Learning Topic Models: Identifiability and Finite-Sample Analysis

Yinyin Chen*a, Shishuang He*b, Yun Yangb, and Feng Liangb

*Meta Platforms, Inc., Menlo Park, CA; bDepartment of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL

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 article, 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. Supplementary materials for this article are available online.

1. Introduction

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

$$U_{V \times d} = C_{V \times k} W_{k \times d},$$

where all matrices are column-stochastic. In particular, $U_{V \times d}$ is the true term-document matrix whose columns are the true underlying word frequencies for the $d$ documents; $C_{V \times k}$ is the topic matrix whose columns are the multinomial parameters (i.e., word frequencies) for the $k$ topics; and $W_{k \times d}$ 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, that is, to estimate the collection’s topic matrix $C$. Despite the popularity and success of topic models, work on the estimation accuracy of $C$ 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 $C$ based on $U$ is equivalent to finding a $k$-vertex convex polytope that encloses all columns of $U$; the vertices of this $k$-vertex convex polytope form the columns of $C$. Apparently, such a convex polytope may not be unique; see Figure 1(a). In statistical language, topic models parameterized by $(C, W)$ without any further constraints are not identifiable (modulo column permutations).

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

1. Identifiability. Under what conditions is a topic model parameterized by $(C, W)$ 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 $C$ 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.
1.1. Related Work

Topic models have been studied under two settings: one in which the mixing weights, columns of $W$, 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 full support on the simplex $\Delta^{k-1}$. Therefore, identifiability can be guaranteed under very mild conditions; for example, one such condition is just that $C$ 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 article, 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 (Papadimitriou et al. 2000; McSherry 2001; Azar et al. 2001; Donoho and Stodden 2004; Kleinberg and Sandler 2003, 2008; Anandkumar, Hsu, and Kakade 2012; Arora, Ge, and Moitra 2012; Recht et al. 2012; Ge and Zou 2015; Ke and Wang 2017). The separability condition can be imposed either on rows of $C$ or on columns of $W$, due to the symmetry between these two matrices in the factorization (1).

When imposed on the topic matrix $C_{V \times k}$, this condition assumes that, after the rows of $C$ have been rearranged, 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 $W_{k \times d}$, this condition again assumes that, after the columns of $W$ 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 $W$ 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 $U$ to form the $k$-vertex polytope that contains all other columns of $U$. In other words, the topic matrix $C_{V \times k}$ 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 $C$ with a convergence rate of the order $1/\sqrt{nd}$ (Arora, Ge, and Moitra 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 $C$; 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 nonregular statistical problem of boundary estimation (see, 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 $U$ has led researchers to start looking at geometric conditions.

When there are multiple $k$-vertex convex polytopes enclosing columns of $U$, 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 con-
text 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 Figures 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, Fu, and Sidiropoulos 2016; Jang and Hero 2019). Huang, Fu, and Sidiropoulos (2016) ensure identifiability under the SS condition by adding the constraint that the determinant of $WW^\top$ is minimized. Jang and Hero (2019) have proved that the SS condition, along with volume minimization on the convex hull of $C$, 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 $C$ is equal to the determinant of $C$ (or to a monotonic function of the determinant of $C^T C$) which holds true only when $V = k$. In addition, neither Huang, Fu, and Sidiropoulos (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 $C$ 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 $C$ is closest to the convex hull formed by the columns of $U$. They show that the minimizer is indeed unique when the separability condition is imposed on $W$; 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 $C$) 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 $C$, based on an integrated likelihood, in which the mixing weights (i.e., columns of $W$) 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 $C$ 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 $C$, of the order $\sqrt{\log(n \vee d)/n}$ under the fixed design setting where the mixing weights $W$ 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 $W$ are independently generated according to some unknown underlying distribution over the simplex, we show that, for sufficiently large $d$, $W$ 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 (Arora, Ge, and Moitra 2012; Ke and Wang 2017) 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, Ng, and Jordan 2003) models, since the proposed estimator is essentially the maximum likelihood estimator of $C$ from the LDA model, with a particular choice of prior on $W$. More generally, the LDA model with other prior choices on $W$ can be interpreted as maximizing the data likelihood while minimizing a weighted volume in which a nonuniform volume element is integrated over the convex hull of $C$ 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 $U = CW$ plays an important role in hyperspectral imaging analysis, in which each column of $U$ 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 $C$ (Winter 1999). Similar models also arise in reinforcement learning (Singh, Jaakkola, and Jordan 1995; Duan, Ke, and Wang 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 $1_k$ denote the all-ones vector of length $k$, and $e_f$ the $f$th column of the $k \times k$ identity matrix $1_k$. Let $\Delta^{k-1} = \{ x \in \mathbb{R}^k : 0 \leq x_1 \leq 1, \sum_{i=1}^k x_i = 1 \}$ denote the $(k-1)$-dimensional probability simplex. For a matrix $A_{p \times q} = (A_1, \ldots, A_q)$, let

$$\text{Conv}(A) = \{ x \in \mathbb{R}^p : x = Ax, \lambda \in \Delta^{q-1} \},$$
$$\text{cone}(A) = \{ x \in \mathbb{R}^p : x = Ax, \lambda \geq 0 \},$$
$$\text{aff}(A) = \{ x \in \mathbb{R}^p : x = Ax, \lambda^T 1_k = 1, \lambda \in \mathbb{R}^q \},$$

