $\mathbf{U}$ are denoted by $\mathbf{C}^0$,$\mathbf{W}^0$,$\mathbf{U}^0$, respectively; $\hat{\mathbf{C}}_n$ is the estimator obtained from $F_{n\times d}(\mathbf{C};\mathbf{X})$. $\hat{\mathbf{W}}_n$ is a valid estimator for the mixing matrix in $\mathbb{R}^{k\times d}$ which we will construct in Lemma D.3 such that $\hat{\mathbf{W}}_n \geqslant 0$, $\hat{\mathbf{W}}_n^T\mathbf{1}_k = \mathbf{1}_d$. $\epsilon_n = C_0\sqrt{\frac{\log(n\vee d)}{n}}$ is a small quantity used to measure the convergence rates. Here C_0 in ϵ_n is a positive constant independent of n and d.

Throughout, we use symbols like C, C', C'', C''', C^*, C_i, $C'_i, i = 1, 2, \ldots,$ and D_1, D_2 as generic notations for large absolute numbers, whose exact values may vary from part to part. Unless stated otherwise, these constants are all independent of n and d.

D.1 Proof of Theorem 2

The following lemmas are useful in the proof of Proposition 10. Their proofs are given in Appendix E.

Lemma D.1. *For a full column rank matrix* $\mathbf{C} \in \mathbb{R}^{V \times k}$,

$$|\operatorname{Conv}(\mathbf{C})| = \frac{\sqrt{\det(\mathbf{C}^T\mathbf{C})}}{h \cdot (k-1)!},$$

where h *is the perpendicular distance from the origin to the hyperplane* $\operatorname{aff}(\mathbf{C})$. *In particular, we have*

$$\frac{|\operatorname{Conv}(\mathbf{C})|}{|\operatorname{Conv}(\bar{\mathbf{C}})|} = \frac{\sqrt{\det(\mathbf{C}^T\mathbf{C})}}{\sqrt{\det(\bar{\mathbf{C}}^T\bar{\mathbf{C}})}},$$

if $\operatorname{aff}(\mathbf{C}) = \operatorname{aff}(\bar{\mathbf{C}})$.

Lemma D.2. *If* $\mathbf{W} \in \mathbb{R}^{k \times d}$ *satisfies Condition (S1), then* $\mathbf{W}$ *is of rank* k *(full row rank), and* $\sigma^+_{\min}(\mathbf{W}) \geqslant \frac{1}{k}$.

We first show that Condition (S1) guarantees that $\operatorname{Conv}(\mathbf{C})$ has the minimal volume.

Proposition 10. *If* $\mathbf{W}$ *satisfies Condition (S1) and* $\mathbf{C}$ *is of rank* k *(full column rank), then* $|\operatorname{Conv}(\bar{\mathbf{C}})| \geqslant |\operatorname{Conv}(\mathbf{C})|$ *must hold for any other set of parameters* $(\bar{\mathbf{C}}, \bar{\mathbf{W}})$ *satisfying* $\mathbf{C}\mathbf{W} = \bar{\mathbf{C}}\bar{\mathbf{W}}$.

Proof of Proposition 10. By Lemma D.2, $\mathbf{W}\mathbf{W}^T \in \mathbb{R}^{k \times k}$ is invertible. Define

$$\mathbf{B}_{k \times k} := \bar{\mathbf{W}}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}.$$

Then $\mathbf{C} = \bar{\mathbf{C}}\mathbf{B}$. Note that

$$\mathbf{B}^T\mathbf{1}_k = (\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}\bar{\mathbf{W}}^T\mathbf{1}_k = (\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}\mathbf{1}_d,$$

which is the solution of the least square (LS) problem $\min_{\mathbf{x} \in \mathbb{R}^k} \|\mathbf{1}_d - \mathbf{x}^T\mathbf{W}\|_2$. Since $\|\mathbf{1}_d - \mathbf{1}_k^T\mathbf{W}\|_2 = 0$ achieves the minimum, the unique LS solution is given by $\mathbf{1}_k$, i.e.,

$$\mathbf{B}^T\mathbf{1}_k = \mathbf{1}_k. \tag{D.3}$$

Thus, columns of $\bar{\mathbf{C}}$ are convex combination of columns of $\mathbf{C}$, which implies $\operatorname{aff}(\mathbf{C}) = \operatorname{aff}(\bar{\mathbf{C}})$. By Lemma D.1, we have

$$\frac{|\operatorname{Conv}(\bar{\mathbf{C}})|}{|\operatorname{Conv}(\mathbf{C})|} = \sqrt{\frac{\det(\bar{\mathbf{C}}^T\bar{\mathbf{C}})}{\det(\mathbf{C}^T\mathbf{C})}} = \sqrt{\frac{\det(\bar{\mathbf{C}}^T\bar{\mathbf{C}})}{\det(\mathbf{B}^T\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{B})}} = \frac{1}{|\det(\mathbf{B})|}.$$

Therefore, it suffices to show

$$|\det(\mathbf{B})| \leqslant 1. \tag{D.4}$$

We first show that for any row of $\mathbf{B}$, we have $\mathbf{B}(f,:) \in cone(\mathbf{W})^* \subseteq \mathcal{K}$. Since $\mathbf{CW} = \bar{\mathbf{C}}\mathbf{B}\mathbf{W} = \bar{\mathbf{C}}\bar{\mathbf{W}}$ and $\bar{\mathbf{C}}^T\bar{\mathbf{C}} \in \mathbb{R}^{k\times k}$ is invertible, we have

$$\mathbf{BW} = (\bar{\mathbf{C}}^T\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{B}\mathbf{W} = (\bar{\mathbf{C}}^T\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^T\bar{\mathbf{C}}\bar{\mathbf{W}} = \bar{\mathbf{W}}.$$

Because $\bar{\mathbf{W}} \geqslant \mathbf{0}_{k\times d}$, we obtain that, for any row of $\mathbf{B}$, $\mathbf{B}(f,:) \in \mathbb{R}^k$,

$$\mathbf{B}^T(f,:)\mathbf{W} = \bar{\mathbf{W}}^T(f,:) \geqslant \mathbf{0}.$$

That is, $\mathbf{B}(f,:) \in cone(\mathbf{W})^*$, which consequently implies that

$$\|\mathbf{B}(f,:)\|_2 \leqslant \mathbf{B}(f,:)^T\mathbf{1}_k. \tag{D.5}$$

Combining (D.4), (D.5) and the Hadamard Inequality and Inequality of Arithmetic and Geometric means (AM-GM), we can show (D.4) as follows:

$$\begin{aligned}|\det(\mathbf{B})| \overset{Hadamard's}{\leqslant} \prod_{f=1}^{k}\|\mathbf{B}(f,:)\|_2 \overset{(D.5)}{\leqslant} \prod_{f=1}^{k}\mathbf{B}^T(f,:)\mathbf{1}_k \overset{AM-GM}{\leqslant} \left(\frac{\sum_{f=1}^{k}\mathbf{B}^T(f,:)\mathbf{1}_k}{k}\right)^k \\ = \left(\frac{\sum\mathbf{B}^T\mathbf{1}_k}{k}\right)^k \overset{(D.3)}{=} 1.\end{aligned} \tag{D.6}$$

□

Next, we give the proof of Theorem 2.

Proof of Theorem 2. Suppose $\mathbf{CW} = \bar{\mathbf{C}}\bar{\mathbf{W}}$ and $|\operatorname{Conv}(\bar{\mathbf{C}})| \leqslant |\operatorname{Conv}(\mathbf{C})|$. Following the notation of the proof of Proposition 10, we aim to show that $\mathbf{B}$ is a permutation matrix.

To complete the proof, we only need to verify the following three conditions on $\mathbf{B}$.

(1.i) Any row of $\mathbf{B}$ belongs to $bd\mathcal{K} \bigcap cone(\mathbf{W})^*$, i.e.,

$$\mathbf{B}(f,:) \in \{\lambda\mathbf{e}_s : s = 1,\cdots,k,\ \lambda \geqslant 0\}, \forall f \in [k].$$

(1.ii) Any row sum of $\mathbf{B}$ is one, which, along with (1.i), implies

$$\mathbf{B}(f,:) \in \{\lambda\mathbf{e}_s : s = 1,\cdots,k,\ \lambda \geqslant 0\}, \forall f \in [k].$$

(1.iii) $\det(\mathbf{B}) = 1$. Along with the previous two conditions, it implies

$$\{\mathbf{B}(1,:), \mathbf{B}(2,:), \cdots, \mathbf{B}(k,:)\} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\};$$

that is, $\mathbf{B}$ must be a permutation matrix.

First, by the condition $|\operatorname{Conv}(\bar{\mathbf{C}})| \leqslant |\operatorname{Conv}(\mathbf{C})|$ and Proposition 10, we have $|\operatorname{Conv}(\bar{\mathbf{C}})| = |\operatorname{Conv}(\mathbf{C})|$, or equivalently $\det(\mathbf{B}) = 1$, i.e., (1.iii) holds.

Consequently, all inequalities in (D.6) become equalities. Specifically,

$$\|\mathbf{B}(f,:)\|_2 = \mathbf{B}^T(f,:)\mathbf{1}_k = 1, \, \forall f \in [k], \tag{D.7}$$

which implies that the row sums of $\mathbf{B}$ are all 1's, i.e., (1.ii) holds.

The above equation (D.7) also implies that $\mathbf{B}(f,:)$ is on the boundary of $\mathcal{K}$, $\mathbf{B}(f,:) \in bd\mathcal{K}$. Together with the fact that $\mathbf{B}(f,:)$ is in $cone(\mathbf{W})^*$ (proved in the proof of Proposition 10), it implies that (1.i) holds. □

D.2 Proof of Theorem 4

The sketch of this proof is as follows:

Step 1: We first show that with high probability, all true word frequency vectors, columns of $\mathbf{U}^0$, are close to the estimated convex polytope $\operatorname{Conv}(\hat{\mathbf{C}}_n)$. More specifically, we show in Lemma D.3 that there exists a $k \times d$ column-stochastic matrix [5] $\hat{\mathbf{W}}_n$ such that

$$\mathbf{U}^0 = \mathbf{C}^0\mathbf{W}^0 = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_n + \mathbf{E}_n \tag{D.8}$$

and $\max_i \|\mathbf{E}_n(:,i)\|_2 \leqslant C\epsilon_n$.

Step 2: We then work with a subset of s documents. Let $\mathbf{W}_1^0 \in \mathbb{R}^{k\times s}$ be the collection of the s columns of $\mathbf{W}^0$ that are (α,β)-SS; let $\hat{\mathbf{W}}_{n1}$ and $\mathbf{E}_{n1}$ be the corresponding sub-matrices of $\hat{\mathbf{W}}_n$ and $\mathbf{E}_n$, respectively. As a consequence of (D.8), we have

$$\mathbf{C}^0\mathbf{W}_1^0 = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_{n1} + \mathbf{E}_{n1}. \tag{D.9}$$

We can upper bound the estimation error by the summation of the following two terms:

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant \|\mathbf{E}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\|_2 + \min_{\mathbf{\Pi}} \sqrt{k}\|\mathbf{B} - \mathbf{\Pi}\|_2, \tag{D.10}$$

[5] We say a matrix is column-stochastic, if its entries are non-negative and columns sum to one.

where $\mathbf{B} = \hat{\mathbf{W}}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}$. By Lemma D.3, the first term is upper bounded.

Step 3: We show that for all $f = 1, \cdots, k$, $\mathbf{B}(f,:)$ satisfies:

$$\mathbf{B}(f,:) \in [cone(\mathbf{W}_1^0)^*]^{C_1\epsilon_n} \bigcap [bd\mathcal{K}]^{C_1\epsilon_n} \tag{D.11}$$

Then by the definition of (α, β)-SS, $\mathbf{B}(f,:)$'s are all close to indicator vectors. Using Lemma D.4 and letting $\alpha = C_1\epsilon_n$, we can prove that the matrix $\mathbf{B}$ is close to a permutation matrix. So the second term in (D.10) can be bounded. Putting all the steps together, we obtain that with high probability,

$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) \leqslant D_1\sqrt{\frac{\log(n \vee d)}{n}} + D_2\beta.$$

In the following, we provide the details of the above-mentioned steps.

Proof of Theorem 4.

Step 1: The following lemma is useful; its proof is given in Appendix E.

Lemma D.3. *With probability at least* $(1 - 3/(n \vee d)^c)^d$, *there exists a matrix* $\hat{\mathbf{W}}_n \in \mathbb{R}^{k\times d}$ *satisfying* $\hat{\mathbf{W}}_n \geqslant 0$, $\hat{\mathbf{W}}_n^T\mathbf{1}_k = \mathbf{1}_d$ *such that*

$$\mathbf{U}^0 = \mathbf{C}^0\mathbf{W}^0 = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_n + \mathbf{E}_n$$

and each column of $\mathbf{E}_n$ *satisfies*

$$\|\mathbf{E}_n(:,i)\|_2 \leqslant C\epsilon_n \tag{D.12}$$

for all $i = 1, \cdots, d$, *where* $c, C > 0$ *are constants independent of* n *and* d.

Step 2: By Lemma D.3, we have

$$\mathbf{C}^0\mathbf{W}_1^0 = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_{n1} + \mathbf{E}_{n1},$$

and $\|\mathbf{E}_{n1}\|_2 \leqslant C\sqrt{s}\epsilon_n$.

Let $\mathbf{B} = \hat{\mathbf{W}}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}$. Then,

$$\mathbf{C}^0 = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1} + \mathbf{E}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1} = \hat{\mathbf{C}}_n\mathbf{B} + \tilde{\mathbf{E}}_{n1} \tag{D.13}$$

where $\tilde{\mathbf{E}}_{n1} = \mathbf{E}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}$. We can bound $\|\tilde{\mathbf{E}}_{n1}\|_2$ by

$$\|\tilde{\mathbf{E}}_{n1}\|_2 \leqslant [\sigma^+_{\min}(\mathbf{W}_1^0)]^{-1}\|\mathbf{E}_{n1}\|_2 \leqslant k \cdot C\sqrt{s}\epsilon_n = C'\sqrt{s}\epsilon_n.$$

Then, we have

$$\begin{aligned}\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}^0) = \min_{\mathbf{\Pi}} \|\hat{\mathbf{C}}_n\mathbf{\Pi} - \mathbf{C}^0\|_2 &= \min_{\mathbf{\Pi}} \|\hat{\mathbf{C}}_n\mathbf{\Pi} - \hat{\mathbf{C}}_n\mathbf{B} - \tilde{\mathbf{E}}_{n1}\|_2 \\ &\leqslant \min_{\mathbf{\Pi}} \|\hat{\mathbf{C}}_n\|_2\|\mathbf{B} - \mathbf{\Pi}\|_2 + \|\tilde{\mathbf{E}}_{n1}\|_2 \\ &\leqslant \min_{\mathbf{\Pi}} \sqrt{k}\|\mathbf{B} - \mathbf{\Pi}\|_2 + \|\tilde{\mathbf{E}}_{n1}\|_2\end{aligned}$$

Step 3: Now, it suffices to show that for some permutation matrix $\mathbf{\Pi}$,

$$\|\mathbf{B} - \mathbf{\Pi}\|_2 \leqslant C''\beta.$$

We will use the following Lemma D.4 to prove the above inequality. The proof of Lemma D.4 is deferred to Appendix E.

Lemma D.4. *For a matrix $\mathbf{B} \in \mathbb{R}^{k\times k}$, if it satisfies the following conditions*

(2.i) $\mathbf{B}^T\mathbf{1}_k = \mathbf{1}_k$;

(2.ii) $\|\mathbf{B}\|_2 \leqslant M$;

(2.iii) any row of $\mathbf{B}$ belongs to $[bd\mathcal{K}]^\alpha \bigcap [cone(\mathbf{W}_1^0)^*]^\alpha$, so that

$$\mathbf{B}(f,:) \in \{\lambda\mathbf{e}_l + \boldsymbol{\epsilon} : l = 1, \cdots, k,\ \lambda \geqslant 0, \|\boldsymbol{\epsilon}\|_2 \leqslant \beta\lambda\}, f = 1, \cdots, k;$$

then there exists a permutation matrix $\mathbf{\Pi}$, such that

$$\|\mathbf{B} - \mathbf{\Pi}\|_2 \leqslant C'''M\beta,$$

where C''' is a constant independent of n and d.

Next, we verify the conditions in Lemma D.4.

Firstly, the proof of (2.i) $\mathbf{B}^T\mathbf{1}_k = \mathbf{1}_k$ is similar to the proof of Proposition 10 equation (D.3), so we omit it here.

Secondly, (2.ii) holds because

$$\|\mathbf{B}\|_2 \leqslant \sigma_{\max}(\hat{\mathbf{W}}_{n1})[\sigma^+_{\min}(\mathbf{W}_1^0)]^{-1} \leqslant \sqrt{s}\|\hat{\mathbf{W}}_{n1}\|_1[\sigma^+_{\min}(\mathbf{W}_1^0)]^{-1} \leqslant \sqrt{s}\cdot k = M.$$

Thirdly, to prove (2.iii), it suffices to verify the followings hold for any $f \in [k]$,

1. $\mathbf{B}(f,:) \in [cone(\mathbf{W}_1^0)^*]^{C_1\epsilon_n}$, i.e.,

$$\mathbf{B}(f,:)^T\mathbf{W}_1^0 \geqslant -C_1\epsilon_n\|\mathbf{B}(f,:)\|_2\mathbf{1}_s. \tag{D.14}$$

2. $\mathbf{B}(f,:) \in [bd\mathcal{K}]^{C_1\epsilon_n}$, i.e.,

$$\|\mathbf{B}(f,:)\|_2 - \mathbf{B}(f,:)^T\mathbf{1}_k \leqslant C_1\epsilon_n\|\mathbf{B}(f,:)\|_2, \tag{D.15}$$

$$\|\mathbf{B}(f,:)\|_2 - \mathbf{B}(f,:)^T\mathbf{1}_k \geqslant -C_1\epsilon_n\|\mathbf{B}(f,:)\|_2, \tag{D.16}$$

Now, we proceed to verify (D.14), (D.15) and (D.16). The following lemma is useful; its proof is given in Appendix E.

Lemma D.5.

$$|\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)| \leqslant (1 + C'\epsilon_n)|\det(\mathbf{C}^{0T}\mathbf{C}^0)| \tag{D.17}$$

where $C' > 0$ is a constant.

Since

$$\begin{aligned}\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n) &= \det(\mathbf{B}^{-T}(\mathbf{C}^0 - \tilde{\mathbf{E}}_{n1})^T(\mathbf{C}^0 - \tilde{\mathbf{E}}_{n1})\mathbf{B}^{-1})) \\ &= |\det(\mathbf{B}^{-1})|^2\det(\mathbf{C}^{0T}\mathbf{C}^0 - \mathbf{C}^{0T}\tilde{\mathbf{E}}_{n1} - \tilde{\mathbf{E}}_{n1}^T\mathbf{C}^0 + \tilde{\mathbf{E}}_{n1}^T\tilde{\mathbf{E}}_{n1}) \\ &= |\det(\mathbf{B}^{-1})|^2\det(\mathbf{C}^{0T}\mathbf{C}^0)\det(\mathbf{I} - \mathbf{F}_n),\end{aligned} \tag{D.18}$$

where $\mathbf{F}_n = (\mathbf{C}^{0T}\mathbf{C}^0)^{-1}\mathbf{C}^{0T}\tilde{\mathbf{E}}_{n1} + (\mathbf{C}^{0T}\mathbf{C}^0)^{-1}\tilde{\mathbf{E}}_{n1}^T\mathbf{C}^0 - (\mathbf{C}^{0T}\mathbf{C}^0)^{-1}\tilde{\mathbf{E}}_{n1}^T\tilde{\mathbf{E}}_{n1}$. Then $\|\mathbf{F}_n\|_2 \leqslant C_5\epsilon_n$. We order the singular values σ_i of $\mathbf{I} - \mathbf{F}_n$ as $\sigma_1 \leqslant \sigma_2 \leqslant \cdots \leqslant \sigma_k$. By Weyl's inequality in matrix theory (Weyl, 1912), $|1 - \sigma_i| \leqslant \|\mathbf{F}_n\|_2 \leqslant C_5\epsilon_n$ for all $i = 1, \cdots, k$. Therefore

$$\det(\mathbf{I} - \mathbf{F}_n) = \prod_{i=1}^{k}\sigma_i \geqslant (1 - C_5\epsilon_n)^k \geqslant 1 - kC_5\epsilon_n. \tag{D.19}$$

By (D.18) and (D.19), we have

$$\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n) \geqslant |\det(\mathbf{B}^{-1})|^2\det(\mathbf{C}^{0T}\mathbf{C}^0)(1 - C_5'\epsilon_n) \tag{D.20}$$

By (D.17) and (D.20), we have

$$|\det(\mathbf{B})| \geqslant 1 - C_6\epsilon_n. \tag{D.21}$$

- *Verify* (D.14).

