A Provably Efficient Algorithm for Separable Topic Discovery

Weicong Ding, Prakash Ishwar, Senior Member, IEEE, and Venkatesh Saligrama, Senior Member, IEEE

Abstract—We develop necessary and sufficient conditions and a novel provably consistent and efficient algorithm for discovering topics (latent factors) from observations (documents) that are realized from a probabilistic mixture of shared latent factors that have certain properties. Our focus is on the class of topic models in which each shared latent factor contains a novel word that is unique to that factor, a property that has come to be known as separability. Our algorithm is based on the key insight that the novel words correspond to the extreme points of the convex hull formed by the row-vectors of a suitably normalized word co-occurrence matrix. We leverage this geometric insight to establish polynomial computational and sample complexity bounds based on a few isotropic random projections of the rows of the normalized word co-occurrence matrix. Our proposed random-projection-based algorithm is naturally amenable to an efficient distributed implementation and is attractive for modern web-scale distributed data mining applications.

Index Terms—Topic Modeling, Separability, Random Projection, Solid Angle, Necessary and Sufficient Conditions.

I. INTRODUCTION

TOPIC modeling refers to a family of generative models and associated algorithms for discovering the (latent) topical structure shared by a large corpus of documents. They are important for organizing, searching, and making sense of a large text corpus [1]. In this paper we describe a novel geometric approach, with provable statistical and computational efficiency guarantees, for learning the latent topics in a document collection. This is a culmination of a series of recent publications on certain structure-leverageing methods for topic modeling with provable theoretical guarantees [2]–[5].

We consider a corpus of $M$ documents, indexed by $m = 1, \ldots, M$, each composed of words from a fixed vocabulary of size $W$. The distinct words in the vocabulary are indexed by $w = 1, \ldots, W$. Each document $m$ is viewed as an unordered “bag of words” and is represented by an empirical $W \times 1$ word-counts vector $X^m$, where $X_{w,m}$ is the number of times that word $w$ appears in document $m$ [1], [5]–[7]. The entire document corpus is then represented by the $W \times M$ matrix $X = [X^1, \ldots, X^M]$.

A “topic” is a $W \times 1$ distribution over the vocabulary. A topic model posits the existence of $K$ latent topics that are shared among all $M$ documents in the corpus. The topics can be collectively represented by the $K$ columns $\beta^1, \ldots, \beta^K$ of a $W \times K$ column-stochastic “topic matrix” $\beta$. Each document $m$ is conceptually modeled as being generated independently of all other documents through a two-step process: 1) first draw a $K \times 1$ document-specific distribution over topics $\theta^m$ from a prior distribution $Pr(\theta)$ on the probability simplex with some hyper-parameters $\alpha$; 2) then draw $N$ iid words from a $W \times 1$ document-specific word distribution over the vocabulary given by $A^m = \sum_{k=1}^{K} \theta^k \beta_{k,m}$ which is a convex combination (probabilistic mixture) of the latent topics. Our goal is to estimate $\beta$ from the matrix of empirical observations $X$. To appreciate the difficulty of the problem, consider a typical benchmark dataset such as a news article collection from the New York Times (NYT) [8] that we use in our experiments. In this dataset, after suitable pre-processing, $W = 14,943$, $M = 300,000$, and on average $N = 298$. Thus, $N \ll W \ll M$, $X$ is very sparse, and $M$ is very large. Typically, $K \approx 100 \ll \min(W, M)$.

This estimation problem in topic modeling has been extensively studied. The prevailing approach is to compute the MAP/ML estimate [1]. The true posterior of $\beta$ given $X$, however, is intractable to compute and the associated MAP and ML estimation problems are in fact NP-hard in the general case [9], [10]. This necessitates the use of sub-optimal methods based on approximations and heuristics such as Variational Bayes and Markov Chain Monte Carlo (MCMC) [6], [11]–[13]. While they produce impressive empirical results on many real-world datasets, guarantees of asymptotic consistency or efficiency for these approaches are either weak or non-existent. This makes it difficult to evaluate model fidelity: failure to produce satisfactory results in new datasets could be due to the use of approximations and heuristics or due to model misspecification which is more fundamental. Furthermore, these sub-optimal approaches are computationally intensive for large text corpora [5], [7].