denote the convex polytope, simplicial cone and affine space generated by (the $q$ columns of) $A$, respectively. For $A \in \mathbb{R}^{p \times q}(p \geq q)$, we define $|\text{Conv}(A)|$ as the $(q-1)$-dimensional volume of Conv(A) on aff(A), which can be computed by the Cayley–Menger determinant or Lemma D.1 in Appendix D, supplementary materials. For any vector $x$, $x \geq a$ means $x$ is element-wisely greater than or equal to $a$. Denote $a \lor b$ and $a \land b$ as the larger and smaller number between $a$ and $b$, respectively. For any cone $C$, let $C^* = \{ x : x^T y \geq 0, \forall y \in C \}$ denote its dual cone. Recall some useful facts of dual cones (Donoho and Stodden 2004): (i) $\text{cone}(A)^* = \{ x \in \mathbb{R}^p : x^T A \geq 0 \}$; (ii) if $A$ and $\bar{A}$ are convex cones, and $A \subseteq \bar{A}$, then $A^* \subseteq \bar{A}^*$. 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 article 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 begin by defining the following distance metric between two topic matrices $C$ and $\tilde{C}$:

$$D(C, \tilde{C}) = \min_{\Pi} \| \tilde{C} - \Pi C \|_2,$$

where $\| \cdot \|_2$ denotes the spectral norm and $\Pi$ is a permutation matrix. Note that $D(C, \tilde{C}) = 0$ if and only if $\tilde{C} = C$, that is, $C$ and $\tilde{C}$ 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 $D$ coincides with the 2-Wasserstein distance between column vectors of $C$ and $\tilde{C}$.

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

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

$$CW = \bar{C}W$$
and $$|\text{Conv}(\bar{C})| \leq |\text{Conv}(C)|,$$  \[ \text{(3)} \]

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

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

Proposition 1. If the separability condition is satisfied on $W$, then $(C, W)$ 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 $K$, its boundary $bK$, and its dual cone $K^*$, which are defined below:

$$K = \{ x \in \mathbb{R}^k : \|x\|_2 \leq x^T 1_k \},$$
$$bK = \{ x \in \mathbb{R}^k : \|x\|_2 = x^T 1_k \},$$
and $$K^* = \{ x \in \mathbb{R}^k : x^T 1_k \geq \sqrt{k-1}\|x\|_2 \}.$$

Definition 2 (SS Condition). A matrix $W$ is sufficiently scattered, if it satisfies:

(S1) $\text{cone}(W)^* \subseteq K$, or equivalently, $\text{cone}(W) \supseteq K^*$;
(S2) $\text{cone}(W)^* \cap bK \subseteq \{ x \geq f = 1, \ldots, k, \lambda \geq 0 \}$.

It is easy to verify that the separability condition on $W$ implies $W$ to be sufficiently scattered. In fact, the separability condition on $W$ means that Conv($W$) = $\Delta^{k-1}$ fills up the entire simplex, and that cone($W)^* = \text{cone}(\Delta^{k-1})$ 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 $W$ is sufficiently scattered and $C$ is of rank $k$ (full column rank), then $(C, W)$ is identifiable.
Proof of Theorem 2 is given in the supplementary material (Section D.1). Here we give a sketch of the proof. Suppose \( CW = \tilde{CW} \). We have \( C = \tilde{C}B \), where \( B = WW^T (WW^T)^{-1} \). It suffices to show \( B \) is a permutation matrix, which we prove by verifying that any row of \( B \) is in \( \text{cone}(W)^* \cap bdK = \{ \lambda e, \lambda \geq 0 \} \) 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 \( \text{Conv}(C) \) (i.e., columns of \( C \)) to the convex hull of \( U \). Their notion of identifiability is equivalent to verifying that any row of \( B \) is in \( \text{cone}(W)^* \cap \text{bdK} = \{ \lambda e, \lambda \geq 0 \} \) and is also of unit length.

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 \( C \) should be a probability weight vector belonging to the simplex). See Figure 1(d) for an example \( (V = k = 3 \text{ and } C = I_2) \) where the SS condition does not hold but the model is identifiable. Since any alternative topic matrix \( C \) as a convex polytope with three vertices must be inside \( \Delta^2 \), due to the parameter constraint, \( I_3 \) is the only topic matrix enclosing all columns of \( U \) and is within simplex \( \Delta^2 \). However, the SS condition does not hold since, apparently, \( \text{cone}(W) \supseteq \mathcal{K}^* \) 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 \( \Delta^2 \), what is shown in Figure 2 is the intersection of the cones with the hyperplane \( x^T I_3 = 1 \). The mixing weights, \( w_1, \ldots, w_d \), are represented as blue dots. Other items related to Definition 2 are: \( bdK \) is the red circle, \( \mathcal{K}^* \) is the dark brown ball inscribed in the triangle, and \( \text{cone}(W)^* \) 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.

An equivalent form of Condition (S1) is \( \text{cone}(W) \supseteq \mathcal{K}^* \). 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 Figures 2(d) for a violation of (S1). In particular, the separability condition on \( W \) implies that the three vertices (blue circles) of the triangle are included in \( W \). As a consequence, \( \text{cone}(W)^* = \text{cone}(W) \) is the entire triangle, which is the most extreme/superfluous instance that satisfies the SS condition.