  Right-multiplying $\mathbf{W}^0$ on both sides of (D.13), we have

$$\hat{\mathbf{C}}_n\hat{\mathbf{W}}_{n1} + \mathbf{E}_{n1} = \mathbf{C}^0\mathbf{W}_1^0 = \hat{\mathbf{C}}_n\mathbf{B}\mathbf{W}_1^0 + \mathbf{E}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0.$$

Then, left-multiply $(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T$ on both sides of the above equation:

$$\begin{aligned}\hat{\mathbf{W}}_{n1} + (\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\mathbf{E}_{n1} &= \mathbf{B}\mathbf{W}_1^0 + (\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\mathbf{E}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0 \\ \mathbf{B}\mathbf{W}_1^0 &= \hat{\mathbf{W}}_{n1} + (\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\mathbf{E}_{n1}(\mathbf{I} - \mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0) \\ &\geqslant -C_7\epsilon_n, \end{aligned} \tag{D.22}$$

The last inequality holds because $\hat{\mathbf{W}}_{n1} \geqslant 0$ and

$$\begin{aligned}&\|(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\mathbf{E}_{n1}(\mathbf{I} - \mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0)\|_F \\ \leqslant&\sqrt{k}\|(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\mathbf{E}_{n1}(\mathbf{I} - \mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0)\|_2 \\ \leqslant&\sqrt{k}\cdot\|(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)^{-1}\hat{\mathbf{C}}_n^T\|_2\cdot\|\mathbf{E}_{n1}\|_2\cdot\|\mathbf{I} - \mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}\mathbf{W}_1^0\|_2 \\ \leqslant&\sqrt{k}\cdot[\sigma_{\min}^+(\hat{\mathbf{C}}_n)]^{-1}\cdot C\sqrt{s}\epsilon_n\cdot 1 \\ \leqslant&C'\sqrt{s}\epsilon_n, \end{aligned} \tag{D.23}$$

where in the last inequality we use the fact that $\sigma_{\min}^+(\hat{\mathbf{C}}_n)$ is lower-bounded by a positive constant. That is because by (D.20),

$$\begin{aligned}\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n) &\geqslant |\det(\mathbf{B}^{-1})|^2\det(\mathbf{C}^{0T}\mathbf{C}^0)\,(1 - C_5'\epsilon_n) \\ &\geqslant \|\mathbf{B}\|_2^{-2k}\det(\mathbf{C}^{0T}\mathbf{C}^0)\,(1 - C_5'\epsilon_n) \\ &\geqslant M^{-2k}\det(\mathbf{C}^{0T}\mathbf{C}^0)\,(1 - C_5'\epsilon_n) \\ &\geqslant \frac{\det(\mathbf{C}^{0T}\mathbf{C}^0)}{2\cdot M^{2k}} \end{aligned} \tag{D.24}$$

At the same time,

$$\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n) \leqslant \|\hat{\mathbf{C}}_n\|_2^{2(k-1)}[\sigma_{\min}^+(\hat{\mathbf{C}}_n)]^2 \leqslant k^{k-1}[\sigma_{\min}^+(\hat{\mathbf{C}}_n)]^2 \tag{D.25}$$

Combining (D.20) and (D.25), we get a lower bound for $\sigma_{\min}^+(\hat{\mathbf{C}}_n)$.

- *Verify* (D.15).

  Since $\mathbf{1}_k^T\mathbf{W}_1^0 = \mathbf{1}_s^T$, by (D.22), we have

$$(\mathbf{B} + C_7\epsilon_n\mathbf{1}_{k\times k})\mathbf{W}_1^0 = \mathbf{B}\mathbf{W}_1^0 + C_7\epsilon_n\mathbf{1}_{k\times s} \geqslant 0, \tag{D.26}$$

  which implies that for any row of $\mathbf{B}$, $\mathbf{B}(f,:)$,

$$(\mathbf{B}(f,:) + C_7\epsilon_n\mathbf{1}_k) \in cone(\mathbf{W}_1^0)^* = \{\mathbf{x} : \mathbf{x}^T\mathbf{W}_1^0 \geqslant \mathbf{0}\} \subseteq \mathcal{K} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T\mathbf{1}\},$$

where we use the condition (S1) in the definition of SS condition. $(\mathbf{B}(f,:)+C_7\epsilon_n\mathbf{1}_k) \in \{\mathbf{x} : \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T\mathbf{1}\}$ implies that

$$\|\mathbf{B}(f,:)+C_7\epsilon_n\mathbf{1}_k\|_2 \leqslant (\mathbf{B}(f,:)+C_7\epsilon_n\mathbf{1}_k)^T\mathbf{1}_k$$
$$\|\mathbf{B}(f,:)\|_2 \leqslant \mathbf{B}(f,:)^T\mathbf{1}_k + C_8\epsilon_n. \tag{D.27}$$

- *Verify* (D.16).

  By Hadamard's inequality, Inequality of AM-GM, and (D.21), we have

$$\left(\frac{1}{k}\sum_{f=1}^{k}\|\mathbf{B}(f,:)\|_2\right)^k \overset{AM-GM}{\geqslant} \prod_{f=1}^{k}\|\mathbf{B}(f,:)\|_2 \overset{Hadamard's}{\geqslant} |\det(\mathbf{B})| \overset{(D.21)}{\geqslant} 1-C_6\epsilon_n. \tag{D.28}$$

  Consequently,

$$\frac{1}{k}\sum_{p=1}^{k}\|\mathbf{B}(p,:)\|_2 \geqslant (1-C_6\epsilon_n)^{1/k}$$
$$\frac{1}{k}\sum_{p=1}^{k}\|\mathbf{B}(p,:)\|_2 \geqslant (1-C_6\epsilon_n)^{1/k}\cdot\frac{1}{k}\sum_{p=1}^{k}[\mathbf{B}(p,:)^T\mathbf{1}_k] \qquad \text{by } \mathbf{B}^T\mathbf{1}_k=\mathbf{1}_k$$
$$\sum_{p=1}^{k}\|\mathbf{B}(p,:)\|_2 \geqslant \sum_{p=1}^{k}[\mathbf{B}(p,:)^T\mathbf{1}_k] - C_9\epsilon_n$$
$$\|\mathbf{B}(f,:)\|_2 + \sum_{p\neq f}^{k}\|\mathbf{B}(p,:)\|_2 \geqslant \mathbf{B}(f,:)^T\mathbf{1}_k + \sum_{p\neq f}^{k}\mathbf{B}(p,:)^T\mathbf{1}_k - C_9\epsilon_n$$
$$\|\mathbf{B}(f,:)\|_2 \geqslant \mathbf{B}(f,:)^T\mathbf{1}_k - \sum_{p\neq f}^{k}[\|\mathbf{B}(p,:)\|_2 - \mathbf{B}(p,:)^T\mathbf{1}_k] - C_9\epsilon_n$$
$$\|\mathbf{B}(f,:)\|_2 \geqslant \mathbf{B}(f,:)^T\mathbf{1}_k - (k-1)C_8\epsilon_n - C_9\epsilon_n \qquad \text{by (D.27)}$$
$$\geqslant \mathbf{B}(f,:)^T\mathbf{1}_k - C_{10}\epsilon_n, \quad \forall f=1,\cdots,k \tag{D.29}$$

- *Check* $\|\mathbf{B}(f,:)\|_2$ *is lower-bounded.*

  Now we show that, $\|\mathbf{B}(f,:)\|_2$ is lower-bounded, using Inequality of AM-GM and (D.27),

$$\|\mathbf{B}(f,:)\|_2\left(\frac{1}{k-1}\sum_{p\neq f}\|\mathbf{B}(p,:)\|_2\right)^{k-1} \overset{AM-GM}{\geqslant} \|\mathbf{B}(f,:)\|_2\prod_{p\neq f}\|\mathbf{B}(p,:)\|_2$$
$$\|\mathbf{B}(f,:)\|_2\left(\frac{\sum_{p\neq f}[\mathbf{B}(p,:)^T\mathbf{1}+C_8\epsilon_n]}{k-1}\right)^{k-1} \geqslant \|\mathbf{B}(f,:)\|_2\prod_{p\neq f}\|\mathbf{B}(p,:)\|_2 \qquad \text{by } (D.27)$$
$$\|\mathbf{B}(f,:)\|_2\left(\frac{k-\|\mathbf{B}(f,:)\|_2+kC_8\epsilon_n}{k-1}\right)^{k-1} \geqslant 1-C_6\epsilon_n \qquad \text{by (D.28)}$$

$$(1+C_8'\epsilon_n)\|\mathbf{B}(f,:)\|_2\left(\frac{k}{k-1}\right)^{k-1} \geqslant 1-C_6\epsilon_n$$

$$\|\mathbf{B}(f,:)\|_2 \geqslant e^{-1}(1-C_8''\epsilon_n) \qquad \text{by } \left(1+\frac{1}{x}\right)^x \leqslant e$$

$$\geqslant e^{-1}/2 \quad \forall f=1,\cdots,k. \tag{D.30}$$

- Now we put all the previous derivations together. From (D.26) and (D.30), we have (D.14) holds,

$$\mathbf{B}(f,:)^T\mathbf{W}_1^0 \geqslant -C_7\epsilon_n\mathbf{1}_s \geqslant -2e\cdot C_7\|\mathbf{B}(f,:)\|_2\epsilon_n\mathbf{1}_s.$$

  Similarly, from (D.27), (D.29) and (D.30), we have

$$\mathbf{B}(f,:)^T\mathbf{1}_k - 2e\cdot C_{10}\|\mathbf{B}(f,:)\|_2\epsilon_n \leqslant \|\mathbf{B}(f,:)\|_2 \leqslant \mathbf{B}(f,:)^T\mathbf{1}_k + 2e\cdot C_8\|\mathbf{B}(f,:)\|_2\epsilon_n.$$

  Therefore, (D.15) and (D.16) hold.

□

D.3 Proof of Theorem 8

Proof. This proof consists of two major steps:

Step 1: We apply Chernoff bound to show that with probability at least $1-D_1's/d$, for any $\mathbf{w}_i^\sharp$, there exists at least one sample $\mathbf{w}_{(i)}^1$, such that

$$\|\mathbf{w}_{(i)}^1 - \mathbf{w}_i^\sharp\|_2 \leqslant r_d, \quad \forall i=1,\cdots,s,$$

where $r_d = \left(\frac{\log d}{d}\right)^{\frac{1}{k-1}}$.

Step 2: Let $\mathbf{W}_1^0 = \{\mathbf{w}_{(i)}^1\}_{i=1}^s$, $\hat{\mathbf{W}}_n = \{\hat{\mathbf{w}}_{n(i)}\}_{i=1}^d$, and $\mathbf{B} = \hat{\mathbf{W}}_{n1}\mathbf{W}_1^{0T}(\mathbf{W}_1^0\mathbf{W}_1^{0T})^{-1}$. We show with probability at least $1-D_2'd/(n\vee d)^c$, for all $f=1,\cdots,k$, $\mathbf{B}(f,:)$ satisfies:

$$\mathbf{B}(f,:) \in [cone(\mathbf{W}_1^0)^*]^{C_1'\epsilon_n} \bigcap [bd\mathcal{K}]^{C_1'\epsilon_n} \tag{D.31}$$

Then using the conclusion from Theorem 4, we get the desired bound.

In the following, we provide the details of the above-mentioned steps.

Step 1: Let X_i denote a random variable representing the number of documents falls into the ball $B(\mathbf{w}_i^\sharp, r_d)$ $(r_d \leqslant r_0)$ in a sample of size d drawn from $\mathcal{P}$,

$$X_i \sim \text{Binomial}(d, p_i)$$

where $p_i \geqslant (k-1)! \cdot c_0 \cdot r_d^{k-1}$. Since $\mathcal{P}$ is an (α, β)-SS distribution, we have

$$p_i = \mathbb{P}\left(\|\mathbf{w} - \mathbf{w}_i^\sharp\|_2 \leqslant r_d\right) \geqslant (k-1)! \cdot c_0 \cdot r_d^{k-1} = C_3 \frac{\log d}{d} \tag{D.32}$$

According to Chernoff bound, for $0 < \delta < 1$,

$$\mathbb{P}\left(X_i \leqslant (1-\delta) C_3 \log d\right) \overset{\text{(D.32)}}{\leqslant} \mathbb{P}\left(X_i \leqslant (1-\delta) d p_i\right) \overset{\text{Chernoff}}{\leqslant} \exp\left(-\frac{\delta^2 d p_i}{2}\right) \overset{\text{(D.32)}}{\leqslant} \exp\left(-\delta^2 C_3 \log d/2\right).$$

Therefore, when d is large enough, such that for some $0 < \delta_0 < 1, (1-\delta_0) C_3 \log d \geqslant \frac{1}{2}$, we have

$$\mathbb{P}\left(X_i \leqslant \frac{1}{2}\right) \leqslant \exp\left(-\delta_0^2 C_3 \log d/2\right) = D_1' \frac{1}{d}, \quad \forall i = 1, \cdots, s$$

Then, we can bound the probability of the event $\left\{\min_{i=1,\cdots,s} X_i \leqslant \frac{1}{2}\right\}$,

$$P\left(\min_{i=1,\cdots,s} X_i \leqslant \frac{1}{2}\right) \leqslant \sum_{i=1}^{s} P\left(X_i \leqslant \frac{1}{2}\right) \leqslant D_1' \frac{s}{d}.$$

In other words, with probability at least $1 - D_1' s/d$, there exist s different samples $\mathbf{w}_{(1)}^1, \cdots, \mathbf{w}_{(s)}^1$, such that $\|\mathbf{w}_{(i)}^1 - \mathbf{w}_i^\sharp\|_2 \leqslant r_d$.

Step 2: Denote $\mathbf{W}_1^0 = \{\mathbf{w}_{(i)}^1\}_{i=1}^s$, $\mathbf{W}^0 = (\mathbf{W}_1^0, \mathbf{W}_2^0) \in \mathbb{R}^{k \times d}$, and $\hat{\mathbf{W}}_n = \{\hat{\mathbf{w}}_{n(i)}\}_{i=1}^d$. We have

$$\mathbf{C}^0 \mathbf{W}^0 = \hat{\mathbf{C}}_n \hat{\mathbf{W}}_n + \mathbf{E}_n$$

Therefore,

$$\mathbf{C}^0 \mathbf{W}_1^0 = \hat{\mathbf{C}}_n \hat{\mathbf{W}}_{n1} + \mathbf{E}_{n1}, \quad \mathbf{W}_1^0 = \mathbf{W}^\sharp + \mathbf{E}_d', \tag{D.33}$$

where $\hat{\mathbf{W}}_{n1}$ and $\mathbf{E}_{n1}$ are the collections of the corresponding columns from $\hat{\mathbf{W}}_n$ and $\mathbf{E}_n$ respectively. Moreover, $\|\mathbf{E}_{n1}(:, j)\|_2 \leqslant C_3 \epsilon_n$ and $\|\mathbf{E}_d'(:, j)\|_2 \leqslant r_d$, for all $j = 1, \cdots, s$.

Now we show that $[cone(\mathbf{W}_1^0)^*]^{\alpha - r_d} = \{\mathbf{x} : \mathbf{x}^T \mathbf{W}_1^0 \geqslant -(\alpha - r_d)\|\mathbf{x}\|_2\} \subseteq [cone(\mathbf{W}^\sharp)^*]^\alpha = \{\mathbf{x} : \mathbf{x}^T \mathbf{W}^\sharp \geqslant -\alpha \|\mathbf{x}\|_2\}$. For any $\mathbf{x} \in \mathbb{R}^k$ and $\mathbf{x}^T \mathbf{W}_1^0 \geqslant -(\alpha - r_d)\|\mathbf{x}\|_2$,

$$-\alpha \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T \mathbf{W}_1^0 - r_d \|\mathbf{x}\|_2 \leqslant \mathbf{x}^T \mathbf{W}_1^0 + \mathbf{x}^T (\mathbf{W}^\sharp - \mathbf{W}_1^0) = \mathbf{x}^T \mathbf{W}^\sharp,$$

where in the second inequality we apply (E.78) and Cauchy–Schwarz inequality. Therefore, by definition, if $\mathbf{W}^\sharp$ is (α, β)-SS, $\mathbf{W}_1^0$ is $(\alpha - r_d, \beta)$-SS.

Back to our case, since $\mathbf{W}^\sharp$ is $(C_1'\sqrt{\frac{\log(n\vee d)}{n}}+r_d,\beta)$-SS, $\mathbf{W}_1^0$ is $(C_1'\sqrt{\frac{\log(n\vee d)}{n}},\beta)$-SS. Then by Theorem 4, we obtain that with probability at least $1-D_1's/d-D_2'd/(n\vee d)^c$,

$$\mathcal{D}(\hat{\mathbf{C}}_n,\mathbf{C}^0)\leqslant D_3'\sqrt{\frac{\log(n\vee d)}{n}}+D_4'\beta.$$

□

E Proofs of Technical Lemmas and Propositions

In this section, we provide proofs of propositions and all technical lemmas. From now on, we use $\mathbf{u}^0$ to denote the true word frequency; $\hat{\mathbf{u}}$ the the sample word frequency; $\tilde{\mathbf{u}} = \hat{\mathbf{C}}_n\hat{\mathbf{w}}$ the estimated word frequency in $\mathrm{Conv}(\hat{\mathbf{C}}_n)$. We use the superscript (i) to denote the i-th document. For example, $\mathbf{u}^{0(i)}$ denotes the true word frequency of the i-th document and $\mathbf{x}^{(i)}$ denotes the observation of the i-th document. We write $\mathbf{U}^0 = \mathbf{C}^0\mathbf{W}^0 = (\mathbf{u}^{0(1)}, \cdots, \mathbf{u}^{0(d)}) \in \mathbb{R}^{V\times d}$ and $\tilde{\mathbf{U}}_n = \hat{\mathbf{C}}_n\hat{\mathbf{W}}_n = (\tilde{\mathbf{u}}^{(1)}, \cdots, \tilde{\mathbf{u}}^{(d)}) \in \mathbb{R}^{V\times d}$.

We use $f^{(i)}(\mathbf{u})$ as a shorthand notation of $f_n(\mathbf{u}; \mathbf{x}^{(i)})$. By Pinsker's inequality, we have

$$\frac{f^{(j)}(\mathbf{u})}{f^{(j)}(\hat{\mathbf{u}}^{(j)})} \leqslant \exp\left(-\frac{n}{2}\|\hat{\mathbf{u}}^{(j)} - \mathbf{u}\|_2^2\right), \tag{E.34}$$

for any $\mathbf{u} \in \Delta^{V-1}$. By the reverse Pinsker's inequality (Götze et al., 2019), we have

$$\frac{f^{(j)}(\mathbf{u})}{f^{(j)}(\hat{\mathbf{u}}^{(j)})} \geqslant \exp\left(-C_6 n\|\hat{\mathbf{u}}^{(j)} - \mathbf{u}\|_2^2\right), \tag{E.35}$$

where $C_6 = (\min_{i\in[V]} u_i)^{-1}$ depends on the minimum element of $\mathbf{u}$.

E.1 Proof of Lemma D.1

Proof. Write $\mathbf{C} = (\mathbf{c}_1, \cdots, \mathbf{c}_k) \in \mathbb{R}^{V\times k}$ and $\tilde{\mathbf{C}} = (\tilde{\mathbf{c}}_1, \cdots, \tilde{\mathbf{c}}_{k-1}) \in \mathbb{R}^{V\times(k-1)}$, where $\tilde{\mathbf{c}}_j = \mathbf{c}_j - \mathbf{c}_k$ with $j \in [k-1]$. Write

$$\mathbf{G} = \mathbf{C}^T\mathbf{C}, \quad \tilde{\mathbf{G}} = \tilde{\mathbf{C}}^T\tilde{\mathbf{C}}.$$

The volume of the k-dimensional parallelepiped spanned by $\mathbf{c}_1, \cdots, \mathbf{c}_k \in \mathbb{R}^V$ is given by $\sqrt{\det(\mathbf{G})}$ (Boyd and Vandenberghe, 2004). Therefore $\sqrt{\det(\tilde{\mathbf{G}})} = (k-1)!|\,\mathrm{Conv}(\mathbf{C})|$, since $\sqrt{\det(\tilde{\mathbf{G}})}$ measures the volume of the $(k-1)$-dimensional parallelepiped spanned by columns of $\tilde{\mathbf{C}}$ in $\mathbb{R}^V$, which is $(k-1)!$ times larger than the volume of $\mathrm{Conv}(\mathbf{C})$. It suffices to show that

$$h^2\det(\tilde{\mathbf{G}}) = \det(\mathbf{G}).$$

Denote by $\mathbf{v}$ the perpendicular vector to $\mathrm{aff}(\mathbf{C})$, represented as

$$\mathbf{v} = \sum_{j=1}^{k-1} t_j(\mathbf{c}_j - \mathbf{c}_k) + \mathbf{c}_k,$$

so that

$$(\mathbf{c}_j - \mathbf{c}_k)^T\mathbf{v} = 0 \ \text{ and } \ \mathbf{c}_k^T\mathbf{v} = \|\mathbf{v}\|_2 = h^2.$$

Further, we construct a system of k linear equations for k unknowns, $t_1, \cdots, t_{k-1}, h^2$,

$$\sum_{j=1}^{k-1} t_j \tilde{\mathbf{c}}_i^T \tilde{\mathbf{c}}_j = -\tilde{\mathbf{c}}_i^T \mathbf{c}_k,\ i = 1, \cdots, k-1 \tag{E.36}$$

$$\sum_{j=1}^{k-1} t_j \tilde{\mathbf{c}}_j^T \mathbf{c}_k - h^2 = -\mathbf{c}_k^T \mathbf{c}_k. \tag{E.37}$$

By Cramer's rule, we have

$$h^2 = \frac{1}{\det\left(\begin{bmatrix} & \tilde{\mathbf{G}} & & 0 \\ & & & \vdots \\ & & & 0 \\ \tilde{\mathbf{c}}_1^T\mathbf{c}_k & \cdots & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k & -1 \end{bmatrix}\right)} \det\left(\begin{bmatrix} & \tilde{\mathbf{G}} & & -\tilde{\mathbf{c}}_1^T\mathbf{c}_k \\ & & & \vdots \\ & & & -\tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k \\ \tilde{\mathbf{c}}_1^T\mathbf{c}_k & \cdots & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k & -\mathbf{c}_k^T\mathbf{c}_k \end{bmatrix}\right)$$

Then, we see that for the denominator,

$$\det\left(\begin{bmatrix} & \tilde{\mathbf{G}} & & 0 \\ & & & \vdots \\ & & & 0 \\ \tilde{\mathbf{c}}_1^T\mathbf{c}_k & \cdots & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k & -1 \end{bmatrix}\right) = -\det(\tilde{\mathbf{G}})$$

The numerator is

$$\begin{aligned} & -\det\left(\begin{bmatrix} \tilde{\mathbf{c}}_1^T\tilde{\mathbf{c}}_1 & \cdots & \tilde{\mathbf{c}}_1^T\tilde{\mathbf{c}}_{k-1} & \tilde{\mathbf{c}}_1^T\mathbf{c}_k \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{\mathbf{c}}_{k-1}^T\tilde{\mathbf{c}}_1 & \cdots & \tilde{\mathbf{c}}_{k-1}^T\tilde{\mathbf{c}}_{k-1} & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k \\ \tilde{\mathbf{c}}_1^T\mathbf{c}_k & \cdots & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k & \mathbf{c}_k^T\mathbf{c}_k \end{bmatrix}\right) \\ \xlongequal{\text{add last column to others}} & -\det\left(\begin{bmatrix} \tilde{\mathbf{c}}_1^T\mathbf{c}_1 & \cdots & \tilde{\mathbf{c}}_1^T\mathbf{c}_{k-1} & \tilde{\mathbf{c}}_1^T\mathbf{c}_k \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_1 & \cdots & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_{k-1} & \tilde{\mathbf{c}}_{k-1}^T\mathbf{c}_k \\ \mathbf{c}_1^T\mathbf{c}_k & \cdots & \mathbf{c}_{k-1}^T\mathbf{c}_k & \mathbf{c}_k^T\mathbf{c}_k \end{bmatrix}\right) \\ \xlongequal{\text{add last row to others}} & -\det\left(\begin{bmatrix} \mathbf{c}_1^T\mathbf{c}_1 & \cdots & \mathbf{c}_1^T\mathbf{c}_{k-1} & \mathbf{c}_1^T\mathbf{c}_k \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{c}_{k-1}^T\mathbf{c}_1 & \cdots & \mathbf{c}_{k-1}^T\mathbf{c}_{k-1} & \mathbf{c}_{k-1}^T\mathbf{c}_k \\ \mathbf{c}_1^T\mathbf{c}_k & \cdots & \mathbf{c}_{k-1}^T\mathbf{c}_k & \mathbf{c}_k^T\mathbf{c}_k \end{bmatrix}\right) \end{aligned}$$

$$= -\det(\mathbf{G}).$$

□

E.2 Proof of Lemma D.2

Proof. We will show that for any $\mathbf{x} \in \mathbb{R}^k, \mathbf{x} \neq \mathbf{0}$, there exists $\boldsymbol{\beta} \in \mathbb{R}^d$ such that

$$\frac{\mathbf{x}^T(\mathbf{W}\boldsymbol{\beta})}{\|\mathbf{x}\|_2} \geqslant \frac{1}{k}, \text{ and } \|\boldsymbol{\beta}\|_2 \leqslant 1. \tag{E.38}$$

Therefore,

$$\frac{\|\mathbf{x}^T\mathbf{W}\|_2}{\|\mathbf{x}\|_2} \geqslant \frac{1}{\|\boldsymbol{\beta}\|_2}\frac{\mathbf{x}^T\mathbf{W}\boldsymbol{\beta}}{\|\mathbf{x}\|_2} \geqslant \frac{1}{k}.$$

In the following, we will find $\boldsymbol{\beta}$ satisfying (E.38).

First, decompose $\mathbf{x}$ as

$$\mathbf{x} = \frac{\lambda}{k}\mathbf{1}_k + \boldsymbol{\gamma},$$

for some $\lambda \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathbb{R}^k$ such that $\boldsymbol{\gamma}^T\mathbf{1}_k = 0$.

Second, let

$$\mathbf{y} = \frac{sign(\lambda)}{k} \cdot \mathbf{1}_k + \frac{1}{\sqrt{k(k-1)}\|\boldsymbol{\gamma}\|_2} \cdot \boldsymbol{\gamma}$$

where $sign(\cdot)$ is the sign function. Next we verify that

$$\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2} \geqslant \frac{1}{k}.$$

This is because

$$\begin{aligned}
\mathbf{x}^T\mathbf{y} &= \frac{|\lambda|}{k} + \frac{1}{\sqrt{k(k-1)}}\|\boldsymbol{\gamma}\|_2 \\
\|\mathbf{x}\|_2 &= \sqrt{\frac{\lambda^2}{k} + \|\boldsymbol{\gamma}\|_2^2} \leqslant \begin{cases} |\lambda| & \text{if } |\lambda| \geqslant \sqrt{\frac{k}{k-1}}\|\boldsymbol{\gamma}\|_2 \\ \sqrt{\frac{k}{k-1}}\|\gamma\|_2 & \text{if } |\lambda| < \sqrt{\frac{k}{k-1}}\|\boldsymbol{\gamma}\|_2 \end{cases},
\end{aligned}$$

and

$$\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2} \geqslant \begin{cases} \frac{1}{k} + \frac{1}{\sqrt{k(k-1)}}\frac{\|\boldsymbol{\gamma}\|_2}{|\lambda|} & \text{if } |\lambda| \geqslant \sqrt{\frac{k}{k-1}}\|\boldsymbol{\gamma}\|_2 \\ \frac{|\lambda|}{\|\boldsymbol{\gamma}\|_2}\frac{\sqrt{k-1}}{k\sqrt{k}} + \frac{1}{k} & \text{if } |\lambda| < \sqrt{\frac{k}{k-1}}\|\boldsymbol{\gamma}\|_2. \end{cases}$$

Third, we verify that $[sign(\lambda) \cdot \mathbf{y}] \in [\Delta^{k-1} \bigcap \mathcal{K}^*]$. This is because

$$[sign(\lambda) \cdot \mathbf{y}]^T\mathbf{1}_k = \frac{1}{k}\mathbf{1}_k^T\mathbf{1}_k = 1,$$

$$\|\mathbf{y}\|_2 = \sqrt{\frac{1}{k} + \frac{1}{k(k-1)}} = \frac{1}{\sqrt{k-1}}.$$

Since $\mathcal{K}^* \subseteq cone(\mathbf{W})$, we have

$$[\Delta^{k-1} \bigcap \mathcal{K}^*] \subseteq [\Delta^{k-1} \bigcap cone(\mathbf{W})] = \{\mathbf{x} \in \Delta^{k-1} : \mathbf{x} = \mathbf{W}\boldsymbol{\lambda}, \boldsymbol{\lambda} \geqslant 0\} = \text{Conv}(\mathbf{W}).$$

Therefore, $[sign(\lambda) \cdot \mathbf{y}] \in \text{Conv}(\mathbf{W})$, meaning that there exists $\boldsymbol{\beta}' \in \Delta^{d-1}$ such that

$$\mathbf{y} = sign(\lambda) \cdot \mathbf{W}\boldsymbol{\beta}' = \mathbf{W}[sign(\lambda)\boldsymbol{\beta}'] = \mathbf{W}\boldsymbol{\beta},$$

and

$$\|\boldsymbol{\beta}\|_2 = \|\boldsymbol{\beta}'\|_2 \leqslant \boldsymbol{\beta}'^T \mathbf{1}_k = 1.$$

□

E.3 Proof of Lemma D.3

We arrange the proof as Lemma E.1 and Lemma E.2. First, in Lemma E.1, we derive a lower bound for the integrated likelihood function, $F_{n\times d}(\hat{\mathbf{C}}_n; \mathbf{X})$. Then, we prove equation (D.12) in Lemma E.2.

We first define the δ-enlargement convex polytope below, which is useful later in the proof.

Definition 5 (δ-enlargement convex polytope)**.** *For a convex polytope,* $\text{Conv}(\mathbf{C}) \subseteq \mathbb{R}^V$*, with* k *linearly independent vertices* $\mathbf{C} = \{\mathbf{c}_f\}_{f=1}^k \in \mathbb{R}^{V\times k}$*. The* ***$\delta$-enlargement convex polytope*** *of* $\text{Conv}(\mathbf{C})$*, denoted as* $\text{Conv}(\mathbf{C}^\delta)$*, is defined such that each column of* $\mathbf{C}^\delta$*,*

$$\mathbf{c}_f^\delta = (1 + \rho(\mathbf{C})\delta)(\mathbf{c}_f - \bar{\mathbf{c}}) + \bar{\mathbf{c}}, \quad \forall f = 1, \cdots k,$$

where $\rho(\mathbf{C}) = \frac{k}{\sigma_{\min}^+(\mathbf{C})}$*, and* $\bar{\mathbf{c}} = \frac{1}{k}\sum_{f=1}^k \mathbf{c}_f \in \mathbb{R}^V$ *is the center of the* k *columns of* $\mathbf{C}$*.* $\mathbf{C}^\delta$ *is called the* ***δ-enlargement matrix*** *of* $\mathbf{C}$*.*

Proposition 11. $\text{Conv}(\mathbf{C}^\delta)$ *satisfies the following properties.*

1. *It composes of* k *vertices,* $\mathbf{C}^\delta = \{\mathbf{c}_f^\delta\}_{f=1}^k \in \mathbb{R}^{V\times k}$*;*
2. $|\text{Conv}(\mathbf{C}^\delta)| = (1 + \rho(\mathbf{C})\delta)^{k-1} |\text{Conv}(\mathbf{C})|$.

Lemma E.1. *With probability at least* $(1-3\cdot(n\vee d)^{-c})^d$*, the integrated likelihood is lower-bounded:*

$$F_{n\times d}(\hat{\mathbf{C}}_n;\mathbf{X}) \geqslant C\cdot A_{n,d}\cdot(n\vee d)^{-C_{10}d},$$

where $A_{n,d}:=\prod_{i=1}^d f_n(\hat{\mathbf{u}}^{(i)};\mathbf{x}^{(i)})$ *and* C, C_{10} *are constants.*

Proof. The integrated likelihood function can be written as

$$\begin{aligned}
F_{n\times d}(\mathbf{C};\mathbf{X}) =& \prod_{i=1}^d \frac{1}{|\operatorname{Conv}(\mathbf{C})|}\int_{\operatorname{Conv}(\mathbf{C})} f_n(\mathbf{u};\mathbf{x}^{(i)})d\mathbf{u}\\
=& \prod_{i=1}^d f_n(\hat{\mathbf{u}}^{(i)};\mathbf{x}^{(i)})\int_{\operatorname{Conv}(\mathbf{C})}\frac{1}{|\operatorname{Conv}(\mathbf{C})|}\frac{f_n(\mathbf{u};\mathbf{x}^{(i)})}{f_n(\hat{\mathbf{u}}^{(i)};\mathbf{x}^{(i)})}d\mathbf{u}\\
=& A_{n,d}\cdot\prod_{i=1}^d\int_{\operatorname{Conv}(\mathbf{C})}\frac{1}{|\operatorname{Conv}(\mathbf{C})|}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}
\end{aligned}$$

where $f^{(i)}(\mathbf{u})$ is a shorthand notation of $f_n(\mathbf{u};\mathbf{x}^{(i)})$.

By Devroye et al. (1983), for each document i, it holds with probability at least $1-3\cdot e^{-cx^2}$ that $\|\mathbf{u}^{0(i)}-\hat{\mathbf{u}}^{(i)}\|_2\leqslant\frac{5\sqrt{c}x}{\sqrt{n}}$ for all $x>0$. By a simple union bound argument, we have that with probability at least $(1-3\cdot(n\vee d)^{-c})^d$, $\|\mathbf{u}^{0(i)}-\hat{\mathbf{u}}^{(i)}\|_2\leqslant 5\sqrt{c}\cdot\sqrt{\frac{\log(n\vee d)}{n}}=:C_1\sqrt{\frac{\log(n\vee d)}{n}}$, for any $i\in[d]$, by choosing x to be a large multiple of $\sqrt{\log(n\vee d)}$. Let $\mathcal{B}(\mathbf{u}^{0(i)};C_1\epsilon_n)$ denote the Euclidean ball centered at $\mathbf{u}^{0(i)}$ with radius $C_1\epsilon_n$. Consequently, with high probability, for any $\mathbf{u}\in\mathcal{B}(\mathbf{u}^{0(i)};C_1\epsilon_n)$,

$$\|\mathbf{u}-\hat{\mathbf{u}}^{(i)}\|_2\leqslant\|\mathbf{u}-\mathbf{u}^{0(i)}\|_2+\|\mathbf{u}^{0(i)}-\hat{\mathbf{u}}^{(i)}\|_2\leqslant 2C_1\epsilon_n. \tag{E.39}$$

Next, by the definition of MLE, we have

$$\begin{aligned}
F_{n\times d}(\hat{\mathbf{C}}_n;\mathbf{X}) &\geqslant F_{n\times d}(\mathbf{C}^0;\mathbf{X})\\
&= \frac{A_{n,d}}{|\operatorname{Conv}(\mathbf{C}^0)|^d}\cdot\prod_{i=1}^d\int_{\operatorname{Conv}(\mathbf{C}^0)}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
&\geqslant \frac{A_{n,d}}{|\operatorname{Conv}(\mathbf{C}^0)|^d}\cdot\prod_{i=1}^d\int_{\operatorname{Conv}(\mathbf{C}^0)\bigcap\mathcal{B}(\mathbf{u}^{0(i)};C_1\epsilon_n)}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
&\geqslant \frac{A_{n,d}}{|\operatorname{Conv}(\mathbf{C}^0)|^d}\cdot\prod_{i=1}^d\int_{\operatorname{Conv}(\mathbf{C}^0)\bigcap\mathcal{B}(\mathbf{u}^{0(i)};C_1\epsilon_n)}\exp\left(-C_6n\|\mathbf{u}-\hat{\mathbf{u}}^{(i)}\|_2^2\right)d\mathbf{u} \qquad \text{(E.40)}\\
&\geqslant \frac{A_{n,d}}{|\operatorname{Conv}(\mathbf{C}^0)|^d}\cdot\prod_{i=1}^d\left[C_8(C_1\epsilon_n)^{k-1}\cdot\exp(-C_6n(2C_1\epsilon_n)^2)\right] \qquad \text{(E.41)}\\
&\geqslant C\cdot A_{n,d}\cdot(n\vee d)^{-C_{10}d}.
\end{aligned}$$

Inequality (E.40) follows from the reverse Pinsker's inequality (E.35) since the columns of $\mathbf{C}^0$ are interior points in Δ^{V-1}, i. Inequality (E.41) follows from (E.39).

□

Definition 6 (Distance between a vector and a convex polytope)**.** *The distance between a vector* $\mathbf{x}$ *and a convex polytope* $\mathrm{Conv}(\mathbf{C})$ *is defined as*

$$d(\mathbf{x}, \mathrm{Conv}(\mathbf{C})) = \min_{\mathbf{y} \in \mathrm{Conv}(\mathbf{C})} \|\mathbf{x} - \mathbf{y}\|_2.$$

Lemma E.2. *With probability at least* $(1 - 3 \cdot (n \vee d)^{-c})^d$*, we have*

$$d(\mathbf{u}^{0(i)}, \mathrm{Conv}(\hat{\mathbf{C}}_n)) \leqslant C\epsilon_n \tag{E.42}$$

for any $i \in [d]$*. Therefore, there exists a matrix* $\hat{\mathbf{W}}_n$*, such that* $\hat{\mathbf{W}}_n \geqslant 0$*,* $\hat{\mathbf{W}}_n^T \mathbf{1}_k = \mathbf{1}_d$*, and*

$$\mathbf{U}^0 = \mathbf{C}^0 \mathbf{W}^0 = \hat{\mathbf{C}}_n \hat{\mathbf{W}}_n + \mathbf{E}_n = \tilde{\mathbf{U}}_n + \mathbf{E}_n$$

and $\max_i \|\mathbf{E}_n(:, i)\|_2 \leqslant C\epsilon_n$*. Here constants* c *and* C *are independent of* n *and* d*.*

Proof. We prove the lemma by contradiction. Suppose the i-th document violates (E.42):

$$d(\mathbf{u}^{0(i)}, \mathrm{Conv}(\hat{\mathbf{C}}_n)) \geqslant C\epsilon_n.$$

First, we claim that there exist at least $C_1 d$ columns of $\mathbf{U}^0$ such that

$$d(\mathbf{u}^{0(i)}, \mathrm{Conv}(\hat{\mathbf{C}}_n)) \geqslant C_2 C \epsilon_n,$$

where $C_1, C_2 \in (0, 1)$ are constants independent of n and d. We prove this claim at the end.

Then, by Devroye et al. (1983), with probability at least $(1 - 3 \cdot (n \vee d)^{-c})^d$, we have $\|\mathbf{u}^{0(i)} - \hat{\mathbf{u}}^{(i)}\|_2 \leqslant O\left(\sqrt{\frac{\log(n \vee d)}{n}}\right)$ hold for all $i = 1, \cdots, d$. By making the constant C large enough, we have

$$d(\hat{\mathbf{u}}^{(j)}, \mathrm{Conv}(\hat{\mathbf{C}}_n)) \geqslant (C_2 C - 1)\epsilon_n.$$

Therefore,

$$\begin{aligned}
F_{n \times d}(\hat{\mathbf{C}}_n; \mathbf{X}) =& A_{n,d} \prod_{i=1}^d \int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \frac{1}{|\mathrm{Conv}(\hat{\mathbf{C}}_n)|} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})} d\mathbf{u} \\
\leqslant& \frac{A_{n,d}}{|\mathrm{Conv}(\hat{\mathbf{C}}_n)|^d} \prod_{i=1}^d \int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \exp\left(-\frac{n}{2}\|\hat{\mathbf{u}}^{(i)} - \mathbf{u}\|_2^2\right) d\mathbf{u} \\
=& A_{n,d} \prod_{i=1}^d \exp\left(-\frac{n}{2}\|\hat{\mathbf{u}}^{(i)} - \mathbf{u}^{*(i)}\|_2^2\right) \\
\leqslant& A_{n,d} \cdot \exp\left(-\frac{n}{2} \sum_{i=1}^d d^2(\hat{\mathbf{u}}^{(i)}, \mathrm{Conv}(\hat{\mathbf{C}}_n))\right) \\
\leqslant& A_{n,d} \cdot \exp\left(-\frac{n}{2} \cdot C_1 d \cdot (C_2 C - 1)^2 \epsilon_n^2\right)
\end{aligned}$$

$$= A_{n,d} \cdot (n \vee d)^{-\frac{1}{2}C_1(C_2C-1)^2C_0^2 d},$$

where the first inequality follows (E.34) and the second inequality is due to the mean value theorem for integrals with $\mathbf{u}^{*(i)}$'s being some points in $\text{Conv}(\hat{\mathbf{C}}_n)$. By choosing C large enough, we can make

$$F_{n\times d}(\hat{\mathbf{C}}_n; \mathbf{X}) \leqslant A_{n,d} \cdot (n \vee d)^{-(C_{10}+1)d},$$

which contradicts with Lemma E.1. So we conclude that

$$d(\mathbf{u}^{0(i)}, \text{Conv}(\hat{\mathbf{C}}_n)) \leqslant C\epsilon_n$$

for all $i = 1, \cdots, d$.

It remains to prove the claim we made at the beginning. When $\text{aff}(\mathbf{C}^0)$ is parallel to $\text{aff}(\hat{\mathbf{C}}_n)$, the claim is trivial by making C_2 small. When $\text{aff}(\mathbf{C}^0)$ is not parallel to $\text{aff}(\hat{\mathbf{C}}_n)$, again we prove it by contradiction. Suppose there are at least $(1 - C_1)d$ columns of $\mathbf{U}^0$ such that

$$d(\mathbf{u}^{0(j)}, \text{Conv}(\hat{\mathbf{C}}_n)) \leqslant C_2C\epsilon_n$$

and let $\mathcal{S}$ be their column index set.