Condition (S1) ensures that \( \text{Conv}(C) \) 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 Figures 2(b)-(c).

The SS condition was first introduced by Huang, Fu, and Sidiropoulos (2016) to study the identifiability of topic models, where identifiability is ensured under the SS condition along with a minimal determinant on \( WW^T \). This condition is used differently in their work and ours: Huang, Fu, and Sidiropoulos (2016) impose the SS condition on rows of \( C \); we impose this condition on columns of \( W \). Although volume is not discussed in Huang, Fu, and Sidiropoulos (2016), imposing the SS condition on rows of \( C \) 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 \( \text{cone}(W) \supseteq \mathcal{K}^* \) in the SS condition is equivalent to verifying whether a convex polytope contains a ball (after being projected to \( \Delta^{k-1} \)), which is in general an NP-complete problem in computational geometry (Freund and Orlin 1985; Huang, Sidiropoulos, and Swami 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
Conv(W) \supseteq bdK^* \cap \Delta^{k-1}, we can independently choose M sample points uniformly from bdK^* \cap \Delta^{k-1} and check whether all of them are in Conv(W). If W satisfies the SS condition, then the M sampled points should belong to Conv(W); if W does not satisfy the SS condition, then, since the probability of each sampled point falling in Conv(W) is a fixed number, the probability of making a mistake decays exponentially in M. For real datasets where Conv(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 Conv(W).

3. Maximum Integrated Likelihood Estimation

Before introducing the proposed estimator for topic matrix C, let us describe some more notations and the data generating process. Let X = \{x(1), \ldots, 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 (C, W), word sequences from different documents are independent, with the word sequence from the ith document, x(i) = (x(i,1), \ldots, x(i,n)), being n iid samples from the categorical distribution \text{Cat}(u_i), where u_i = Cw_i is the V-dimensional probability vector in \Delta^{V-1}, and w_i = (w_{ij}, \ldots, w_{ik}) denotes the ith column of matrix W. We use f_n(\cdot | u_i) to denote the multinomial likelihood function of the ith document. Let c_j denote the jth topic vector, that is, the jth column of matrix C, for j = 1, 2, \ldots, k. Under this notation, we can express the word frequency vector u_i = \sum_{j=1}^k w_{ij}c_j associated with the ith document as a convex combination of the topic vectors, where w_i serves as the mixing weight vector.

3.1. Implicit Volume Minimization

Since our primary interest is on the topic matrix C, we can profile out the nuisance parameters w_i’s by integrating them with respect to some distribution, resulting an integrated likelihood function of C. After that, we can estimate C by maximizing the integrated likelihood (Barr, Liseo, and Wolpert 1999). We propose to integrate out w_i’s with respect to the uniform distribution over simplex \Delta^{k-1}, which induces a uniform distribution on u_i = Cw_i over Conv(C). This is because the linear transformation w \mapsto Cw has a constant Jacobian. The integrated likelihood can be formally written as follows:

\[ F_{n \times d}(C; X) = \prod_{i=1}^d \int_{\text{Conv}(C)} f_n(x(i) | u) \frac{du}{|\text{Conv}(C)|} \]

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

\[ \hat{C}_n = \arg \max_C F_{n \times d}(C; X), \]

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 w with respect to a Dirichlet(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 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 → ∞), in which we “observe” the true word-frequency vectors for the d documents: u_1^0, \ldots, u_d^0. In this ideal setting, from a standard Laplace approximation argument, the ith integral inside the product in (4) after rescaling by a factor of order n^{(V-1)/2} converges to \( 1 |u_i^0 \in \text{Conv}(C) \), and the MLE \( \hat{C} \) becomes:

\[
\arg \max_C \frac{\prod_{i=1}^d \mathbb{1}(u_i^0 \in \text{Conv}(C))}{|\text{Conv}(C)|} = \arg \max_C \frac{\prod_{i=1}^d \mathbb{1}(u_i^0 \in \text{Conv}(C))}{|\text{Conv}(C)|},
\]

where \( \mathbb{1}(-) \) is the indicator function. Therefore, maximizing the integrated likelihood function (4) is asymptotically equivalent to minimizing the volume of Conv(C) subject to the constraint that Conv(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 C can be viewed as a two-stage procedure: in the first stage, we estimate the (k-1)-dimensional hyperplane aff(C) in which the convex polytope of C lies; then in the second stage, we determine the boundary of Conv(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 Papadimitriou et al. 2000; Azar et al. 2001; McSherry 2001; Kleinberg and Sandler 2003, 2008; Anandkumar, Hsu, and Kakade 2012; Arora, Ge, and Moitra 2012; Ke and Wang 2017) can be explained under this general two-stage framework. For example, some papers (Azar et al. 2001; Kleinberg and Sandler 2003, 2008) aim only at recovering the column span of 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 Papadimitriou et al. 2000; Azar et al. 2001; McSherry 2001; Kleinberg and Sandler 2003, 2008; Anandkumar, Hsu, and Kakade 2012; Arora, Ge, and Moitra 2012; Ke and Wang 2017) can be explained under this general two-stage framework. For example, some papers (Azar et al. 2001; Kleinberg and Sandler 2003, 2008) aim only at recovering the column span of 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 Papadimitriou et al. 2000; Azar et al. 2001; McSherry 2001; Kleinberg and Sandler 2003, 2008; Anandkumar, Hsu, and Kakade 2012; Arora, Ge, and Moitra 2012; Ke and Wang 2017) can be explained under this general two-stage framework. For example, some papers (Azar et al. 2001; Kleinberg and Sandler 2003, 2008) aim only at recovering the column span of 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.
on anchor words, Arora, Ge, and Moitra (2012) or documents (separability condition on pure topic documents, Papadimitriou et al. 2000; McSherry 2001; Anandkumar, Hsu, and Kakade 2012) in their first stage, and then in their second stage recover the population-level term-document matrix (or the hyperplane $\text{aff}(C)$) 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 rely 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 (see, 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:

$$\frac{1}{|\text{Conv}(C)|^d} \prod_{i=1}^{d} \left( \int_{\text{Conv}(C)} \exp \left\{ -n D_{\text{KL}}(\hat{u}^{(i)} \mid \mid u) \right\} du \right),$$

(7)

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

$$\exp \left\{ -n \sum_{i=1}^{d} \| (1_V - P_C) \hat{u}^{(i)} \|_2^2 \right\} \cdot \frac{1}{|\text{Conv}(C)|^d} \times \prod_{i=1}^{d} \left( \int_{\text{Conv}(C)} \exp \left\{ -n \|u - P_C \hat{u}^{(i)}\|_2^2 \right\} du \right),$$

(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 $\hat{u}^{(i)}$.

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 $C$ that is closest to $\hat{u}^{(i)}$'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 (2003), Kleinberg and Sandler (2008), and 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 $\text{Conv}(C)$ that encloses the projections of $\hat{u}^{(i)}$'s onto the hyperplane $\text{aff}(C)$, so that the second term in (8) is maximized.

With the separability condition on $C$ or $W$, 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 $C$ has at least $(k - 1)$ zeros; consequently, columns of $C$ should be chosen from the intersection of $\text{aff}(C)$ and the simplex $\Delta^{V-1}$ 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 nonregular 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
3.3. Computing Maximum Integrated Likelihood Estimator

For computation, we employ an MCMC-EM algorithm to find the maximizer \( \hat{C}_o \) of the integrated likelihood objective (4) by augmenting the model with a set of latent variables \( Z = \{ Z_{ij} : i = 1, 2, \ldots, d, j = 1, 2, \ldots, n \} \), where, given the mixing weights \( w_i \), \( Z_{ij} \in \{1, 2, \ldots, k\} \) follows \( \text{Cat}(w_i) \) and is interpreted as the topic indicating variable for the \( i \)-th word \( x_i^{(j)} \) 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 \( \log p(X, Z | C) \), where the expectation is with respect to the distribution of latent variable \( Z \) after marginalizing out \( W \), and then an M-step of maximizing the expected log-likelihood function over topic matrix \( C \). An MCMC scheme is introduced in the E-step for sampling \( (Z, W) \) pairs from the joint conditional distribution of \( p(Z, W | X, C) \) 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, Ng, and Jordan 2003) with a particular choice of priors on \( W \). 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, Ng, and Jordan 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) and Terenin et al. (2018). According to the asymptotic results of stochastic EM algorithms in Nielsen (2000), the estimation of the topic matrix produced by our algorithm is guaranteed to converge to the proposed MLE, provided that \( W^0 \) 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 article, we confine the details, including derivations for the full algorithm, to the supplementary materials.

4. Finite-Sample Error Analysis

In this section, we study the finite-sample error bound and its implied asymptotic consistency of the proposed estimator \( \hat{C}_o \). We consider the fixed design setting where columns of \( W \) can take arbitrary positions in \( \Delta^{k-1} \) as long as a perturbed version of the SS condition described in the following is satisfied. For the stochastic setting where columns of \( W \) are generated from some distribution, the error analysis and consistency can be found from Section B in the supplementary materials. To avoid ambiguity, we use \( C^0, W^0, U^0 \) to denote the ground truth, and leave \( C, W, U \) 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 \((\alpha, \beta)\)-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 (\((\alpha, \beta)\)-SS Condition).** A matrix \( W \) is \((\alpha, \beta)\)-sufficiently scattered for some \( \alpha, \beta \geq 0 \), if it satisfies (S1) and

\[
\text{(S3)} \quad [\text{cone}(W)^*]_f^q \cap [bdK]^q \subseteq \{ x : ||x - \lambda e_f||_2 \leq \beta \lambda, \lambda \geq 0 \},
\]

where

\[
[\text{cone}(W)^*]_f^q = \{ x : x^T \lambda W \geq -\alpha ||x||_2 \} \quad \text{and} \quad [bdK]^q = \{ x : ||x||_2 \leq \alpha ||x||_2 \}
\]

are the \( \alpha \)-enlargements of \( \text{cone}(W)^* \) and \( bdK \), respectively.

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

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

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

While the SS condition requires cone($W$)$^*$ and $bdK$ to intersect exactly at the positive semiaxis rays $\{\lambda e_i, \lambda \geq 0\}$, the ($\alpha, \beta$)-SS condition requires the intersection of [cone($W$)$^*$]$^\alpha$ and $[bdK]^\beta$—the perturbed versions of cone($W$)$^*$ and $bdK$, respectively, with noise level $\alpha$—to be within distance $\beta$ away from the semiaxis rays. Note that ($\alpha, \beta$)-SS degenerates to the SS condition when $\alpha = \beta = 0$.