Denote r as the distance from $\mathbf{u}^{0(i)}$ to the intersection of $\text{aff}(\mathbf{C}^0)$ and $\text{aff}(\hat{\mathbf{C}}_n)$, i.e.,

$$r = d(\mathbf{u}^{0(i)}, \text{aff}(\mathbf{C}^0) \bigcap \text{aff}(\hat{\mathbf{C}}_n)),$$

where $\mathbf{u}^{0(i)}$ is the vector such that $d(\mathbf{u}^{0(i)}, \text{Conv}(\hat{\mathbf{C}}_n)) \geqslant C\epsilon_n$. Since $d(\mathbf{u}^{0(j)}, \text{Conv}(\hat{\mathbf{C}}_n)) \leqslant C_2C\epsilon_n$ for all $j \in \mathcal{S}$, we know that

$$d(\mathbf{u}^{0(j)}, \text{aff}(\mathbf{C}^0) \bigcap \text{aff}(\hat{\mathbf{C}}_n)) \leqslant \frac{C_2C\epsilon_n}{C\epsilon_n} \cdot r = C_2 r, \ \forall j \in \mathcal{S}.$$

At the same time,

$$r - C_2 r \leqslant \max_{j\in\mathcal{S}} \|\mathbf{u}^{0(i)} - \mathbf{u}^{0(j)}\|_2 \leqslant \max_{i,j\in[k]} \|\mathbf{C}^{0(i)} - \mathbf{C}^{0(j)}\|_2.$$

Since the RHS is a constant, we know that r is upper bounded.

Let $\mathbf{b}_n$ be the unit normal vector of $\text{aff}(\mathbf{C}^0) \bigcap \text{aff}(\hat{\mathbf{C}}_n)$ on the hyperplane $\text{aff}(\mathbf{C}^0)$. Since $\mathbf{b}_n \in \text{aff}(\mathbf{C}^0)$, there exists $\boldsymbol{\lambda}_n \in \mathbb{R}^k$ and $\boldsymbol{\lambda}_n^T \mathbf{1}_k = 1$ such that $\mathbf{b}_n = \mathbf{C}^0 \boldsymbol{\lambda}_n$.

On the one hand, the variance of all $\mathbf{u}^{0(i)}$'s on the direction of $\mathbf{b}_n$ can be upper bounded:

$$Var_{\mathbf{b}_n}(\mathbf{U}^0) \leqslant \frac{1}{d}\left[(1 - C_1)d \cdot C_2^2 r^2 + C_1 d \cdot r^2\right] = \left((1 - C_1)C_2^2 + C_1\right) r^2 \tag{E.43}$$

On the other hand, since the minimum eigenvalue of $\mathbf{W}_c \mathbf{W}_c^T$ is lower bounded, we have

$$Var_{\mathbf{b}_n}(\mathbf{U}^0) \geqslant \frac{1}{d} \mathbf{b}_n^T \mathbf{U}_c \mathbf{U}_c^T \mathbf{b}_n = \frac{1}{d} \mathbf{b}_n^T \mathbf{C}^0 \mathbf{W}_c \mathbf{W}_c^T \mathbf{C}^{0T} \mathbf{b}_n$$

$$\geqslant C_3\|\mathbf{C}^{0T}\mathbf{b}_n\|_2^2 = C_3\cdot\boldsymbol{\lambda}_n^T\mathbf{C}^{0T}\mathbf{C}^0\mathbf{C}^{0T}\mathbf{C}^0\boldsymbol{\lambda}_n$$
$$\geqslant C_3\left[\sigma_{\min}^+(\mathbf{C}^0)\right]^4\|\boldsymbol{\lambda}_n\|_2^2$$
$$\geqslant C_3\left[\sigma_{\min}^+(\mathbf{C}^0)\right]^4\frac{1}{k} \tag{E.44}$$

In (E.43), by choosing the constants C_1 and C_2 small enough, we can make

$$\left((1-C_1)C_2^2+C_1\right)r^2 < C_3\left[\sigma_{\min}^+(\mathbf{C}^0)\right]^4\frac{1}{k}.$$

Therefore, we get a contradiction from (E.43) and (E.44), which finishes the proof of the claim.

As a conclusion, let $\tilde{\mathbf{u}}^{(i)} = \arg\min_{\mathbf{u}\in\mathrm{Conv}(\hat{\mathbf{C}}_n)} d(\mathbf{u},\mathbf{u}^{0(i)})$ for $i=1,\cdots,d$ and $\tilde{\mathbf{U}}_n = \{\tilde{\mathbf{u}}^{(1)},\cdots,\tilde{\mathbf{u}}^{(d)}\}$, then we have shown that w.h.p. $\|\mathbf{u}^{0(i)}-\tilde{\mathbf{u}}^{(i)}\|_2 \leqslant C\epsilon_n$. Further, by the definition of $\mathrm{Conv}(\hat{\mathbf{C}}_n)$, there exists $\hat{\mathbf{w}}^{(i)}\in\Delta^{k-1}$, such that $\tilde{\mathbf{u}}^{(i)} = \hat{\mathbf{C}}_n\hat{\mathbf{w}}^{(i)}$, for any $i=1,\cdots,d$. Let $\hat{\mathbf{W}}_n = \{\hat{\mathbf{w}}^{(1)},\cdots,\hat{\mathbf{w}}^{(d)}\}$, we have $\hat{\mathbf{C}}_n\hat{\mathbf{W}}_n = \tilde{\mathbf{U}}_n$ and

$$\|\mathbf{E}_n(:,i)\|_2 = \|\mathbf{u}^{0(i)}-\hat{\mathbf{C}}_n\hat{\mathbf{w}}^{(i)}\|_2 = \|\mathbf{u}^{0(i)}-\tilde{\mathbf{u}}^{(i)}\|_2 \leqslant C\epsilon_n.$$

□

E.4 Proof of Lemma D.4

Proof. Since $\mathbf{B}(f,:) = \lambda_f\mathbf{e}_{(f)}+\boldsymbol{\epsilon}_f, \|\boldsymbol{\epsilon}_f\|_2\leqslant\lambda_f\beta$ and $\|\mathbf{B}\|_2\leqslant M$, we can bound λ_f by C_2M.

$$M\geqslant\|\mathbf{B}\|_2\geqslant\|\mathbf{B}(f,:)\|_2\geqslant\|\lambda_f\mathbf{e}_{(f)}\|_2-\|\boldsymbol{\epsilon}_f\|_2\geqslant\lambda_f-\beta\lambda_f,\quad f=1,\cdots,k.$$

$$\lambda_f\leqslant\frac{M}{1-\beta}\leqslant C_2M,\quad f=1,\cdots,k.$$

We write $\mathbf{T} = (\lambda_1\mathbf{e}_{(1)},\cdots\lambda_k\mathbf{e}_{(k)})^T$, $\mathbf{E} = (\boldsymbol{\epsilon}_1,\cdots,\boldsymbol{\epsilon}_k)^T$, such that $\mathbf{T}+\mathbf{E} = \mathbf{B}$.

Next, we show that the column sums of $\mathbf{T}$ are close to 1, using the fact that the column sums of $\mathbf{B}$ are all 1's.

$$\begin{aligned}\sum_{s=1}^k\left|\sum_{f=1}^k\mathbf{T}(f,s)-1\right| &= \sum_{s=1}^k\left|\sum_{f=1}^k\mathbf{T}(f,s)-\sum_{f=1}^k\mathbf{B}(f,s)\right|\\ &\leqslant\sum_{s=1}^k\sum_{f=1}^k|\mathbf{B}(f,s)-\mathbf{T}(f,s)| = \sum_{f=1}^k\sum_{s=1}^k|\mathbf{B}(f,s)-\mathbf{T}(f,s)|\\ &= \sum_{f=1}^k\|\mathbf{B}(f,:)-\mathbf{T}(f,:)\|_1\leqslant\sqrt{k}\sum_{f=1}^k\|\mathbf{B}(f,:)-\mathbf{T}(f,:)\|_2\\ &= \sqrt{k}\sum_{f=1}^k\|\boldsymbol{\epsilon}_f\|_2\leqslant\sqrt{k}\sum_{f=1}^k\lambda_f\beta\leqslant C_3M\beta.\end{aligned} \tag{E.45}$$

Let $\mathbf{\Pi} = (\mathbf{e}_{(1)}, \cdots \mathbf{e}_{(k)})^T$. Then $\mathbf{\Pi}$ must be a permutation matrix. Otherwise, there exists at least one column p, such that all the entries in the p-th column of $\mathbf{\Pi}$ are 0, i.e., $\mathbf{e}_{(1),p} = \cdots \mathbf{e}_{(k),p} = 0$, where $\mathbf{e}_{(f),p}$ denotes the p-th element in $\mathbf{e}_{(f)}$. Then the sum of p-th column of $\mathbf{T}$ is 0, i.e., $\sum_{f=1}^{k} \mathbf{T}(f,p) = \sum_{f=1}^{k} \lambda_f \mathbf{e}_{(f),p} = 0$, which contradicts with (E.45).

Furthermore, since $\mathbf{T} = \mathbf{\Pi} \cdot \mathrm{diag}(\lambda_1, \cdots, \lambda_k)$ and $\mathbf{\Pi}$ is a permutation matrix, each column of $\mathbf{T}$ should include one and only one of $\lambda_1, \cdots, \lambda_k$, so that

$$\sum_{s=1}^{k} \left| \sum_{f=1}^{k} \mathbf{T}(f,s) - 1 \right| = \sum_{f=1}^{k} |\lambda_f - 1| \leqslant C_3 M\beta.$$

Consequently,

$$\|\mathbf{B} - \mathbf{\Pi}\|_2 \leqslant \|\mathbf{T} - \mathbf{\Pi}\|_2 + \|\mathbf{B} - \mathbf{T}\|_2 \leqslant \sum_{f=1}^{k} |\lambda_f - 1| + \|\mathbf{E}\|_2 \leqslant C_3' M\beta.$$

where the last inequality holds because

$$\|\mathbf{E}\|_2 \leqslant \|\mathbf{E}\|_F = \left(\sum_{f=1}^{k} \|\boldsymbol{\epsilon}_f\|_2^2 \right)^{\frac{1}{2}} \leqslant \sqrt{k} C_2 M\beta$$

□

E.5 Proof of Lemma D.5

Recall that $\epsilon_n = C_0 \sqrt{\frac{\log(n \vee d)}{n}}$ where $C_0 > 0$ is a constant. We aim to show that

$$|\det(\hat{\mathbf{C}}_n^T \hat{\mathbf{C}}_n)| \leqslant (1 + C'' \epsilon_n) |\det(\mathbf{C}^{0T} \mathbf{C}^0)|.$$

Let $\mathrm{aff}(\hat{\mathbf{C}}_n)$ and $\mathrm{aff}(\mathbf{C}^0)$ be the $(k-1)$-dim hyperplanes obtained by expanding $\mathrm{Conv}(\hat{\mathbf{C}}_n)$ and $\mathrm{Conv}(\mathbf{C}^0)$, respectively. By Lemma D.1,

$$\frac{|\det(\hat{\mathbf{C}}_n^T \hat{\mathbf{C}}_n)|}{|\det(\mathbf{C}^{0T} \mathbf{C}^0)|} = \frac{\hat{h}_n}{h^0} \cdot \frac{|\mathrm{Conv}(\hat{\mathbf{C}}_n)|}{|\mathrm{Conv}(\mathbf{C}^0)|}.$$

where $\hat{h}_n$ is the perpendicular distance from the origin to $\mathrm{aff}(\hat{\mathbf{C}}_n)$, and h^0 is the perpendicular distance from the origin to $\mathrm{aff}(\mathbf{C}^0)$.

Therefore, it suffices to show the following two inequalities,

$$\hat{h}_n \leqslant (1 + C_1 \epsilon_n) h^0, \tag{E.46}$$

and

$$|\mathrm{Conv}(\hat{\mathbf{C}}_n)| \leqslant (1 + C_2 \epsilon_n) |\mathrm{Conv}(\mathbf{C}^0)|. \tag{E.47}$$

We first prove the projection matrix associated with $\mathrm{aff}(\hat{\mathbf{C}}_n)$ converges to the one associated with $\mathrm{aff}(\mathbf{C}^0)$ in the order of ϵ_n in Lemma E.3. Then (E.46) is proved in Corollary 11.1, as a special case of Lemma E.3.

To compare $|\operatorname{Conv}(\hat{\mathbf{C}}_n)|$ and $|\operatorname{Conv}(\mathbf{C}^0)|$, we introduce three more convex polytopes:

- $\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})$, an enlarged convex polytope of $\operatorname{Conv}(\mathbf{C}^0)$ (defined in Definition 5). Here $\gamma > 0$ is a constant.
- $\operatorname{Conv}(\mathbf{C}^\sharp)$, the projection of $\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})$ on $\mathrm{aff}(\hat{\mathbf{C}}_n)$.
- $\operatorname{Conv}(\mathbf{C}^*)$, the smallest k-vertex convex polytope on $\mathrm{aff}(\hat{\mathbf{C}}_n) \bigcap \Delta^{V-1}$ containing $\mathcal{S} = \operatorname{Conv}(\hat{\mathbf{C}}_n) \bigcap \left\{ \bigcup_{i=1}^{d} \mathcal{B}(\mathbf{u}^{0(i)}; C_4\epsilon_n) \right\}$. Here $\mathcal{B}(\mathbf{u}^{0(i)}; C_4\epsilon_n)$ is the Euclidean ball centered at $\mathbf{u}^{0(i)}$ with radius $C_4\epsilon_n$. The formal definition is given in Definition 7.

We then prove (E.47) by the following steps.

1. In Lemma E.4 and Lemma E.5, we show
$$\left(1 - \frac{1}{n}\right) |\operatorname{Conv}(\hat{\mathbf{C}}_n)| \leqslant |\operatorname{Conv}(\mathbf{C}^*)|. \tag{E.48}$$
2. In Lemma E.6 to Lemma E.8, we show that $\operatorname{Conv}(\mathbf{C}^\sharp)$ is a k-vertex convex polytope within Δ^{V-1} containing $\mathcal{S}$. Therefore, by the definition of $\operatorname{Conv}(\mathbf{C}^*)$, we have
$$|\operatorname{Conv}(\mathbf{C}^*)| \leqslant |\operatorname{Conv}(\mathbf{C}^\sharp)|. \tag{E.49}$$
3. In Lemma E.9, we prove (E.47) by summarizing the the above inequalities, i.e.,
$$\left(1 - \frac{1}{n}\right) |\operatorname{Conv}(\hat{\mathbf{C}}_n)| \overset{\text{(E.48)}}{\leqslant} |\operatorname{Conv}(\mathbf{C}^*)| \overset{\text{(E.49)}}{\leqslant} |\operatorname{Conv}(\mathbf{C}^\sharp)| \overset{\text{Definition of } \operatorname{Conv}(\mathbf{C}^\sharp)}{\leqslant} |\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})|$$
$$\overset{\text{Proposition 11}}{\leqslant} \left(1 + \rho(\mathbf{C}^0)\gamma\epsilon_n\right)^{k-1} |\operatorname{Conv}(\mathbf{C}^0)|.$$

Next, we provide detailed proof.

First, we show that the projection matrix and any projected vector of aff($\hat{\mathbf{C}}_n$) converges to the ones of aff($\mathbf{C}^0$) in the order of $\sqrt{\frac{\log(n \vee d)}{n}}$.

Let $(\mathbf{u}^{(1)}, \cdots, \mathbf{u}^{(k)})$ be any k linearly independent vectors from aff($\mathbf{C}$). Then, the **projection matrix** of aff($\mathbf{C}$) can be written as

$$\mathbf{P}_{\mathbf{C}} = \mathbf{U}'(\mathbf{U}'^T\mathbf{U}')^{-1}\mathbf{U}'^T, \tag{E.50}$$

where $\mathbf{U}' = \left(\mathbf{u}^{(2)} - \mathbf{u}^{(1)}, \cdots, \mathbf{u}^{(k)} - \mathbf{u}^{(1)}\right)$.

For any vector $\mathbf{y}$, its **projection onto aff($\mathbf{C}$)** is given by

$$\hat{\mathbf{y}}_{\mathbf{C}} = \mathbf{P}_{\mathbf{C}}(\mathbf{y} - \mathbf{u}^{(1)}) + \mathbf{u}^{(1)} = \mathbf{P}_{\mathbf{C}}\mathbf{y} + (\mathbf{I} - \mathbf{P}_{\mathbf{C}})\mathbf{u}^{(1)}. \tag{E.51}$$

Lemma E.3.

$$\|\mathbf{P}_{\hat{\mathbf{C}}_n} - \mathbf{P}_{\mathbf{C}^0}\|_2 \leqslant C\epsilon_n \tag{E.52}$$

$$\|\hat{\mathbf{y}}_{\hat{\mathbf{C}}_n} - \hat{\mathbf{y}}_{\mathbf{C}^0}\|_2 \leqslant C\|\mathbf{y}\|_2\epsilon_n + C'\epsilon_n, \tag{E.53}$$

for any $\mathbf{y} \in \mathbb{R}^V$, where C and C' are positive constants.

Proof. By assumption (A3), let $(i_1, \cdots, i_k)$ denote the index set of the columns of $\mathbf{W}^{0*}$ in $\mathbf{W}^0$, where the k columns of $\mathbf{W}^{0*}$ are affinely independent and have minimum positive singular value lower bounded. Let $\mathbf{U}^{0*} = \mathbf{C}^0\mathbf{W}^{0*} = (\mathbf{u}^{0(i_1)}, \cdots, \mathbf{u}^{0(i_k)})$ and $\tilde{\mathbf{U}}_n^* = (\tilde{\mathbf{u}}^{(i_1)}, \cdots, \tilde{\mathbf{u}}^{(i_k)})$, where $\tilde{\mathbf{u}}^{(i)} = \arg\min_{\mathbf{u} \in \mathrm{Conv}(\hat{\mathbf{C}}_n)} d(\mathbf{u}, \mathbf{u}^{0(i)})$ is the projection of $\mathbf{u}^{0(i)}$ onto Conv($\hat{\mathbf{C}}_n$).

By Lemma E.2, we have

$$\|\mathbf{U}^{0*} - \tilde{\mathbf{U}}_n^*\|_2 \leqslant \|\mathbf{U}^{0*} - \tilde{\mathbf{U}}_n^*\|_F = \left(\sum_{j=1}^{k} \|\mathbf{u}^{0(i_j)} - \tilde{\mathbf{u}}^{(i_j)}\|_2^2\right)^{\frac{1}{2}} \leqslant C_2\epsilon_n,$$

and

$$\|\mathbf{u}^{0(i_1)} - \tilde{\mathbf{u}}^{(i_1)}\|_2 \leqslant C_3\epsilon_n.$$

By (E.50), we have

$$\mathbf{P}_{\mathbf{C}^0} = \mathbf{U}^{0\prime}(\mathbf{U}^{0\prime T}\mathbf{U}^{0\prime})^{-1}\mathbf{U}^{0\prime T}, \quad \mathbf{P}_{\hat{\mathbf{C}}_n} = \tilde{\mathbf{U}}'_n(\tilde{\mathbf{U}}'^T_n\tilde{\mathbf{U}}'_n)^{-1}\tilde{\mathbf{U}}'^T_n$$

where $\mathbf{U}^{0\prime} = \mathbf{U}^{0*}\mathbf{Q}$, $\tilde{\mathbf{U}}'_n = \tilde{\mathbf{U}}_n^*\mathbf{Q}$, and $\mathbf{Q}_{k\times(k-1)} = \begin{bmatrix} -\mathbf{1}_{k-1} & \mathbf{I}_{k-1} \end{bmatrix}^T$.

By Weyl's inequality in matrix theory (Weyl, 1912),

$$\sigma^+_{\min}(\mathbf{U}^{0\prime}) - \sigma^+_{\min}(\tilde{\mathbf{U}}'_n) \leqslant \|\mathbf{U}^{0\prime} - \tilde{\mathbf{U}}'_n\|_2 \leqslant \|\mathbf{U}^{0*} - \tilde{\mathbf{U}}_n^*\|_2\|\mathbf{Q}\|_2 \leqslant C'_2\epsilon_n.$$

Therefore,

$$\sigma^+_{\min}(\tilde{\mathbf{U}}'_n) \geqslant \sigma^+_{\min}(\mathbf{U}^{0\prime}) - C'_2\epsilon_n \geqslant \frac{\sigma^+_{\min}(\mathbf{U}^{0\prime})}{2}. \tag{E.54}$$

Moreover,

$$\sigma^+_{\min}(\mathbf{U}^{0\prime}) = \sigma^+_{\min}(\mathbf{U}^{0*}\mathbf{Q}) = \sigma^+_{\min}(\mathbf{C}^0\mathbf{W}^{0*}\mathbf{Q}) \geqslant \sigma^+_{\min}(\mathbf{C}^0)\sigma^+_{\min}(\mathbf{W}^{0*})\sigma^+_{\min}(\mathbf{Q}) \geqslant C_3. \tag{E.55}$$

So the columns of $\tilde{\mathbf{U}}^*_n$ are also affinely independent.