Intuitively, if a matrix $W$ has vertices sensitivity coefficient $\beta$ under noise level $\alpha$, 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 ($\alpha, \beta$)-SS conditions under different combinations of ($\alpha, \beta$).

**Proposition 3.** The followings are some properties of ($\alpha, \beta$)-SS condition and SS condition.

(i) If $\alpha \geq \alpha'$ and $\beta \leq \beta'$, then ($\alpha, \beta$)-SS implies ($\alpha', \beta'$)-SS.

(ii) If $W$ is ($\alpha, \beta$)-SS and Conv($W$) $\subseteq$ Conv($\hat{W}$), then $W$ is also ($\alpha, \beta$)-SS.

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

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

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

### 4.2. Error Analysis and Consistency

In this section, we consider the setting where columns of $W$ are fixed, and satisfy a set of conditions related to the noise perturbed SS condition discussed in the previous section. Note that the results in this section 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 (see, Section B in the supplementary materials). Before presenting our main results on the finite-sample error bound of the estimator $\hat{C}_n$, let us first state our assumptions.

**Assumptions.** Assume the following:

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

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

(A3) There exist $s$ columns of $W^0$ which are ($\alpha, \beta$)-SS with $\alpha \geq C_1 \sqrt{s \log(sd)/n}$, 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 $(1 - 3/(n \sqrt{d}))^d$,

$$D(\hat{C}_n, C^0) \leq D_1 \sqrt{s \log(n \lor d)}/n + D_2 \sqrt{s} \beta,$$

where $c, D_1$, and $D_2$ are positive constants. In particular, if $\beta \leq C_2 \sqrt{\log(sd)/n}$ where $C_2$ is a constant, then

$$D(\hat{C}_n, C^0) \leq D_1 \sqrt{s \log(n \lor d)}/n.$$

![Figure 4. Geometric view of ($\alpha, \beta$)-SS sliced at the hyperplane $x^T 1_k = 1$ (for $k = 3$). $W$ is the same in (a), (b), and (c) while the values of $\alpha$ and $\beta$ are different. In (b) and (c), we highlight the region (the dashed circle) that are in [cone($W$)$^*$]$^\alpha$ $\cap$ $[bdK]^\beta$ but not in [cone($\hat{W}$)$^*$] $\subseteq$ cone($\hat{W}$).](image-url)
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 $\mathbf{W}$ is that Conv$(U_0)$ 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) and Javadi and Montanari (2020).

Next, we discuss Assumption (A3) in detail. First, note that a subset of columns of $\mathbf{W}$ satisfying the $(\alpha, \beta)$-SS condition immediately implies the full matrix $\mathbf{W}$ 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 submatrix $\mathbf{W}$ 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) and Javadi and Montanari (2020).

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

Third, we discuss the parameter $s$, the smallest number of columns in $\mathbf{W}$ that are $(\alpha, \beta)$-SS, in Assumption (A3). The following proposition shows that when the columns of $\mathbf{W}$ are stochastically generated according to some underlying distribution over $\Delta^{k-1}$ 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 $n^{\alpha \beta} / \log(nv\sqrt{d})$ in the asymptotic setting where $(n, d) \to \infty$.

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

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

Assumptions. Assume the following:

(A3) For any sufficiently small $\varepsilon > 0$, there exists some $\beta_\varepsilon$ such that $\beta_\varepsilon \to 0$ as $\varepsilon \to 0$, and there are $s$ columns of $\mathbf{W}$ satisfying the $(\varepsilon, \beta_\varepsilon)$-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

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

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

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

Note that Proposition 5 again provides a set of sufficient conditions for Assumption (A3'). However, our current condition on $\mathbf{W}$ 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 of grows, which is seemingly weaker than some existing results, such as Arora, Ge, and Moitra (2012), Bansal, Bhattacharyya, and Kannan (2014), Anandkumar et al. (2014), Ke and Wang (2017), and Wang (2019). In particular, under the anchor word assumption, Arora, Ge, and Moitra (2012) and Ke and Wang (2017) showed an 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 2003, 2008; Ke and Wang 2017) to find a $(k-1)$-dimensional hyperplane best approximating the data whose error bound is $1/\sqrt{vd}$. 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 the $(k-1)$-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) and Brunel, Klusowski, and Yang (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 \( \mathbb{P}_C^{(i)} \hat{u}^{(i)} \) onto \( \text{aff}(C) \) of the sample word frequency vector \( \hat{u}^{(i)} \) for document \( i \), for \( i = 1, \ldots, d \), plays the role of a noisy measurement from the convex polytope \( \text{Conv}(C) \). 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 \( \text{aff}(C) \) obtained in the first stage as well as the location of \( \mathbb{P}_C^{(i)} \hat{u}^{(i)} \) on the data simplex \( \Delta^{V-1} \). 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 \( \beta_0 = 1_k \) 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: \( k = V = 3 \), \( C^0 = I_3 \), and the number of words for each document is sampled from Poisson(2000). For the true matrix \( W^0 \), we consider four different configurations for \( w^0_k \): (a) concentrated in the center of \( \Delta^2 \); (b) concentrated in the bottom right; (c) satisfying the SS condition; (d) spread around three vertices. The four configurations are displayed in Figure 5, where the black dots denote \( \hat{w}^0_1 \) and the large black triangle represents \( \text{Conv}(C^0) = \Delta^2 \). In cases (a), (b), and (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 \( \text{Conv}(\hat{C}_n) \) 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)), \( \text{Conv}(C^0) \) 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 \( w^0 \) 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 \( W^0 \) is stochastically generated from a distribution other than uniform.

In this simulation study we consider the following setup: \( k = 3, V = 1000, d = 200, C^0 \sim \text{Dirichlet}_1(1) \), and the number of words for each document is generated from Poisson(20000). The true mixing weights \( W^0 \) are stochastically generated from the following distributions: (a) Dirichlet_1(1); (b) uniformly from 10 Euclidean balls whose centers satisfy the SS condition; (c) a mixture of Dirichlet distributions: \( 0.2 \times \text{Dirichlet}(10, 1, 1) + 0.2 \times \text{Dirichlet}(0.1, 1, 1) + 0.2 \times \text{Dirichlet}(10, 10, 1) + 0.2 \times \text{Dirichlet}(0.1, 0.1, 1) + 0.2 \times \text{Dirichlet}(1, 2, 3) \).

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 \( \hat{C}_n \) over 100 replications. The results are reported in Table 1. We can see that in all three cases, our proposed estimator outperforms other estimators.

5.1.3. Convergence of the Estimation

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

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 \( n = \)
and the 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.

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

We observe that Conv($\hat{C}_n$) tends to cover these density balls while maintaining its volume small. Recall that $\hat{C}_n = \arg \max_{C} \prod_{i=1}^{d} \frac{f_n(x^{(i)} | u = C w_i)}{\text{Conv}(C)} du$. Conv($\hat{C}_n$) 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 $\hat{x}^{(i)}$. Therefore, Conv($\hat{C}_n$) tends to trade off its volume for a larger coverage of the density balls. In this case, the true means $u^{(i)}$ are all located on the boundary of Conv($C^0$); to fulfill the SS condition, a fraction of each circle thus lies outside Conv($C^0$). Consequently, the averaged likelihood density over Conv($\hat{C}_n$) is larger than that of Conv($C^0$), though $|\text{Conv}(\hat{C}_n)| > |\text{Conv}(C^0)|$. As proved in Theorem 4, the convergence rate of $\hat{C}_n$ in the order of $\sqrt{\log(n + d)/n}$ is slightly slower than that of $\hat{u}^{(i)}$, 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.

---

**Table 1. Relative RMSE (Simulation 2).**

| Priors | $(1, 1, 1)$ | $(0.1, 0.1, 0.1)$ | $(10, 1, 1)$ | $(0.1, 0.1, 1)$ | $(1, 0.1, 1)$ | $(0.1, 1, 1)$ | $(1, 0.1, 0.1)$ | $(0.1, 1, 0.1)$ | $(1, 2, 3)$ | $(3, 3, 3)$ |
|--------|-------------|-----------------|--------------|---------------|--------------|--------------|---------------|--------------|-------------|-------------|
| 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 |           |           |

The best results are highlighted in bold-face.
5.2.1 Text Datasets

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

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

| NIPS | NIPS | NIPS | NIPS | NIPS |
|------|------|------|------|------|
| k = 5 | k = 10 | k = 15 | k = 20 | k = 25 |
| AnchorF | GDM | Gibbs | pcLDA | MC2-EM |
| -904 | -1954 | -2935 | -3664 | -04 |
| -501 | -1083 | -1770 | -2409 | -67 |
| -365 | -960 | -1648 | -2314 | -43 |
| -355 | -942 | -1599 | -2373 | -25 |
| -342 | -1659 | -1573 | -2254 | -05 |
| -699 | -1551 | -2727 | -3942 | -21 |
| -643 | -1708 | -2307 | -3182 | -25 |
| -752 | -1609 | -2465 | -3840 | -25 |
| -709 | -1614 | -2380 | -3115 | -29 |
| -723 | -1573 | -2411 | -3299 | -07 |

| Similarity counts | Perplexity score |
|-------------------|-----------------|
| k = 5 | k = 10 | k = 15 | k = 20 | k = 25 |
| Topic coherence | AnchorF | GDM | Gibbs | pcLDA | MC2-EM |
| 24 | 69 | 102 | 154 | 4431 |
| 25 | 44 | 98 | 161 | 2955 |
| 26 | 63 | 99 | 134 | 2256 |
| 26 | 67 | 99 | 155 | 2183 |
| 24 | 63 | 102 | 147 | 2182 |
| 24 | 66 | 78 | 153 | 1755 |
| 25 | 90 | 103 | 143 | 1758 |
| 25 | 57 | 90 | 1342 | 1724 |
| 25 | 90 | 1342 | 1342 | 107 |

The best results are highlighted in bold-face.

5.2.1. Text Datasets

The NIPS dataset contains $V = 11463$ unique words and $d = 5811$ NIPS conference papers, with an average document length of 1902 words. The Daily Kos dataset 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 preprocessing procedures.