According to Davis-Kahan theorem (Chen et al., 2016; Davis and Kahan, 1970), we have

$$\begin{aligned}
\|\mathbf{P}_{\hat{\mathbf{C}}_n} - \mathbf{P}_{\mathbf{C}^0}\|_2 &\overset{Davis-Kahan}{\leqslant} \max\left(\frac{1}{\sigma^+_{\min}(\tilde{\mathbf{U}}'_n)}, \frac{1}{\sigma^+_{\min}(\mathbf{U}^{0\prime})}\right)\|\tilde{\mathbf{U}}'_n - \mathbf{U}^{0\prime}\|_2 \\
&\leqslant \max\left(\frac{1}{\sigma^+_{\min}(\tilde{\mathbf{U}}'_n)}, \frac{1}{\sigma^+_{\min}(\mathbf{U}^{0\prime})}\right)\|\tilde{\mathbf{U}}^*_n - \mathbf{U}^{0*}\|_2\|\mathbf{Q}\|_2 \\
&\leqslant C_4\epsilon_n,
\end{aligned}$$

where the last inequality is due to (E.54) and (E.55).

Finally, for any $\mathbf{y} \in \mathbb{R}^V$,

$$\begin{aligned}
\|\hat{\mathbf{y}}_{\hat{\mathbf{C}}_n} - \hat{\mathbf{y}}_{\mathbf{C}^0}\|_2 &\leqslant \|\mathbf{P}_{\hat{\mathbf{C}}_n} - \mathbf{P}_{\mathbf{C}^0}\|_2\|\mathbf{y}\|_2 + \|\mathbf{P}_{\hat{\mathbf{C}}_n} - \mathbf{P}_{\mathbf{C}^0}\|_2\|\mathbf{u}^{0(i_1)}\|_2 + \|\mathbf{u}^{0(i_1)} - \tilde{\mathbf{u}}^{(i_1)}\|_2 \\
&\leqslant C\|\mathbf{y}\|_2\epsilon_n + C'\epsilon_n
\end{aligned}$$

□

Corollary 11.1. *Denote the perpendicular distance between origin, $\mathbf{0} = (0, 0, \cdots, 0)$, and aff$(\mathbf{C}^0)$ by h^0, and the perpendicular distance between origin and aff$(\hat{\mathbf{C}}_n)$ by $\hat{h}_n$. The followings hold,*

1. $|\hat{h}_n - h^0| \leqslant C'\epsilon_n$,
2. $h_0 > C''$.

where C' and C'' are positive constants.

Proof. The perpendicular distance of aff$(\mathbf{C})$ is the length of the projected vector of $\mathbf{0}$ on aff$(\mathbf{C})$. Specifically,

$$\hat{h}_n = \|\hat{\mathbf{0}}_{\hat{\mathbf{C}}_n}\|_2, \quad h^0 = \|\hat{\mathbf{0}}_{\mathbf{C}^0}\|_2,$$

Therefore,

$$\left|\hat{h}_n - h^0\right| = \left|\|\hat{\mathbf{0}}_{\hat{\mathbf{C}}_n}\|_2 - \|\hat{\mathbf{0}}_{\mathbf{C}^0}\|_2\right| \leqslant \|\hat{\mathbf{0}}_{\hat{\mathbf{C}}_n} - \hat{\mathbf{0}}_{\mathbf{C}^0}\|_2 \leqslant (C\|\mathbf{0}\|_2\epsilon_n + C'\epsilon_n) \leqslant C'\epsilon_n.$$

Furthermore, since $\hat{\mathbf{0}}_{\mathbf{C}^0}$ is on $\mathrm{aff}(\mathbf{C}^0)$, we can represent $\hat{\mathbf{0}}_{\mathbf{C}^0}$ by $\mathbf{C}^0\mathbf{w}_h$ for some $\mathbf{w}_h \in \Delta^{k-1}$.

$$h^0 = \|\hat{\mathbf{0}}_{\mathbf{C}^0}\|_2 \geqslant \sigma^+_{\min}(\mathbf{C}^0)\|\mathbf{w}_h\|_2 \geqslant \sigma^+_{\min}(\mathbf{C}^0)\|\mathbf{w}_h\|_1/\sqrt{k} = \sigma^+_{\min}(\mathbf{C}^0)/\sqrt{k}.$$

□

With the result from Lemma E.2, in the following Lemma E.4, we show that most of the mass of $f^{(i)}(\mathbf{u})$ on $\mathrm{Conv}(\hat{\mathbf{C}}_n)$ is concentrated on $\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}(\mathbf{u}^{0(i)}; C_4\epsilon_n)$.

Lemma E.4. *For any $i \in [d]$,*

$$\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}(\mathbf{u}^{0(i)};C_4\epsilon_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u} \geqslant (1-\frac{1}{n})\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}.$$

Proof. It suffices to show that for any $i \in [d]$,

$$\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}^C(\mathbf{u}^{0(i)};C_4\epsilon_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u} \leqslant \frac{1}{n}\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}. \tag{E.56}$$

For the LHS of (E.56),

$$\begin{aligned}
&\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}^C(\mathbf{u}^{0(i)};C_4\epsilon_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
\leqslant &\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}^C(\mathbf{u}^{0(i)};C_4\epsilon_n)} \exp\left(-\frac{n}{2}\|\hat{\mathbf{u}}^{(i)}-\mathbf{u}\|_2^2\right)d\mathbf{u}\\
\leqslant &\exp\left(-\frac{n}{4}(C_4-1)^2\epsilon_n^2\right)\int_{\mathrm{aff}(\hat{\mathbf{C}}_n)} \exp\left(-\frac{n}{4}\|\hat{\mathbf{u}}^{(i)}-\mathbf{u}\|_2^2\right)d\mathbf{u}\\
= &\exp\left(-\frac{n}{4}(C_4-1)^2\epsilon_n^2\right)\exp\left(-\frac{n}{4}d^2(\hat{\mathbf{u}}^{(i)},\mathrm{aff}(\hat{\mathbf{C}}_n))\right)\int_{\mathrm{aff}(\hat{\mathbf{C}}_n)} \exp\left(-\frac{n}{4}\|\hat{\mathbf{u}}^{(i)}_{\hat{\mathbf{C}}_n}-\mathbf{u}\|_2^2\right)d\mathbf{u}\\
\leqslant &\exp\left(-\frac{n}{4}(C_4-1)^2\epsilon_n^2\right)\cdot 1\cdot \frac{C_5}{n^{\frac{k-1}{2}}},
\end{aligned}$$

where the first inequality is due to the Pinsker's inequality (E.34), the third inequality is from the normalizing constant for a multivariate Gaussian distribution.

For the integration in the RHS of (E.56),

$$\begin{aligned}
&\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
\geqslant &\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}(\mathbf{u}^{0(i)};\sqrt{2}C\epsilon_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
\geqslant &\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}(\mathbf{u}^{0(i)};\sqrt{2}C\epsilon_n)} \exp\left(-C_7 n\|\hat{\mathbf{u}}^{(i)}-\mathbf{u}\|_2^2\right)d\mathbf{u}\\
\geqslant &C_8(C\epsilon_n)^{k-1}\cdot\exp\left(-C_7 n(\sqrt{2}C+1)^2\epsilon_n^2\right)
\end{aligned}$$

where C is the constant from (E.42). Since $\mathbf{u}^{0(i)}$'s are interior points in Δ^{V-1}, when n is large enough, the second inequality follows from the reverse Pinsker's inequality (E.35).

By choosing C_4 large enough, we can ensure

$$\exp\left(-\frac{n}{4}(C_4-1)^2\epsilon_n^2\right)\frac{C_5}{n^{\frac{k-1}{2}}} \leqslant \frac{1}{n}\cdot C_8 C^{k-1}\epsilon_n^{k-1}\exp\left(-C_7 n(\sqrt{2}C+1)^2\epsilon_n^2\right)$$

Consequently, we have

$$\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap \mathcal{B}^C(\mathbf{u}^{0(i)};C_4\epsilon_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u} \leqslant \frac{1}{n}\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)} \frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}.$$

□

Definition 7. *Define* $\mathbf{C}^* = [\mathbf{c}_1^*,\cdots,\mathbf{c}_k^*] \in \mathbb{R}^{V\times k}$, *such that,* $\mathbf{c}_f^* \in$ *aff*$(\hat{\mathbf{C}}_n)\bigcap\Delta^{V-1}$, $\forall f\in[k]$, *and* $\mathrm{Conv}(\mathbf{C}^*)$ *is* **the smallest (volume) convex polytope with** k **vertices on aff**$(\hat{\mathbf{C}}_n)\bigcap\Delta^{V-1}$ *that contains the set* $\mathcal{S} = \mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap\left\{\bigcup_{i=1}^d \mathcal{B}(\mathbf{u}^{0(i)};C_4\epsilon_n)\right\}$.

Note that $\mathrm{Conv}(\hat{\mathbf{C}}_n)$ is a convex polytope with k vertices on $\mathrm{aff}(\hat{\mathbf{C}}_n)\bigcap\Delta^{V-1}$ containing $\mathcal{S}$. So $\mathbf{C}^*$ must exist and it satisfies $|\mathrm{Conv}(\mathbf{C}^*)| \leqslant |\mathrm{Conv}(\hat{\mathbf{C}}_n)|$. In the following lemma, we show that $|\mathrm{Conv}(\hat{\mathbf{C}}_n)|$ cannot be much larger than $|\mathrm{Conv}(\mathbf{C}^*)|$.

Lemma E.5.

$$\left(1-\frac{1}{n}\right)|\mathrm{Conv}(\hat{\mathbf{C}}_n)| \leqslant |\mathrm{Conv}(\mathbf{C}^*)|$$

Proof. Since $\mathbf{c}_f^* \in \Delta^{V-1}$, $\forall f\in[k]$, $\mathbf{C}^*$ is a valid parameter of $F_{n\times d}(\mathbf{C};\mathbf{X})$, and $F_{n\times d}(\hat{\mathbf{C}}_n;\mathbf{X}) \geqslant F_{n\times d}(\mathbf{C}^*;\mathbf{X})$. From Lemma E.4, we have

$$\begin{aligned}
&\frac{A_{n,d}}{|\mathrm{Conv}(\hat{\mathbf{C}}_n)|^d}\cdot\prod_{i=1}^d\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap\mathcal{B}(\mathbf{u}^{0(i)};C_4\epsilon_n)}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
\geqslant&\left(1-\frac{1}{n}\right)^d F_{n\times d}(\hat{\mathbf{C}}_n;\mathbf{X}) \geqslant \left(1-\frac{1}{n}\right)^d F_{n\times d}(\mathbf{C}^*;\mathbf{X})\\
=&\left(1-\frac{1}{n}\right)^d\frac{A_{n,d}}{|\mathrm{Conv}(\mathbf{C}^*)|^d}\cdot\prod_{i=1}^d\int_{\mathrm{Conv}(\mathbf{C}^*)}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u}\\
\geqslant&\left(1-\frac{1}{n}\right)^d\frac{A_{n,d}}{|\mathrm{Conv}(\mathbf{C}^*)|^d}\cdot\prod_{i=1}^d\int_{\mathrm{Conv}(\hat{\mathbf{C}}_n)\bigcap\mathcal{B}(\mathbf{u}^{0(i)};C_4\epsilon_n)}\frac{f^{(i)}(\mathbf{u})}{f^{(i)}(\hat{\mathbf{u}}^{(i)})}d\mathbf{u},
\end{aligned}$$

where the last inequality is due to the definition of $\mathrm{Conv}(\mathbf{C}^*)$. Therefore,

$$\frac{1}{|\mathrm{Conv}(\hat{\mathbf{C}}_n)|^d} \geqslant \frac{\left(1-\frac{1}{n}\right)^d}{|\mathrm{Conv}(\mathbf{C}^*)|^d},$$

$$\left(1-\frac{1}{n}\right)|\mathrm{Conv}(\hat{\mathbf{C}}_n)| \leqslant |\mathrm{Conv}(\mathbf{C}^*)|.$$

□

Next, we compare $|\operatorname{Conv}(\mathbf{C}^0)|$ and $|\operatorname{Conv}(\mathbf{C}^*)|$. We will construct an enlarged convex polytope, $\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})$, and then project it to $\operatorname{aff}(\hat{\mathbf{C}}_n)$ to obtain a projected convex polytope, $|\operatorname{Conv}(\mathbf{C}^\sharp)|$. We will show that $|\operatorname{Conv}(\mathbf{C}^\sharp)|$ contains the set $\mathcal{S}$, so that $|\operatorname{Conv}(\mathbf{C}^*)| \leqslant |\operatorname{Conv}(\mathbf{C}^\sharp)|$.

Definition 8. *Let $\mathbf{C}^\sharp = (\mathbf{c}_1^\sharp, \cdots, \mathbf{c}_k^\sharp) \in \mathbb{R}^{V\times k}$ such that $\mathbf{c}_f^\sharp$ is the projected vector of the f-th vertex of $\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})$ on aff$(\hat{\mathbf{C}}_n)$, $\forall f = [k]$. Here $\gamma > 0$ is a constant.*

Lemma E.6. *When n is large enough, $\operatorname{Conv}(\mathbf{C}^\sharp)$ is in Δ^{V-1}.*

Proof. It suffices to show that for any $f \in [k]$, (1) $\mathbf{c}_f^{\sharp T}\mathbf{1}_V = 1$ and (2) $\mathbf{c}_f^\sharp \geqslant 0$.

By the definition of aff$(\hat{\mathbf{C}}_n)$, $\mathbf{c}_f^\sharp = \hat{\mathbf{C}}_n\boldsymbol{\lambda}_f$ and $\boldsymbol{\lambda}_f^T\mathbf{1}_k = 1$. Therefore, (1) holds because

$$\mathbf{c}_f^{\sharp T}\mathbf{1}_V = \boldsymbol{\lambda}_f^T\hat{\mathbf{C}}_n^T\mathbf{1}_V = \boldsymbol{\lambda}_f^T\mathbf{1}_k = 1.$$

By Lemma E.3, we have

$$\|\mathbf{c}_f^\sharp - (\mathbf{c}_f^0)^{\gamma\epsilon_n}\|_2 \leqslant C\epsilon_n.$$

Therefore, to show (2), it suffices to verify that $(\mathbf{c}_f^0)^{\gamma\epsilon_n} > C_1$, for any $f \in [k]$.

Note that $\mathbf{c}_1^0, \cdots, \mathbf{c}_k^0 \geqslant C_2$, since $\mathbf{c}_1^0, \cdots, \mathbf{c}_k^0$ are strict inner points of Δ^{V-1}.

By the definition of the enlarged convex polytope (Definition 5), we have

$$\begin{aligned}(\mathbf{c}_f^0)^{\gamma\epsilon_n} &= (1+\rho\gamma\epsilon_n)(\mathbf{c}_f^0 - \bar{\mathbf{c}}^0) + \bar{\mathbf{c}}^0\\ &= (1+\rho\gamma\epsilon_n)\mathbf{c}_f^0 - \rho\gamma\epsilon_n\bar{\mathbf{c}}^0\\ &\geqslant (1+\rho\gamma\epsilon_n)C_2 - \rho\gamma\epsilon_n \geqslant C_1,\end{aligned}$$

where $\rho = \rho(\mathbf{C}^0)$ and $\bar{\mathbf{c}}^0 = \frac{1}{k}\sum_{f=1}^k \mathbf{c}_f^0$. □

Next, we want to prove that $\operatorname{Conv}(\mathbf{C}^\sharp)$ contains the set $\mathcal{S}$. We first study the property of the boundary points of the δ-enlargement convex polytope in Lemma E.7 and show that the distance between any boundary point and the original convex polytope is at least δ. Using this fact, we know that any boundary point of $\operatorname{Conv}(\mathbf{C}^\sharp)$ is at least $\gamma\epsilon_n$ away from any $\mathbf{u}^{0(i)} \in \operatorname{Conv}(\mathbf{C}^0)$. By letting γ large enough, we can have $\operatorname{Conv}(\mathbf{C}^\sharp)$ contain the set $\mathcal{S}_i' = \mathcal{B}(\mathbf{u}^{0(i)}, C_4\epsilon_n)\bigcap \operatorname{aff}(\hat{\mathbf{C}}_n)$ for any $i \in [d]$. Therefore, $\operatorname{Conv}(\mathbf{C}^\sharp)$ contains the set $\mathcal{S}' = \bigcup_{i=1}^d \mathcal{S}_i'$, which is a superset of the set $\mathcal{S}$. The detailed proof is in Lemma E.8.

Lemma E.7. *For any point* $\mathbf{x}$ *on the boundary of* $\mathrm{Conv}(\mathbf{C}^\delta)$,

$$\delta \leqslant d(\mathbf{x}, \mathrm{Conv}(\mathbf{C})) \leqslant \kappa(\mathbf{C}) k\delta.$$

where $\kappa(\mathbf{C}) = \frac{\sigma_{\max}(\mathbf{C})}{\sigma^+_{\min}(\mathbf{C})}$ *is the conditional number of* $\mathbf{C}$.

Proof. By the definition of δ-enlargement in Definition 5,

$$\begin{aligned}
\mathbf{x} &= \sum_{f=1}^{k} \alpha_f \mathbf{c}_f^\delta = \sum_{f=1}^{k} \alpha_f [(1+\rho\delta)(\mathbf{c}_f - \bar{\mathbf{c}}) + \bar{\mathbf{c}}] \\
&= \sum_{f=1}^{k} \alpha_f (1+\rho\delta)[\mathbf{c}_f - \frac{1}{k}\sum_{s=1}^{k} \mathbf{c}_s] + \sum_{f=1} \alpha_f \bar{\mathbf{c}} \\
&= \sum_{f=1}^{k} \alpha_f (1+\rho\delta)\mathbf{c}_f - \frac{1+\rho\delta}{k}\sum_{s=1}^{k} \mathbf{c}_s + \bar{\mathbf{c}} \\
&= \sum_{f=1}^{k} (1+\rho\delta)\left(\alpha_f - \frac{1}{k}\right)\mathbf{c}_f + \bar{\mathbf{c}}
\end{aligned}$$

where $\rho = \frac{k}{\sigma^+_{\min}(\mathbf{C})}$, and $\boldsymbol{\alpha} = \{\alpha_f\}_{f=1,\cdots,k} \in \Delta^{k-1}$. Since $\mathbf{x}$ is a boundary point, there exists at least one $f \in [k]$, such that $\alpha_f = 0$. WLOG, we assume $\alpha_k = 0$.

Let $\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \frac{1}{k}\mathbf{1}$. We have

$$\mathbf{x} = \mathbf{C}(1+\rho\delta)\boldsymbol{\alpha}' + \bar{\mathbf{c}}.$$

At the same time, any point $\mathbf{y}$ in $\mathrm{Conv}(\mathbf{C})$ can be represented by

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} = \sum_{f=1}^{k} \beta_f(\mathbf{c}_f - \bar{\mathbf{c}}) + \bar{\mathbf{C}} = \mathbf{C}\boldsymbol{\beta}' + \bar{\mathbf{c}},$$

where $\boldsymbol{\beta} = \{\beta_f\}_{f=1,\cdots,k} \in \Delta^{k-1}$ and $\boldsymbol{\beta}' = \boldsymbol{\beta} - \frac{1}{k}\mathbf{1}$.