We compare the performance of our algorithm (MC2-EM) with the following baseline algorithms: Anchor Free (AnchorF) (Huang, Fu, and Sidiropoulos 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 hyperparameters 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 $\sum_{l=1}^{k} \sum_{i,j} \log \left( \frac{\text{freq}(v_{i},v_{j})+\epsilon}{\text{freq}(v_{j})} \right)$, where $\text{freq}(v_{i},v_{j})$ is the leading 20 words for topic $l$, $\text{freq}(\cdot)$ is the occurrence count, and $\epsilon$ is a small constant added to avoid numerical issues. Similarity Count is used to measure similarity between topics (Arora et al. 2013; Huang, Fu, and Sidiropoulos 2016); it is obtained simply by adding up the overlapped words across $V_{l}$. Perplexity Score is used to measure 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, supplementary materials.)

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 $\hat{U}$. Since the topic matrix $\hat{U}$ is assumed to have full rank (Theorem 2), the true term-document matrix $U$ has rank $k$. By Weyl’s inequality (Weyl 1912), the singular values of $\hat{U}$ are expected to be close to those of $U$. Therefore, we can plot the ordered singular values of $\hat{U}$ versus its index, and then select $k$ by detecting the location of a significant drop of the curve. See Appendix F, supplementary materials for a simulation illustrating this approach.

The results are summarized in Tables 2 and 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, supplementary materials). The best score in each case is highlighted in boldface. Overall, our estimator (MC2-EM) gives promising results. For all three metrics in both datasets, it gives the highest score or a score close to the highest.

https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/
The best results are highlighted in bold-face.

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, Jaakkola, and Jordan 1995) is equivalent to estimating a topic model.

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

\[
P(X_{t+1}|X_t) = \sum_{l=1}^{k} P(Z_t = l|X_t) \cdot P(X_{t+1}|Z_t = l),
\]

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

In this section, we consider a New York taxi-trip\(^4\) dataset. This dataset contains \( \sum_{t=1}^{d} n_t = 7,667,792 \) New York City yellow cab trips in January 2019. The location information is discretized into \( V = 263 \) 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, Ke, and Wang (2019), we aim to merge the \( V = 263 \) taxi zones into meta-states via soft state aggregation.

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, \( P(X_{t+1}|Z_t = l) \). We observe that the drop-off traffic is decomposed into three clusters, (a) downtown Manhattan, (b) west midtown Manhattan, and (c) 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, \( P(Z_t = l|X_t) \). The three meta states consist of (a) Staten Island, Brooklyn, Queens, and downtown Manhattan; (b) uptown Manhattan and Bronx; and (c) east midtown Manhattan. Note that the scales of the estimates for \( C \) and \( W \) are quite different. In specific, the sum of values over each map of the top three is 1 since \( \sum_{l=1}^{3} P(X_{t+1} = l|Z_t = l) = 1, l = 1, 2, 3 \), while the sum of values for each zone over the bottom three maps is 1 since \( \sum_{v=1}^{V} P(Z_t = l|X_t = v) = 1, v = 1, \ldots, V \). 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

---

\(^4\)https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
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 article, 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: (a) exploring a sufficient and necessary condition for model identifiability, as the SS condition is not necessary; (b) providing explicit verifiable sufficient conditions for the \((\alpha, \beta)\)-SS condition—we conjecture that the \((\alpha, \beta)\)-SS condition can be implied by the SS condition; (c) 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.

Supplementary Materials

Additional discussions, analysis, algorithms, technical proofs and computational results can be found in the supplementary materials.