Now we can measure the distance between the boundary point $\mathbf{x}$ and $\mathrm{Conv}(\mathbf{C})$,

$$\begin{aligned}
d(\mathbf{x}, \mathrm{Conv}(\mathbf{C})) &= \min_{\mathbf{y}\in\mathrm{Conv}(\mathbf{C})} \|\mathbf{x}-\mathbf{y}\|_2 = \min_{\boldsymbol{\beta}\in\Delta^{V-1}} \|\mathbf{C}[(1+\rho\delta)\boldsymbol{\alpha}' - \boldsymbol{\beta}']\|_2 \\
&\geqslant \sigma^+_{\min}(\mathbf{C}) \cdot \min_{\boldsymbol{\beta}\in\Delta^{V-1}} \|(1+\rho\delta)\boldsymbol{\alpha}' - \boldsymbol{\beta}'\|_2.
\end{aligned}$$

Write $\boldsymbol{\eta} = (1+\rho\delta)\boldsymbol{\alpha}' - \boldsymbol{\beta}' = (1+\rho\delta)\boldsymbol{\alpha} - \boldsymbol{\beta} - \frac{\rho\delta}{k}\mathbf{1}$. Then, the k-th element of $\boldsymbol{\eta}$ is

$$\eta_k = (1+\rho\delta)\alpha_k - \beta_k - \frac{\rho\delta}{k} = 0 - \beta_k - \frac{\rho\delta}{k} \leqslant -\frac{\rho\delta}{k}$$

because we assume $\alpha_k = 0$, and $\boldsymbol{\beta} \in \Delta^{(k-1)}$ so that $\beta_k \geqslant 0$. Then, we obtain the lower bound,

$$d(\mathbf{x}, \mathrm{Conv}(\mathbf{C})) \geqslant \sigma^+_{\min}(\mathbf{C}) \cdot \min_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2 \geqslant \sigma^+_{\min}(\mathbf{C}) \cdot \min_{\boldsymbol{\eta}} |\eta_k| \geqslant \sigma^+_{\min}(\mathbf{C}) \cdot \frac{\rho\delta}{k} = \delta.$$

For the upper bound, let $\boldsymbol{\beta}' = \boldsymbol{\alpha}'$, we have

$$\begin{aligned}\min_{\mathbf{y}\in\mathrm{Conv}(\mathbf{C})} \|\mathbf{x}-\mathbf{y}\|_2 &\leqslant \|\mathbf{C}[(1+\rho\delta)\boldsymbol{\alpha}' - \boldsymbol{\alpha}']\|_2 \leqslant \sigma_{\max}(\mathbf{C})\cdot\rho\delta\cdot\|\boldsymbol{\alpha}'\|_2 \\ &= \sigma_{\max}(\mathbf{C})\rho\delta\sqrt{\|\boldsymbol{\alpha}\|_2^2 - \frac{1}{k}} \leqslant \sigma_{\max}(\mathbf{C})\rho\delta\sqrt{1-\frac{1}{k}} \\ &\leqslant \kappa(\mathbf{C})k\delta.\end{aligned}$$

□

Lemma E.8. *For some $\gamma > 0$, $\mathrm{Conv}(\mathbf{C}^\sharp)$ covers the set $\mathcal{S}' = \left\{\bigcup_{i=1}^d \mathcal{B}(\mathbf{u}^{0(i)}, C_4\epsilon_n)\right\} \bigcap aff(\hat{\mathbf{C}}_n)$, i.e.,*

$$\mathcal{S} \subseteq \mathcal{S}' \subseteq \mathrm{Conv}(\mathbf{C}^\sharp).$$

Proof. Since $\mathrm{Conv}(\mathbf{C}^\sharp)$ is a closed and simply connected region, it suffices to show that for any boundary point, $\mathbf{x} \in bd\,\mathrm{Conv}(\mathbf{C}^\sharp)$,

$$\min_{i=1\cdots,d} \|\mathbf{x} - \mathbf{u}^{0(i)}\|_2 \geqslant C_4\epsilon_n.$$

In fact, any boundary point of $\mathrm{Conv}(\mathbf{C}^\sharp)$, $\mathbf{x} \in bd\,\mathrm{Conv}(\mathbf{C}^\sharp)$, is projected from a boundary point of $\mathrm{Conv}(\mathbf{C}^0)^{\gamma\epsilon_n}$, denoted as $\mathbf{y} \in bd(\mathrm{Conv}(\mathbf{C}^0)^{\gamma\epsilon_n})$. Then,

$$\mathbf{x} = \mathbf{P}_{\tilde{\mathbf{U}}_n}\mathbf{y} + (\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{U}}_n})\tilde{\mathbf{u}}^{(k)}.$$

By Lemma E.7, we have $\forall \mathbf{y} \in bd(\mathrm{Conv}(\mathbf{C}^0)^{\gamma\epsilon_n})$,

$$\|\mathbf{y} - \mathbf{u}^{0(i)}\|_2 \geqslant d(\mathbf{y}, \mathrm{Conv}(\mathbf{C}^0)) \geqslant \gamma\epsilon_n. \quad \forall i = 1, \cdots, d. \tag{E.57}$$

Denote the projected point of $\mathbf{u}^{0(i)}$ on $\mathrm{aff}(\hat{\mathbf{C}}_n) = \mathrm{aff}(\tilde{\mathbf{U}}_n)$, as $\hat{\mathbf{u}}^{0(i)}_{\tilde{\mathbf{U}}_n}$. We have

$$\begin{aligned}\|\mathbf{x} - \mathbf{u}^{0(i)}\|_2 =& \|(\mathbf{x} - \hat{\mathbf{u}}^{0(i)}_{\tilde{\mathbf{U}}_n}) + (\hat{\mathbf{u}}^{0(i)}_{\tilde{\mathbf{U}}_n} - \mathbf{u}^{0(i)})\|_2 \\ \geqslant& \|\mathbf{x} - \hat{\mathbf{u}}^{0(i)}_{\tilde{\mathbf{U}}_n}\|_2 = \|\mathbf{P}_{\tilde{\mathbf{U}}_n}(\mathbf{y} - \mathbf{u}^{0(i)})\|_2 \\ =& \|\mathbf{P}_{\mathbf{U}^0}(\mathbf{y} - \mathbf{u}^{0(i)}) + (\mathbf{P}_{\tilde{\mathbf{U}}_n} - \mathbf{P}_{\mathbf{U}^0})(\mathbf{y} - \mathbf{u}^{0(i)})\|_2 \\ \geqslant& \|\mathbf{y} - \mathbf{u}^{0(i)}\|_2 - \|\mathbf{P}_{\tilde{\mathbf{U}}_n} - \mathbf{P}_{\mathbf{U}^0}\|_2\|\mathbf{y} - \mathbf{u}^{0(i)}\|_2 \\ \geqslant& (1 - C_5\epsilon_n)\|\mathbf{y} - \mathbf{u}^{0(i)}\|_2 && \text{by Lemma E.3 (E.52)} \\ \geqslant& (1 - C_5\epsilon_n)\gamma\epsilon_n \quad \forall i = 1, \cdots, d. && \text{by (E.57)}\end{aligned}$$

By letting γ large enough, we can make that for any $\mathbf{x} \in bd\,\mathrm{Conv}(\mathbf{C}^\sharp)$,

$$\|\mathbf{x} - \mathbf{u}^{0(i)}\|_2 \geqslant (1 - C_5\epsilon_n)\gamma\epsilon_n \geqslant C_4\epsilon_n \quad \forall i = 1, \cdots, d.$$

□

Now we are ready to piece together all the useful results and conclude the following inequalities.

Lemma E.9.

$$|\operatorname{Conv}(\hat{\mathbf{C}}_n)| \leqslant (1 + C'\epsilon_n)\,|\operatorname{Conv}(\mathbf{C}^0)|$$

$$|\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)| \leqslant (1 + C''\epsilon_n)|\det({\mathbf{C}^0}^T\mathbf{C}^0)|$$

Proof. Since the δ-enlargement is an affine transformation, by Proposition 11,

$$|\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})| = \big(1 + \rho(\mathbf{C}^0)\gamma\epsilon_n\big)^{k-1}\,|\operatorname{Conv}(\mathbf{C}^0)| \leqslant (1 + C_1\epsilon_n)|\operatorname{Conv}(\mathbf{C}^0)|.$$

Then, by Lemma E.6 and Lemma E.8, $\operatorname{Conv}(\mathbf{C}^\sharp)$ is on $\operatorname{aff}(\hat{\mathbf{C}}_n)\bigcap\Delta^{V-1}$ covering $\mathcal{S}$. And by Definition 7, $\operatorname{Conv}(\mathbf{C}^*)$ is the smallest k-vertex convex polytope on $\operatorname{aff}(\hat{\mathbf{C}}_n)\bigcap\Delta^{V-1}$ covering $\mathcal{S}$. So we have ,

$$|\operatorname{Conv}(\mathbf{C}^*)| \leqslant |\operatorname{Conv}(\mathbf{C}^\sharp)| \leqslant |\operatorname{Conv}((\mathbf{C}^0)^{\gamma\epsilon_n})|.$$

The last inequality holds because $\mathbf{C}^\sharp$ is a projection of $(\mathbf{C}^0)^{\gamma\epsilon_n}$ on $\operatorname{aff}(\hat{\mathbf{C}}_n)$, so that any side length of $\operatorname{Conv}(\mathbf{C}^\sharp)$ is shorter than the corresponding one of $(\mathbf{C}^0)^{\gamma\epsilon_n}$, i.e.,

$$\|\mathbf{c}_f^\sharp - \mathbf{c}_p^\sharp\|_2 = \|\mathbf{P}_{\tilde{\mathbf{U}}_n}[(\mathbf{c}_f^0)^{\gamma\epsilon_n} - (\mathbf{c}_p^0)^{\gamma\epsilon_n}]\|_2 \leqslant \|(\mathbf{c}_f^0)^{\gamma\epsilon_n} - (\mathbf{c}_p^0)^{\gamma\epsilon_n}\|_2$$

Together with Lemma E.5, we obtain,

$$\begin{aligned}
|\operatorname{Conv}(\hat{\mathbf{C}}_n)| &\leqslant \left(1 + \frac{1}{n-1}\right)|\operatorname{Conv}(\mathbf{C}^*)| \\
&\leqslant \left(1 + \frac{1}{n-1}\right)(1 + C_1\epsilon_n)\,|\operatorname{Conv}(\mathbf{C}^0)| \\
&\leqslant (1 + C'\epsilon_n)\,|\operatorname{Conv}(\mathbf{C}^0)|.
\end{aligned}$$

Furthermore, by Lemma D.1, we have

$$\frac{|\operatorname{Conv}(\hat{\mathbf{C}}_n)|}{|\operatorname{Conv}(\mathbf{C}^0)|} = \frac{h^0}{\hat{h}_n}\cdot\frac{\sqrt{\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)}}{\sqrt{\det\left({\mathbf{C}^0}^T\mathbf{C}^0\right)}} \leqslant 1 + C'\epsilon_n.$$

By Corollary 11.1, we obtain

$$\frac{\sqrt{\det(\hat{\mathbf{C}}_n^T\hat{\mathbf{C}}_n)}}{\sqrt{\det\left({\mathbf{C}^0}^T\mathbf{C}^0\right)}} \leqslant \frac{\hat{h}_n}{h^0}\cdot(1 + C'\epsilon_n) \leqslant 1 + C''\epsilon_n.$$

□

E.6 Proof of Proposition 3

Proof. (i) If $\alpha \geqslant \alpha'$, $\beta \leqslant \beta'$ and $\mathbf{W}$ is (α, β)-SS,

$$\begin{aligned}&[cone(\mathbf{W})^*]^{\alpha'} \bigcap [bd\mathcal{K}]^{\alpha'} \subseteq [cone(\mathbf{W})^*]^{\alpha} \bigcap [bd\mathcal{K}]^{\alpha} \\ &\subseteq \{\mathbf{x} : \|\mathbf{x} - \lambda \mathbf{e}_f\|_2 \leqslant \beta\lambda, \lambda \geqslant 0\} \subseteq \{\mathbf{x} : \|\mathbf{x} - \lambda \mathbf{e}_f\|_2 \leqslant \beta'\lambda, \lambda \geqslant 0\}.\end{aligned}$$

Then $\mathbf{W}$ is (α', β')-SS.

(ii) If $\mathrm{Conv}(\mathbf{W}) \subseteq \mathrm{Conv}(\bar{\mathbf{W}})$,

$$cone(\bar{\mathbf{W}})^* \subseteq cone(\mathbf{W})^* \subseteq \mathcal{K}. \tag{E.58}$$

Also, since

$$[cone(\mathbf{W})^*]^{\alpha} = \left\{\mathbf{x} : \mathbf{x}^T\mathbf{W} \geqslant -\alpha\|\mathbf{x}\|_2\right\} = \left\{\mathbf{x} : \mathbf{x}^T\mathbf{w} \geqslant -\alpha\|\mathbf{x}\|_2, \forall \mathbf{w} \in \mathrm{Conv}(\mathbf{W})\right\},$$

we have

$$[cone(\bar{\mathbf{W}})^*]^{\alpha} \subseteq [cone(\mathbf{W})^*]^{\alpha}. \tag{E.59}$$

By (E.58), (E.59) and the definition, if $\mathbf{W}$ is (α, β)-SS, $\bar{\mathbf{W}}$ is also (α, β)-SS.

(iii) The proof is trivial by definition.

□

E.7 Proof of Proposition 5

The following lemma is helpful in the proof of Proposition 5.

Lemma E.10. *For any $\mathbf{x} \in \mathbb{R}^k$, if*

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \leqslant \epsilon\|\mathbf{x}\|_1$$

for some $\epsilon \in [0, \frac{1}{2k}]$, then there exists one element x_i of $\mathbf{x}$ such that

$$\|\mathbf{x} - x_i\mathbf{e}_i\|_2 \leqslant 4\sqrt{k-1}\epsilon \cdot |x_i|.$$

Proof. It suffices to show the lemma holds for $\mathbf{x} \geqslant 0$ and $\|\mathbf{x}\|_1 = 1$. Now suppose there exist some elements of $\mathbf{x}$, say x_1, such that

$$x_1 \geqslant 2\epsilon \ \ and \ \ x_1 \leqslant \frac{1}{2}.$$

Since $\|\mathbf{x}_{-1}\|_2$ is a convex function, under the convex constraint

$$\left\{(x_2,\cdots,x_k):\sum_{j=2}^{k}x_j=1-x_1,x_j\geqslant 0,j=2,\cdots,k\right\},$$

it is maximized on the vertex of the constraint. Therefore, $\|\mathbf{x}_{-1}\|_2\leqslant 1-x_1$, which leads to

$$\begin{aligned}\|\mathbf{x}\|_1-\|\mathbf{x}\|_2&\geqslant 1-\sqrt{(1-x_1)^2+x_1^2}\\&=1-(1-x_1)\sqrt{1+\left(\frac{x_1}{1-x_1}\right)^2}\\&>1-(1-x_1)\left[1+\frac{1}{2}\left(\frac{x_1}{1-x_1}\right)^2\right]\\&=x_1-\frac{x_1^2}{2(1-x_1)}\\&\geqslant x_1-x_1^2\geqslant\frac{x_1}{2}\geqslant\epsilon,\end{aligned}$$

where the second inequality is because $\sqrt{1+t}<1+\frac{t}{2}$ for $t>0$. So we get a contradiction. Consequently, there is no element in $[2\epsilon,\frac{1}{2}]$. Since there is at least one element that is larger than or equal to $\frac{1}{k}$ and $2\epsilon\leqslant\frac{1}{k}$, at least one element is larger than or equal to 2ϵ. At the same time, there is at most one element that is larger than $\frac{1}{2}$, so there must be exactly one element that is larger than $\frac{1}{2}$. Let the element be x_i, then all other elements are less than 2ϵ. Therefore,

$$\|\mathbf{x}-x_i\mathbf{e}_i\|_2\leqslant\sqrt{k-1}\cdot 2\epsilon\leqslant\sqrt{k-1}\cdot 2\epsilon\cdot 2x_i$$

□

Now we are ready to present the proof of Proposition 5.

Proof. By Proposition 3(ii), it suffices to prove for the case when all $x_{ij}=m\in\left[0,\frac{1}{k}\right)$. Denote

$$\begin{aligned}A_{\beta_\epsilon}&=\{\mathbf{x}:\|\mathbf{x}-\lambda\mathbf{e}_f\|_2\leqslant\beta_\epsilon\lambda,\lambda\geqslant 0\},\\B_\epsilon&=[bd\mathcal{K}]^\epsilon=\left\{\mathbf{x}:|\|\mathbf{x}\|_2-\mathbf{x}^T\mathbf{1}_k|\leqslant\epsilon\|\mathbf{x}\|_2\right\},\\C_\epsilon&=[cone(\mathbf{W}^0)^*]^\epsilon=\left\{\mathbf{x}:\mathbf{x}^T\mathbf{W}^0\geqslant-\epsilon\|\mathbf{x}\|_2\right\}.\end{aligned}$$

Then it suffices to show that there exist $\beta_\epsilon\to 0$ when $\epsilon\to 0$ such that $B_\epsilon\bigcap C_\epsilon\subseteq A_{\beta_\epsilon}$ for $\epsilon>0$.

Without loss of generality, we assume $\|\mathbf{x}\|_2=1$. For any $\mathbf{x}\in B_\epsilon\bigcap C_\epsilon$, We consider the following three cases:

Case (i). When all elements of $\mathbf{x}$ are nonnegative. Since $\mathbf{x}\in B_\epsilon$, $\|\mathbf{x}\|_1=\sum_{i=1}^k x_i\leqslant 1+\epsilon$. Then

$$\|\mathbf{x}\|_1-\|\mathbf{x}\|_2\leqslant\epsilon\leqslant\epsilon\|\mathbf{x}\|_1. \tag{E.60}$$

Case (ii). When there exist at least negative two elements in $\mathbf{x}$. Suppose one of the negative elements is x_1. Denote

$$\mathcal{P} = \{i : x_i \geqslant 0\}, \ s = \sum_{i \in \mathcal{P}} x_i \geqslant 0.$$

And

$$\mathcal{N} = \{i : x_i < 0, i \neq 1\}, \ t = \sum_{i \in \mathcal{N}} x_i < 0.$$

Since $\mathbf{x} \in C_\epsilon$, for all $i \in \mathcal{N}$,

$$mx_1 + (1-m)x_i \geqslant -\epsilon,$$

which implies

$$0 > x_i \geqslant -\frac{\epsilon}{1-m}$$

and

$$0 > t = \sum_{i \in \mathcal{N}} x_i > -\frac{k}{1-m}\epsilon. \tag{E.61}$$

Also, pick any $i \in \mathcal{N}$,

$$-\epsilon \leqslant (1-m)x_1 + mx_i < (1-m)x_1.$$

Therefore,

$$0 > x_1 > -\frac{\epsilon}{1-m}. \tag{E.62}$$

By (E.61) and (E.62),

$$\sum_{i=1}^{k} x_i = x_1 + s + t > \left(|x_1| - \frac{2\epsilon}{1-m}\right) + s + \left(|t| - \frac{2k}{1-m}\epsilon\right) = \|\mathbf{x}\|_1 - \frac{2k+2}{1-m}\epsilon. \tag{E.63}$$

Since $\mathbf{x} \in B_\epsilon$,

$$\sum_{i=1}^{k} x_i \leqslant 1 + \epsilon. \tag{E.64}$$

By (E.63) and (E.64),

$$\|\mathbf{x}\|_1 \leqslant 1 + \left(1 + \frac{2k+2}{1-m}\right)\epsilon,$$

i.e.,

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \leqslant \left(1 + \frac{2k+2}{1-m}\right)\epsilon \leqslant \left(1 + \frac{2k+2}{1-m}\right)\epsilon \cdot \|\mathbf{x}\|_1. \tag{E.65}$$

Case (iii). When there exists only one negative element in $\mathbf{x}$. Suppose the negative element is x_1. Without loss of generality, we assume $x_k \geqslant |x_j|$ for all $j = 1, \cdots, k-1$.

Denote

$$r = \sum_{i=2}^{k-1} x_i \geqslant 0.$$

Since $\mathbf{x} \in B_\epsilon$, $\sum_{i=1}^{k} x_i \leqslant 1+\epsilon$. At the same time, $\|\mathbf{x}\|_2 = 1$. Combining these two expressions, we have

$$0 \leqslant \sum_{i=1}^{k-1} x_i^2 + \left(\sum_{i=1}^{k-1} x_i\right)^2 - 2(1+\epsilon)\sum_{i=1}^{k-1} x_i + \epsilon^2 + 2\epsilon \tag{E.66}$$

In (E.66), applying the fact that

$$\sum_{i=2}^{k-1} x_i^2 \leqslant \left(\sum_{i=2}^{k-1} x_i\right)^2 = r^2,$$

we have

$$0 \leqslant 2x_1^2 - 2(1+\epsilon)x_1 + 2r^2 - 2(1+\epsilon)r + \epsilon^2 + 2\epsilon. \tag{E.67}$$

Since $\mathbf{x} \in C_\epsilon$, for all $i = 2, \cdots, k-1$,

$$(1-m)x_1 + mx_i \geqslant -\epsilon,$$

which implies

$$0 > x_1 \geqslant -\frac{m}{(k-2)(1-m)}r - \frac{\epsilon}{1-m}. \tag{E.68}$$

From (E.67) and (E.68), we derive that

$$\begin{aligned} 0 \leqslant & 2\left[\left(\frac{m}{(k-2)(1-m)}\right)^2 + 1\right]r^2 + \left[\frac{4m}{(k-2)(1-m)^2}\epsilon + \frac{2m}{(k-2)(1-m)}(1+\epsilon) - 2 - 2\epsilon\right]r \\ & + \left[\frac{2}{(1-m)^2}\epsilon^2 + \frac{2}{1-m}\epsilon + \frac{2}{1-m}\epsilon^2 + \epsilon^2 + 2\epsilon\right] \end{aligned} \tag{E.69}$$

Since $\mathbf{x} \in B_\epsilon$,

$$1 - \epsilon \leqslant \sum_{i=1}^{k} x_i = x_1 + r + x_k \leqslant r + x_k \leqslant (k-1)x_k,$$

so

$$x_k \geqslant \frac{1-\epsilon}{k-1}. \tag{E.70}$$

Since $\mathbf{x} \in C_\epsilon$,

$$(1-m)x_1 + mx_k \geqslant -\epsilon.$$

So

$$x_1 \geqslant -\frac{m}{1-m}x_k - \frac{\epsilon}{1-m}. \tag{E.71}$$

Therefore,

$$\begin{aligned}1+\epsilon \geqslant \sum_{i=1}^{k} x_i &= x_1 + r + x_k \\ &\overset{(E.71)}{\geqslant} -\frac{m}{1-m}x_k - \frac{\epsilon}{1-m} + r + x_k \\ &\overset{(E.70)}{\geqslant} \left(1-\frac{m}{1-m}\right)\frac{1-\epsilon}{k-1} - \frac{\epsilon}{1-m} + r.\end{aligned}$$

In other words,

$$r \leqslant \left(1-\frac{1-2m}{(k-1)(1-m)}\right) + \left(1+\frac{1}{1-m}+\frac{1-2m}{(k-1)(1-m)}\right)\epsilon \tag{E.72}$$

By (E.69) and (E.72), we get $r < 30\epsilon$ when ϵ is small enough. Then combining with (E.68),

$$0 > x_1 \geqslant -\frac{1}{1-m}\left(\frac{30m}{k-2}+1\right)\epsilon.$$

Consequently,

$$\begin{aligned}\sum_{i=1}^{k} x_i = x_1 + r + x_k &> \left[|x_1| - \frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\epsilon\right] + r + x_k \\ &= \|\mathbf{x}\|_1 - \frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\epsilon.\end{aligned} \tag{E.73}$$

Since $\mathbf{x} \in B_\epsilon$,

$$\sum_{i=1}^{k} x_i \leqslant 1+\epsilon. \tag{E.74}$$

By (E.73) and (E.74),

$$\|\mathbf{x}\|_1 \leqslant 1 + \left[1+\frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\right]\epsilon,$$

i.e.,

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \leqslant \left[1+\frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\right]\epsilon \leqslant \left[1+\frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\right]\epsilon\cdot\|\mathbf{x}\|_1. \tag{E.75}$$

Finally, combining the three cases above, by (E.60), (E.65) and (E.75), for any $\mathbf{x} \in B_\epsilon \bigcap C_\epsilon$,

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \leqslant \left[1+\max\left\{\frac{2k+2}{1-m}, \frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\right\}\right]\epsilon\cdot\|\mathbf{x}\|_1. \tag{E.76}$$

Then by Lemma E.10, we know that $B_\epsilon \bigcap C_\epsilon \subseteq A_{\beta_\epsilon}$ for all $\epsilon > 0$ and ϵ small, where

$$\beta_\epsilon = 4\sqrt{k-1}\cdot\left[1+\max\left\{\frac{2k+2}{1-m}, \frac{2}{1-m}\left(\frac{30m}{k-2}+1\right)\right\}\right]\epsilon \to 0$$

when $\epsilon \to 0$. □

E.8 Proof of Proposition 6

Proof. By Step 1 of the proof of Theorem 8 in Section D.3 we know that: if the probability density function satisfies

$$\mathbb{P}(\|\mathbf{w}-\mathbf{w}_i^\sharp\|_2 \leqslant r) \geqslant (k-1)!\cdot c_0\cdot r^{k-1}, \quad \forall\, 0<r\leqslant r_0 \tag{E.77}$$

for the s distinct points $\mathbf{w}_1^\sharp,\cdots,\mathbf{w}_s^\sharp$ in its support and some positive constants r_0, c_0, then with probability at least $1-C_1 s/d$, for any $\mathbf{w}_i^\sharp$, there exists at least one sample $\mathbf{w}_{(i)}^1$, such that

$$\|\mathbf{w}_{(i)}^1-\mathbf{w}_i^\sharp\|_2 \leqslant r_d, \quad \forall i=1,\cdots,s, \tag{E.78}$$

where $r_d=\left(\frac{\log d}{d}\right)^{\frac{1}{k-1}}$.

Now we show that

$$[cone(\mathbf{W}_1^0)^*]^{\alpha-r_d}=\{\mathbf{x}:\mathbf{x}^T\mathbf{W}_1^0\geqslant -(\alpha-r_d)\|\mathbf{x}\|_2\}\subseteq[cone(\mathbf{W}^\sharp)^*]^\alpha=\{\mathbf{x}:\mathbf{x}^T\mathbf{W}^\sharp\geqslant-\alpha\|\mathbf{x}\|_2\}.$$

For any $\mathbf{x}\in\mathbb{R}^k$ and $\mathbf{x}^T\mathbf{W}_1^0\geqslant-(\alpha-r_d)\|\mathbf{x}\|_2$,

$$-\alpha\|\mathbf{x}\|_2\leqslant\mathbf{x}^T\mathbf{W}_1^0-r_d\|\mathbf{x}\|_2\leqslant\mathbf{x}^T\mathbf{W}_1^0+\mathbf{x}^T(\mathbf{W}^\sharp-\mathbf{W}_1^0)=\mathbf{x}^T\mathbf{W}^\sharp,$$

where in the second inequality we apply (E.78) and Cauchy–Schwarz inequality. Therefore, by definition, if $\mathbf{W}^\sharp$ is (α,β)-SS, $\mathbf{W}_1^0$ is $(\alpha-r_d,\beta)$-SS.

We pick $\mathbf{w}_i^\sharp$ to be the vertex $\mathbf{e}_i$ of Δ^{k-1} for $i=1,\cdots,k$. Apparently, when the density function is uniformly larger than a constant on neighborhoods of $\mathbf{e}_i$'s, (E.77) holds. Let

$$\alpha_0=C_2\sqrt{\frac{\log(n\vee d)}{n}}+r_d,$$

then by Proposition 5, $\mathbf{W}^\sharp$ is $(\alpha_0,C_3\alpha_0)$-SS for all n and d if $\frac{\log d}{n}\to 0$. As a result, $\mathbf{W}_1^0$ is $(C_2\sqrt{\frac{\log(n\vee d)}{n}},C_3\alpha_0)$-SS.

When $d\geqslant Cn^{\frac{k-1}{2}}$, we have

$$\alpha_0=C_2\sqrt{\frac{\log(n\vee d)}{n}}+r_d\leqslant C_2'\sqrt{\frac{\log(n\vee d)}{n}}.$$

Then $\mathbf{W}_1^0$ is $(C_2\sqrt{\frac{\log(n\vee d)}{n}},C_3'\sqrt{\frac{\log(n\vee d)}{n}})$-SS for all n and d if $\frac{\log d}{n}\to 0$, which finishes the proof. □

F Additional Simulations and Experiments

F.1 Convergence of the Estimation

We use the Monte Carlo simulation to show the convergence of the integrated likelihood $F_{n\times d}(\mathbf{C})$ and the MLE $\hat{\mathbf{C}}_n$. Consider a simple setup: $k = V = 3, d = 6$,

$$\mathbf{C}^0 = \begin{bmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 2/3 & 1/6 \\ 1/6 & 1/6 & 2/3 \end{bmatrix}, \quad \mathbf{W}^0 = \begin{bmatrix} 5/6 & 0 & 1/6 & 5/6 & 1/6 & 0 \\ 1/6 & 5/6 & 0 & 0 & 5/6 & 1/6 \\ 0 & 1/6 & 5/6 & 1/6 & 0 & 5/6 \end{bmatrix},$$

and the sample size is set to be $n = 60, 600, 6000, 60000$.

In the experiment, we consider the "noiseless" data, i.e., $\mathbf{X} = n\mathbf{C}^0\mathbf{W}^0$. We compare the integrated likelihood among candidate $\mathbf{C}$'s taking the following form:

$$\begin{bmatrix} c & (1-c)/2 & (1-c)/2 \\ (1-c)/2 & c & (1-c)/2 \\ (1-c)/2 & (1-c)/2 & c \end{bmatrix}, \tag{F.79}$$

with c taking values from $[0.5, 1]$. We use Monte Carlo method to evaluate the integrated likelihood (4):

$$\hat{F}_{n\times d,T}(\mathbf{C}) \approx \prod_{i=1}^{d} \left[\frac{1}{T} \sum_{t=1}^{T} f_n(\mathbf{x}^{(i)} | \mathbf{u} = \mathbf{C}\mathbf{w}_t) \right],$$

where $\mathbf{w}_1, \cdots, \mathbf{w}_T$ are i.i.d. random samples from $\mathrm{Dir}_k(\mathbf{1})$ and $T = 50,000$.

The left plot of Figure S3 shows $\hat{F}_{n\times d,T}(\mathbf{C})/\max_{\mathbf{C}} \hat{F}_{n\times d,T}(\mathbf{C})$, the relative value of the estimated integrated likelihood. As n increases, the peak of the likelihood approach the truth (i.e., $c = 2/3$): the optimal c values that maximize $\hat{F}_{n\times d,T}(\mathbf{C})$ for $n = 60, 600, 6000, 60000$, are $0.778, 0.720, 0.701, 0.686$, respectively. The small fluctuations in the curves of $n = 6000, 60000$ are possibly due to numeric issues. The right plot of Figure S3 displays $\mathrm{Conv}(\mathbf{C}^0)$ and the optimal $\mathrm{Conv}(\mathbf{C})$'s for different n.

F.2 Comparison with Other Methods

In this section, we provide additional simulation studies to compare the proposed method (MCMC-EM) with several existing approaches: Anchor Free (AnchorF) (Huang et al., 2016), Geometric Dirichlet Means (GDM) (Yurochkin and Nguyen, 2016), and two MCMC algorithms based on Gibbs sampler (Gibbs) (Griffiths and Steyvers, 2004) and based on partially collapsed Gibbs sampler (pcLDA) (Magnusson et al., 2018; Terenin et al., 2018).

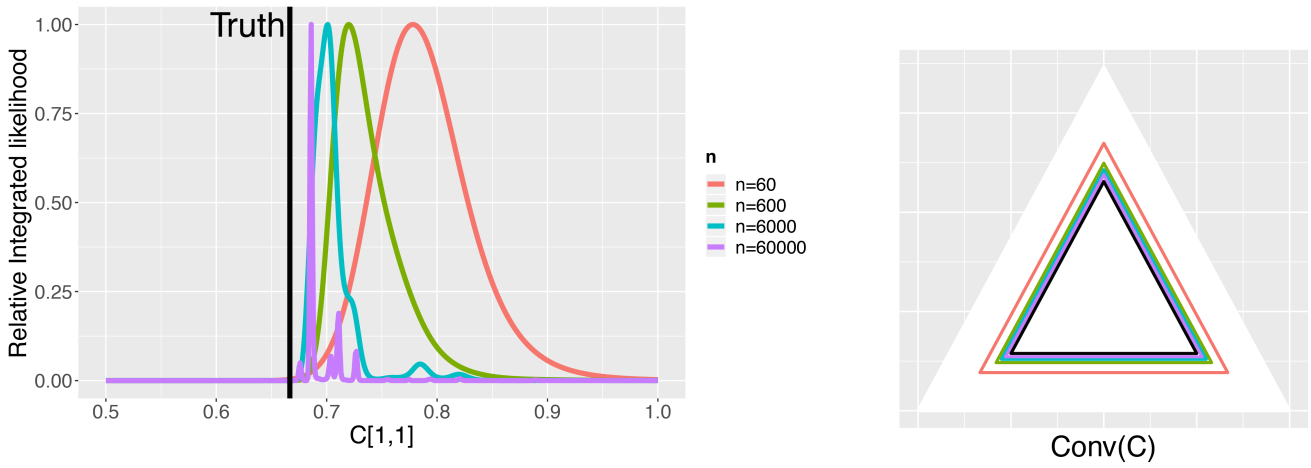


Figure S3: Results of the experiment in Section F.1. Left: the relative integrated likelihood of "noiseless" data. Right: Conv($\mathbf{C}^0$) and the optimal Conv($\mathbf{C}$)'s under different n. The white triangle represents Δ^2; the smallest black triangle is Conv($\mathbf{C}^0$); other colored triangles represent the Conv($\mathbf{C}$)'s that maximize $\hat{F}_{n\times d,T}(\mathbf{C})$ under different n's. The legend in the middle is shared by both plots.

The basic simulation setup is as follows: $V = 1200$, $d = 1000$, $n = 1000$ and $k = 5$, columns of $\mathbf{C}$ are generated from $Dir_V(0.1)$ and columns of $\mathbf{W}$ are from $Dir_k(0.1)$. For our MCMC-EM algorithm, the number of MCMC samples is 20 without burn-in. The EM algorithm stops after 50 iterations; For each simulation, we run the EM algorithm 12 times in parallel with different randomly-initialized parameters and report the result with the highest likelihood value. All hyper-parameters are set as default, except that the prior over mixing weights in Gibbs and pcLDA is set to be uniform, same as ours.

We evaluate the performance by the following four metrics:

- *Relative Error* is defined by $\min_{\mathbf{\Pi}} \|\hat{\mathbf{C}}\mathbf{\Pi} - \mathbf{C}\|_F / \|\mathbf{C}\|_F$, where $\mathbf{\Pi}$ is a permutation matrix.
- *Topic Coherence* is used to measure the single-topic quality, defined as

$$\sum_{l=1}^{k} \sum_{v_1, v_2 \in \mathcal{V}_l} \log\left(\frac{\text{freq}(v_1, v_2) + \epsilon}{\text{freq}(v_2)}\right)$$

  where $\mathcal{V}_l$ is the leading 20 words for topic l, freq(v_1, v_2), freq(v_2) are the co-occurrence count of word v_1 and word v_2 and the occurrence counts of word v_2, respectively, and ϵ is a small constant added to avoid numerical issue. Generally, the higher the topic coherence is, the better the quality of the mined topics is.

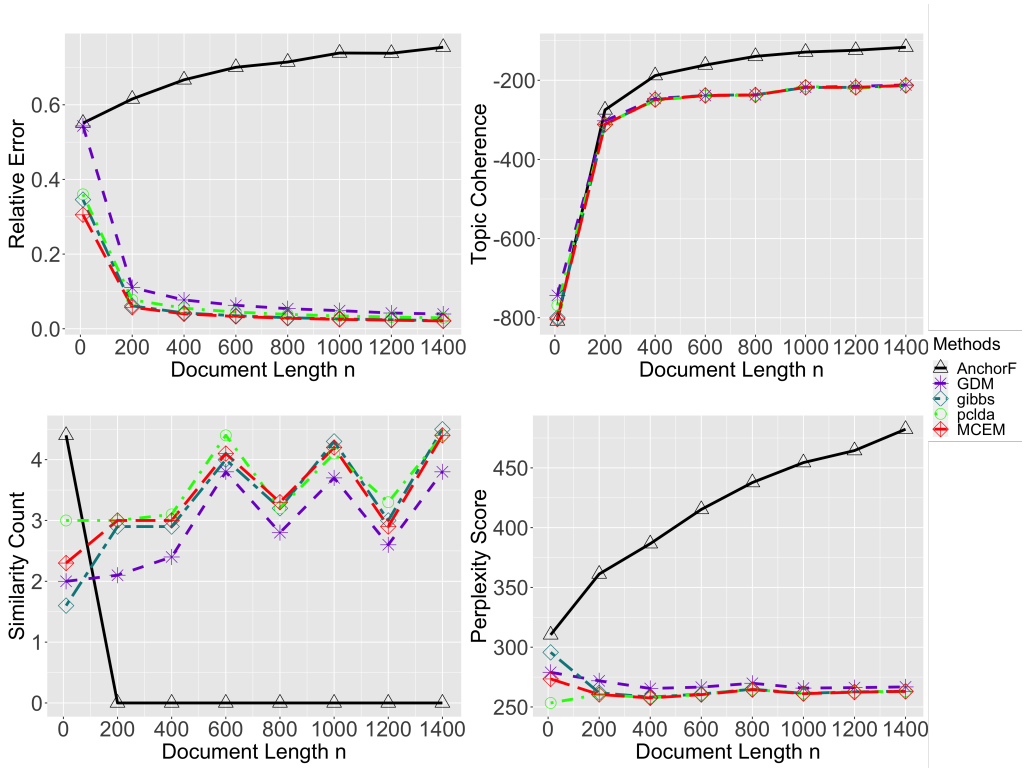


Figure S4: Comparison with existing methods when document length varies ($n = 10, 200, 400, \cdots, 1400$).

- *Similarity Count* is used to measure similarity between topics (Arora et al., 2013; Huang et al., 2016), which is obtained simply by adding up the overlapped words across $\mathcal{V}_l$.

$$\sum_{l_1<l_2} \sum_{v_1 \in V_{l_1}, v_2 \in V_{l_2}} \mathbb{1}(v_1 = v_2).$$

  It focuses on the relationship between mined topics while the topic coherence measures the one within each topic. A smaller similarity count means the mined topics are more distinguishable.

- *Perplexity Score* measures the goodness of fit of the fitted model to the data. It is the multiplicative inverse of the likelihood normalized by the number of words. Sometimes the perplexity score is calculated on the hold-out data. Here, for simplicity, we use the one based on the training data (the whole dataset),

$$\sqrt[\sum_{i=1}^{d} n_i]{\frac{1}{\prod_{i=1}^{d} f_{n_i}(\mathbf{x}^{(i)}|\hat{\mathbf{C}}, \hat{\mathbf{w}}^{(i)})}}.$$

  For a fixed k, a smaller perplexity score implies a better fit of the model.

We investigate the performance of those methods (i) when document length n varies, (ii) when both document length n and number of documents d varies, and (iii) when the parameter α of the Dirichlet

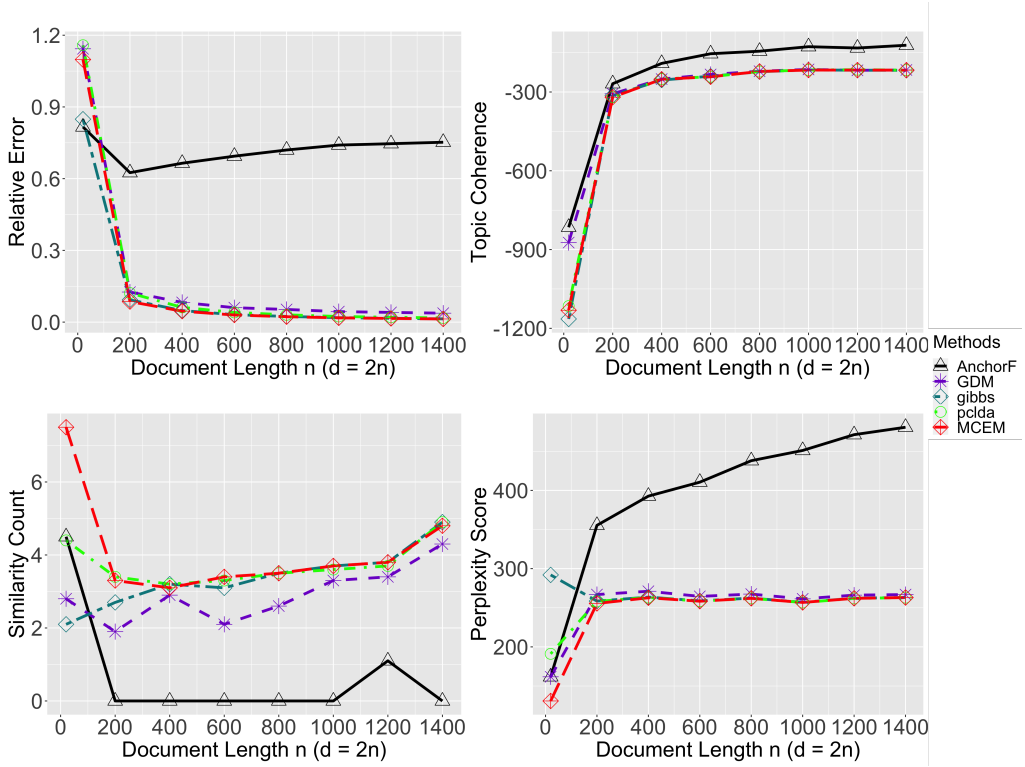


Figure S5: Comparison with existing methods when both document length n and number of documents d vary ($n = 20, 200, 400, \cdots, 1400$, $d = 2n$).

distribution we use to generate $\mathbf{W}$ varies. Results are reported in Figure S4 to Figure S6; each metric reported in those plots is the average over 10 repetitions. Below we summarize our findings:

(i) MCMC-EM, GDM, Gibbs and pcLDA perform very similarly in these three simulation settings in terms of four different evaluation metrics. That is because MCMC-EM, Gibbs and pcLDA have the same objective function and GDM is also a likelihood-based approach. MCMC-EM has the best relative error and perplexity score in most experiments of the first two settings;

(ii) Estimators of MCMC-EM, GDM, Gibbs and pcLDA converge very quickly as n increases or as both n and d increase. Their performance is stable as the Dirichlet parameter α increases;

(iii) The eigenvalue decomposition-based approach AnchorF has better similarity count than other methods in most experiments. However, it performs much worse than others in terms of relative error and perplexity score in almost all experiments. The topic coherence of AnchorF is slightly better than the others in the first two settings, but decreases sharply as the Dirichlet parameter α increases in the third setting.