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. (2012), “A Spectral Algorithm for Latent Dirichlet Allocation,” in Advances in Neural Information Processing Systems, pp. 917–925. [2861]
Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014), “Tensor Decompositions for Learning Latent Variable Models,” Journal of Machine Learning Research, 15, 2773–2832. [2869]
Anandkumar, A., Hsu, D., and Kakade, S. M. (2012), “A Method of Moment for Mixture Models and Hidden Markov Models,” in Conference on Learning Theory, pp. 33–1. JMLR Workshop and Conference Proceedings. [2861,2865,2866]
Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013), “A Practical Algorithm for Topic Modeling with Provable Guarantees,” in International Conference on Machine Learning, pp. 280–288. [2872]
Arora, S., Ge, R., and Moitra, A. (2012), “Learning Topic Models–Going Beyond SVD,” in 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, pp. 1–10. [2861,2862,2866,2869]
Azar, Y., Fiat, A., Karlin, A., McSherry, F., and Saia, J. (2001), “Spectral Analysis of Data,” in Proceedings of the Thirty-Third Annual ACM Symposium on Theory of computing, pp. 619–626. [2861,2862,2866,2869]
Bansal, T., Bhattacharyya, C., and Kannan, R. (2014), “A Provable SVD-based Algorithm for Learning Topics in Dominant Admixture Corpus,” Advances in Neural Information Processing Systems, 27, 1997–2005. [2869]
Berger, J. O., Liseo, B., and Wolpert, R. L. (1999), “Integrated Likelihood Methods for Eliminating Noise Parameters,” Statistical Science, 14, 1–28. [2865]
Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [2860,2862,2867]
Brunel, V.-E., Klusowski, J. M., and Yang, D. (2021), "Estimation of Convex Supports from Noisy Measurements," *Bernoulli*, 27, 772–793. [2869]
Craig, M. D. (1994), "Minimum-Volume Transforms for Remotely Sensed Data," *IEEE Transactions on Geoscience and Remote Sensing*, 32, 542–552. [2861]
Donoho, D., and Stodden, V. (2004), "When does Non-negative Matrix Factorization Give a Correct Decomposition into Parts?" in *Advances in Neural Information Processing Systems*, pp. 1141–1148. [2861,2863]
Duan, Y., Ke, T., and Wang, M. (2019), "State Aggregation Learning from Markov Transition Data," in *Advances in Neural Information Processing Systems*, pp. 4488–4497. [2862,2873]
Freund, R. M., and Orlin, J. B. (1985), "On the Complexity of Four Polyhedral Set Containment Problems," *Mathematical Programming*, 33, 139–145. [2864]
Fu, X., Ma, W.-K., Huang, K., and Sidiropoulos, N. D. (2015), "Blind Separation of Quasi-Stationary Sources: Exploiting Convex Geometry in Covariance Domain," *IEEE Transactions on Signal Processing*, 63, 2306–2320. [2861]
Ge, R., and Zou, J. (2015), "Intersecting Faces: Non-negative Matrix Factorization with New Guarantees," in *International Conference on Machine Learning*, pp. 2295–2303. [2861]
Goldenshluger, A., and Tsybakov, A. (2004), "Estimating the Endpoint of a Distribution in the Presence of Additive Observation Errors," *Statistics & Probability Letters*, 68, 39–49. [2869]
Griffiths, T. L., and Steyvers, M. (2004), "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101, 5228–5235. [2867,2872]
Hofmann, T. (1999), "Probabilistic Latent Semantic Analysis," in *Uncertainty in Artificial Intelligence*, pp. 289–296. [2860]
Huang, K., Fu, X., and Sidiropoulos, N. D. (2016), "Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm," in *Advances in Neural Information Processing Systems*, pp. 1786–1794. [2862,2864,2872]
Huang, K., Sidiropoulos, N. D., and Swami, A. (2014), "Non-negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition," *IEEE Transactions on Signal Processing*, 62, 211–224. [2864]
Jang, B., and Hero, A. (2019), "Minimum Volume Topic Modeling," in *International Conference on Artificial Intelligence and Statistics*, pp. 3013–3021. [2862]
Javadi, H., and Montanari, A. (2020), "Nonnegative Matrix Factorization via Archetypal Analysis," *Journal of the American Statistical Association*, 115, 896–907. [2861,2862,2864,2869]
Ke, Z. T., and Wang, M. (2017), "A New SVD Approach to Optimal Topic Estimation," arXiv preprint arXiv:1704.07016. [2861,2862,2865,2866,2869]
Kleinberg, J., and Sandler, M. (2003), "Convergent Algorithms for Collaborative Filtering," in *Proceedings of the 4th ACM conference on Electronic Commerce*, pp. 1–10. [2861,2865,2866,2869]
Kleinberg, J., and Sandler, M. (2008), "Using Mixture Models for Collaborative Filtering," *Journal of Computer and System Sciences*, 74, 49–69. [2861,2865,2866,2869]
Magnusson, M., Jonsson, L., Villani, M., and Broeman, D. (2018), "Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models," *Journal of Computational and Graphical Statistics*, 27, 449–463. [2867,2872]
McSherry, F. (2001), "Spectral Partitioning of Random Graphs," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE. [2861,2865,2866]
Miao, L., and Qi, H. (2007), "Endmember Extraction from Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization," *IEEE Transactions on Geoscience and Remote Sensing*, 43, 898–910. [2861]
Nguyen, X. (2015), "Posterior Contraction of the Population Polytope in Finite Admixture Models," *Bernoulli*, 21, 618–646. [2861,2869]
Nielsen, S. F. (2000), "The Stochastic EM Algorithm: Estimation and Asymptotic Results," *Bernoulli*, 6, 457–489. [2867]
Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (2000), "Latent Semantic Indexing: A Probabilistic Analysis," *Journal of Computer and System Sciences*, 61, 217–235. [2861,2865,2866]
Recht, B., Re, C., Tropp, J., and Bittorf, V. (2012), "Factoring Nonnegative Matrices with Linear Programs," in *Advances in Neural Information Processing Systems*, pp. 1214–1222. [2861]
Singh, S. P., Jaakkola, T., and Jordan, M. I. (1995), "Reinforcement Learning with Soft State Aggregation," *Advances in Neural Information Processing Systems*, pp. 361–368. [2862,2873]
Tang, J., Meng, Z., Nguyen, X., Mei, Q., and Zhang, M. (2014), "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis," in *International Conference on Machine Learning*, pp. 190–198. [2861]
Terenin, A., Magnusson, M., Jonsson, L., and Draper, D. (2018), "Polya Urn Latent Dirichlet Allocation: A Doubly Sparse Massively Parallel Sampler," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 1709–1719. [2867,2872]
Wang, Y. (2019), "Convergence Rates of Latent Topic Models Under Relaxed Identifiability Conditions," *Electronic Journal of Statistics*, 13, 37–66. [2861,2869]
Weyl, H. (1912), "Das Asymptotische Ver teilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, 71, 441–479. [2872]
Winter, M. E. (1999), "N-findr: An Algorithm for Fast Autonomous Spectral End-member Determination in Hyperspectral Data," in *Imaging Spectrometry V* (Vol. 3753), pp. 266–275, International Society for Optics and Photonics. [2862]
Yurochkin, M., and Nguyen, X. (2016), "Geometric Dirichlet Means Algorithm for Topic Inference," in *Advances in Neural Information Processing Systems*, pp. 2505–2513. [2872]