In Table 1, we report the computation time of our MCMC-EM algorithm and other methods for

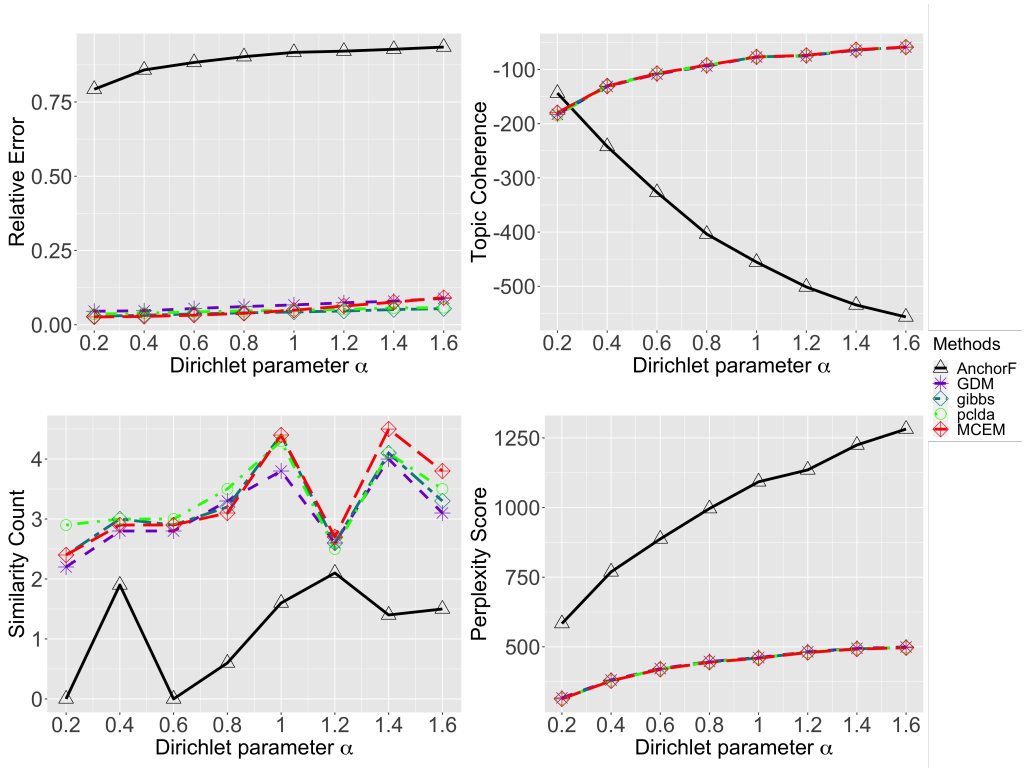


Figure S6: Comparison with existing methods when the Dirichlet parameter α varies. Columns of $\mathbf{W} \sim Dir_k(\alpha)$ with $\alpha = 0.2, 0.4, \cdots, 1.6$. Identity matrix $\mathbf{I}_k$ is appended to the randomly sampled matrix $\mathbf{W}$ to ensure model identifiability.

the experiment in Fig. S6 ($V = 1200$, $d = n = 1000$ and $k = 5$). For our MCMC-EM algorithm, the number of MCMC samples is 20 without burn-in. The EM algorithm stops after 50 iterations. The results show that the computation time of our MCMC-EM algorithm is comparable with the other methods. Our code, which is currently partially implemented in C++, could run faster if being fully implemented in C++; in comparison, the publicly available codes of the competing methods have been mostly highly optimized.

Method	AnchorF	GDM	Gibbs	pcLDA	MC^2-EM
Time/s	6.93	0.27	82.20	34.83	49.15

Table 1: Computational time of the MCMC-EM algorithm and other methods ($V = 1200$, $d = n = 1000$, $k = 5$).

F.3 Selecting the Number of Topics

In practice, the number of topics k is unknown. Below we propose a procedure to select k based on the "effective rank" of the sample term-document matrix $\hat{\mathbf{U}}$ reflected in the spectrum.

In Theorem 2, the topic matrix $\mathbf{C}$ is assumed to have full rank; consequently, the true term-document matrix $\mathbf{U} = \mathbf{CW}$ has rank k. By Weyl's inequality (Weyl, 1912), the singular values of the sample term-document matrix $\hat{\mathbf{U}}$ are expected to be close to those of $\mathbf{U}$. Similar to the elbow method used in selecting the number of components in clustering analysis and in PCA, we plot the ordered singular values of $\hat{\mathbf{U}}$ versus its index, and then select k by detecting the location of a significant drop of the curve.

To test our procedure, we conducted a simulation study where $k = 5$, $V = 1200$, $d = 1000$ and $n = 50$. Columns of $\mathbf{C}$ are randomly generated from $Dir_V(0.1)$ and columns of $\mathbf{W}$ are randomly generated from $Dir_k(0.1)$. We repeated the experiment 10 times and the results are shown in Fig. S7. From the figure we can see that there is a sudden drop between the 5th and the 6th largest singular values. And the singular values after the 6th one are stable. So, we would set $k = 5$, which agrees with the underlying truth.

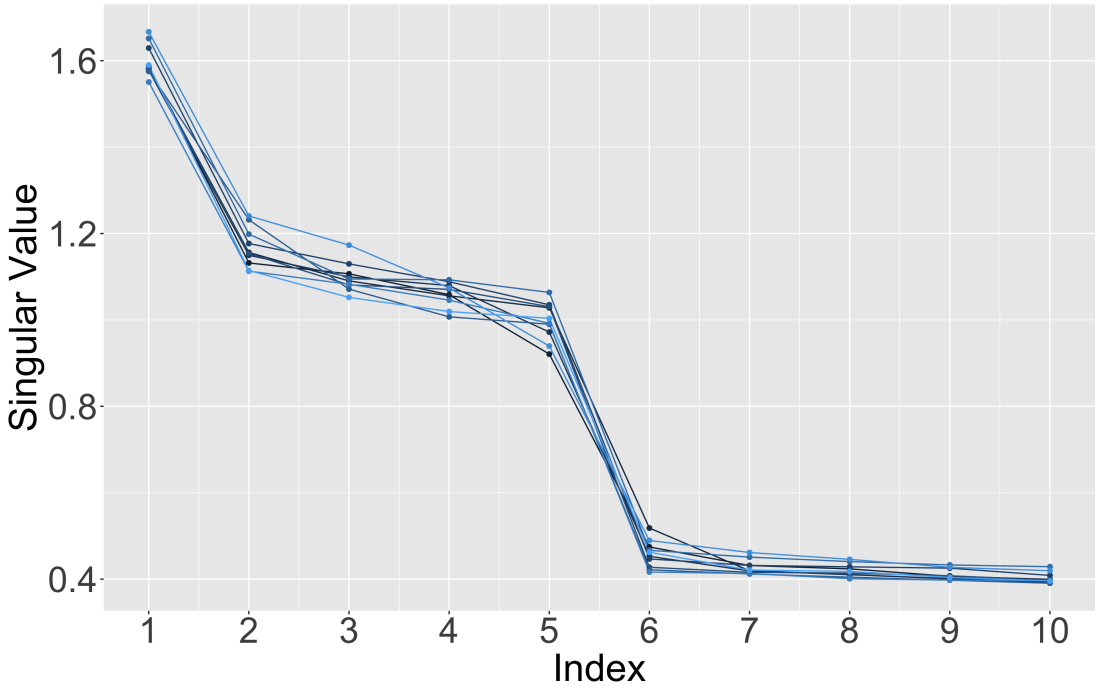


Figure S7: Singular values plot of sample term-document matrices. In 10 repetitions of the experiments, $k = 5$, $V = 1200$, $d = 1000$ and $n = 50$. Columns of $\mathbf{C}$ and $\mathbf{W}$ are generated Dirichlet distributions.

We also apply the approach to the two text data used in the paper – the NIPS and the Daily Kos

datasets. The singular values plots are in Figure S8. For the NIPS dataset, there is a drop between 5th and 6th largest singular values. For the Daily Kos dataset, there is a drop between 7th and 8th largest singular values. So we choose 5 and 7 as the recommended number of topics for the NIPS and the Daily Kos datasets, respectively.

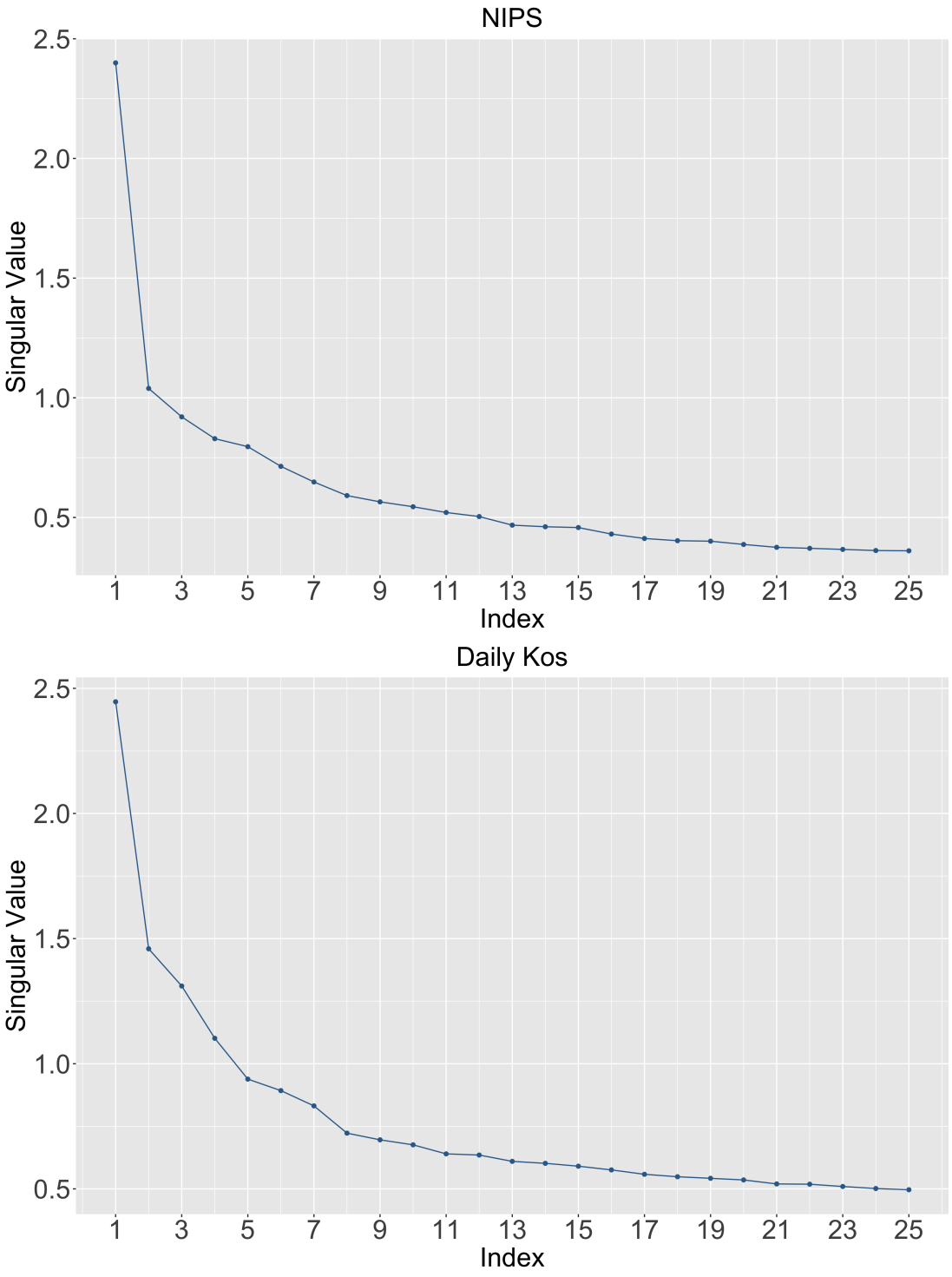


Figure S8: Singular values plot of the NIPS and the Daily Kos datasets.

G Estimated Topics for the NIPS Dataset

The NIPS dataset is originally from Perrone et al. (2016) and is accessible on UCI Machine Learning Repository[6]. It contains $V = 11463$ words and $d = 5811$ NIPS conference papers published between 1987 and 2015, with an average document length of 1902. In this section, we display the top 10 words of mined topics output by our MCMC-EM algorithm at $k = 5, 10, 15, 20$.

[6] https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
network	algorithm	model	training	learning
neural	matrix	models	learning	algorithm
input	function	data	data	state
time	problem	distribution	set	time
model	data	inference	image	function
networks	set	using	features	value
figure	error	parameters	feature	policy
neurons	linear	prior	classification	set
output	let	likelihood	using	action
system	theorem	bayesian	images	optimal

Table 2: Mined topics for NIPS at $k = 5$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
image	learning	data	network	graph	neurons	state	algorithm	matrix	model
images	training	noise	networks	algorithm	model	learning	function	kernel	models
object	data	time	neural	tree	time	policy	theorem	data	distribution
model	classification	model	learning	set	neural	time	bound	problem	data
figure	set	figure	training	node	neuron	action	let	linear	gaussian
features	features	test	input	clustering	input	value	learning	algorithm	inference
using	class	error	output	nodes	spike	function	loss	method	likelihood
visual	feature	performance	layer	number	figure	algorithm	case	sparse	parameters
recognition	label	estimate	units	problem	activity	reward	functions	methods	bayesian
objects	using	using	hidden	time	stimulus	optimal	error	vector	prior

Table 3: Mined topics for NIPS at $k = 10$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
neurons	learning	model	learning	model	graph	network	state
model	data	models	algorithm	time	nodes	neural	learning
time	training	distribution	bound	data	node	networks	policy
neuron	classification	inference	loss	models	graphs	input	action
spike	set	bayesian	bounds	figure	structure	learning	value
neural	features	data	probability	human	network	output	states
activity	feature	parameters	error	prediction	edge	training	time
stimulus	class	likelihood	theorem	task	networks	units	reward
cells	label	latent	let	subjects	edges	layer	function
figure	using	posterior	regret	target	random	hidden	control

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
gaussian	image	tree	algorithm	kernel	matrix	function
data	images	time	algorithms	data	sparse	problem
distribution	object	speech	gradient	space	norm	set
function	features	using	time	points	rank	functions
mean	model	trees	optimization	distance	problem	theorem
noise	objects	source	methods	clustering	matrices	let
variance	using	algorithm	method	linear	data	convex
error	feature	signal	step	dimensional	analysis	case
estimate	recognition	node	cost	point	algorithm	following
estimation	figure	used	convergence	kernels	convex	given

Table 4: Mined topics for NIPS at $k = 15$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
theorem	network	neurons	distribution	function	graph	model	image	policy	matrix
bound	networks	model	posterior	points	tree	data	images	learning	data
let	learning	spike	bayesian	point	nodes	models	object	action	rank
probability	training	neuron	sampling	functions	node	parameters	features	value	matrices
bounds	neural	time	prior	space	algorithm	distribution	model	state	norm
proof	input	activity	inference	error	set	likelihood	recognition	reward	low
sample	layer	neural	process	case	graphs	gaussian	feature	function	algorithm
lemma	units	cells	variational	approximation	variables	variables	using	optimal	dimensional
following	output	input	data	mean	structure	log	objects	actions	pca
distribution	hidden	stimulus	model	given	edge	mixture	vision	decision	analysis
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
algorithm	kernel	noise	time	algorithm	learning	features	clustering	system	model
algorithms	distance	signal	state	optimization	training	feature	word	memory	human
online	data	using	model	gradient	classification	data	cluster	network	figure
regret	kernels	speech	states	problem	class	set	words	neural	task
learning	space	filter	sequence	function	data	regression	clusters	figure	target
set	metric	source	system	convex	examples	method	data	input	subjects
time	learning	signals	markov	methods	label	selection	language	control	brain
number	based	sparse	transition	convergence	set	number	set	time	information
problem	using	coding	dynamics	method	classifier	problem	model	output	subject
bound	similarity	basis	sequences	solution	error	methods	means	systems	experiment

Table 5: Mined topics for NIPS at $k = 20$

H Estimated Topics for the KOS Dataset

The Daily Kos dataset is accessible on UCI Machine Learning Repository Bag of Words Database[7], and its original source is dailykos.com, a group blog and internet forum focused on the Democratic Party and liberal American politics. The KOS dataset contains $V = 6906$ words and $d = 3430$ Daily Kos blog entries, with an average document length of 67. In this section, we display the top 10 words of mined topics output by our MCMC-EM algorithm at $k = 5, 10, 15, 20$.

[7] https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
republican	kerry	bush	november	iraq
senate	bush	president	poll	war
house	dean	people	house	bush
party	poll	kerry	republicans	administration
democrats	percent	media	governor	military
campaign	edwards	bushs	senate	american
democratic	democratic	time	electoral	president
elections	voters	campaign	polls	iraqi
race	primary	general	account	people
state	polls	years	vote	officials

Table 6: Mined topics for KOS at $k = 5$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
kerry	november	senate	general	house	bush	iraq	campaign	bush	people
dean	poll	race	election	bush	kerry	war	media	tax	political
poll	house	house	bush	committee	president	military	party	administration	america
edwards	republicans	elections	states	white	bushs	iraqi	democratic	years	issue
primary	governor	republican	republican	national	administration	american	million	jobs	rights
percent	senate	democrats	voters	delay	general	troops	money	year	time
clark	electoral	state	state	texas	cheney	soldiers	time	health	marriage
democratic	account	district	party	administration	war	saddam	people	percent	conservative
polls	polls	gop	vote	court	iraq	officials	political	million	politics
results	vote	democratic	nader	report	john	people	democrats	economy	gay

Table 7: Mined topics for KOS at $k = 10$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
bush	party	senate	bush	years	law	poll	people
news	campaign	race	kerry	people	republicans	bush	campaign
national	money	elections	president	abu	court	percent	convention
john	million	republican	general	policy	rights	kerry	media
kerry	democratic	state	bushs	blades	republican	voters	nader
war	house	house	john	meteor	marriage	polls	speech
bushs	democrats	democrats	campaign	ghraib	issue	results	tom
general	delay	district	kerrys	american	state	numbers	ballot
service	candidates	seat	debate	government	gay	polling	time
campaign	committee	gop	election	international	political	lead	party

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
election	dean	bush	media	iraq	bush	november
vote	edwards	tax	people	war	administration	poll
general	kerry	jobs	time	iraqi	president	house
voting	primary	administration	ive	military	iraq	account
voters	democratic	year	community	troops	house	governor
republicans	clark	health	blog	american	intelligence	senate
voter	iowa	economy	political	soldiers	white	polls
republican	gephardt	years	dkos	forces	cheney	electoral
ohio	lieberman	billion	read	killed	report	republicans
oct	jan	states	ill	baghdad	war	vote

Table 8: Mined topics for KOS at $k = 15$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
news	military	iraq	tax	party	people	iraq	law	administration	house
media	abu	war	billion	democratic	life	war	court	bush	republicans
john	women	saddam	years	campaign	american	iraqi	marriage	white	delay
campaign	ghraib	troops	year	political	political	baghdad	gay	house	republican
fox	rumsfeld	united	federal	democrats	country	killed	rights	intelligence	democrats
debate	people	iraqi	cuts	candidates	years	american	amendment	president	committee
press	american	american	budget	candidate	family	soldiers	federal	report	senate
sunday	defense	bush	energy	election	white	military	issue	commission	gop
national	health	country	plan	campaigns	america	forces	state	officials	elections
mccain	war	military	health	dnc	politics	city	legal	security	bill
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
poll	dean	november	bush	media	bush	million	bush	bush	senate
percent	edwards	poll	cheney	people	kerry	money	states	democrats	race
kerry	kerry	house	general	time	president	campaign	state	republicans	elections
bush	primary	governor	kerry	ive	bushs	candidates	kerry	jobs	republican
polls	clark	account	service	ill	john	raised	nader	republican	state
voters	democratic	electoral	national	blog	general	house	general	democratic	seat
results	iowa	republicans	bushs	bloggers	campaign	dkos	florida	president	district
polling	gephardt	senate	guard	night	kerrys	fundraising	election	conservative	democrats
numbers	lieberman	polls	military	convention	george	donors	vote	job	gop
lead	poll	vote	president	email	debate	time	ohio	reagan	candidate

Table 9: Mined topics for KOS at $k = 20$

I Mined meta states for the taxi-trip dataset

Figure S9: Estimation of disaggregation distributions for NYC taxi-trip data for $k = 9$: $\hat{\mathbf{C}}_1, \hat{\mathbf{C}}_2, \cdots, \hat{\mathbf{C}}_9 \in \mathbb{R}^V$, where $\hat{\mathbf{C}}_l = \mathbb{P}(X_{t+1}|Z_t = l)$.

Figure S10: Estimation of aggregation distributions for NYC taxi-trip data for $k = 9$: $\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \cdots, \hat{\mathbf{W}}_9 \in \mathbb{R}^V$, where $\hat{\mathbf{W}}_l = \mathbb{P}(Z_t = l|X_t)$.