To overcome the hardness of the topic estimation problem in its full generality, a new approach has emerged to learn the topic model by imposing additional structure on the model parameters [3], [5], [7], [9], [14], [15]. This paper focuses on a key 1Depending on the context, we will use $X_w,m$ to represent either the empirical word-count or, by suitable column-normalization of $X$, the empirical word-frequency.
structural property of the topic matrix $\beta$ called topic separability [3], [5], [7], [9], [15] wherein every latent topic contains at least one word that is novel to it, i.e., the word is unique to that topic and is absent from the other topics (novel words are called anchor words in [7], [9]). This is, in essence, a property of the support of the latent topic matrix $\beta$. The topic separability property can be motivated by the fact that for many real-world datasets, the empirical topic estimates produced by popular Variational Bayes and Gibbs Sampling approaches are approximately separable [5], [7]. Moreover, it has recently been shown that the separability property will be approximately satisfied with high probability when the dimension of the vocabulary scales sufficiently faster than the number of topics $K$ and $\beta$ is a realization of a Dirichlet prior that is typically used in practice [16]. Therefore, separability is a natural approximation for most high-dimensional topic models.

Our approach exploits the following geometric implication of the key separability structure. If we associate each word in the vocabulary with a row-vector of \{an $\ell_1$-normalized\} empirical word co-occurrence matrix, the set of novel words correspond to the extreme points of the convex hull formed by the row-vectors of all words. We leverage this geometric insight and develop a provably consistent and efficient algorithm. Informally speaking, we establish the following result:

**Theorem 1:** If the topic matrix is separable and the mixing weights satisfy a minimum information-theoretically necessary technical condition, then our proposed algorithm runs in polynomial time in $M, W, N, K$, and estimates the topic matrix consistently as $M \to \infty$ with $N \geq 2$ held fixed. Moreover, our proposed algorithm can estimate $\beta$ to within an $\epsilon$ element-wise error with a probability at least $1 - \delta$ if $M \geq \text{Poly}(W, 1/N, K, \log(1/\delta), 1/\epsilon)$.

The asymptotic setting $M \to \infty$ with $N$ held fixed is motivated by text corpora in which the number of words in a single document is small while the number of documents is large. We note that our algorithm can be applied to any family of topic models whose topic mixing weights prior $\Pr(\alpha)$ satisfies a minimum information-theoretically necessary technical condition. In contrast, the standard Bayesian approaches such as Variational Bayes or MCMC need to be hand-designed separately for each specific topic mixing weights prior.

The highlight of our approach is to identify the novel words as extreme points through appropriately defined random projections. Specifically, we project the row-vector of each word in an $\ell_1$-normalized word co-occurrence matrix along a few independent and isotropically distributed random directions. The fraction of times that a word attains the maximum value along a random direction is a measure of its degree of robustness as an extreme point. This process of random projections followed by counting the number of times a word is a maximizer can be efficiently computed and is robust to the perturbations induced by sampling noise associated with having only a very small number of words per document $N$. In addition to being computationally efficient, it turns out that this random projections based approach 1) requires the minimum information-theoretically necessary technical conditions on the topic prior for asymptotic consistency, and 2) can be naturally parallelized and distributed. As a consequence, it can provably achieve the efficiency guarantees of a centralized method while requiring insignificant communication between distributed document collections [5]. This is attractive for web-scale topic modeling of large distributed text corpora.

Another advance of this paper is the identification of necessary and sufficient conditions on the mixing weights for consistent topic estimation in separable topic models. In previous work we showed that a simplicial condition on the mixing weights is both necessary and sufficient for consistently detecting all the novel words [4]. In this paper we complete the characterization by showing that an affine independence condition on the mixing weights is necessary and sufficient for consistently estimating a separable topic matrix. These conditions are satisfied by practical choices of topic priors such as the Dirichlet distribution [6]. All these necessary conditions are information-theoretic and algorithm-independent, i.e., they are irrespective of the specific statistics of the observations or the algorithms that are used. The provable statistical and computational efficiency guarantees of our proposed algorithm hold true under these necessary and sufficient conditions.

The rest of this paper is organized as follows. We review related work on topic modeling as well as the separability property in various domains in Sec. II. We introduce the separability property on $\beta$, the simplicial and affine independence conditions on mixing weights, and the extreme point geometry that motivates our approach in Sec. III. We then discuss how the solid angle can be used to identify robust extreme points to deal with a finite number of samples (words per document) in Sec. IV. We describe our overall algorithm and sketch its analysis in Sec. V. We demonstrate the performance of our approach in Sec. VI on various synthetic and real-world examples. Due to space constraints, proofs of all results are presented in [17].

II. Related Work

The idea of modeling text documents as mixtures of a few semantic topics was first proposed in [18] where the mixing weights were assumed to be deterministic. Latent Dirichlet Allocation (LDA) [6] extended this to a probabilistic setting by modeling topic mixing weights using Dirichlet priors. This setting has been further extended to include other topic priors such as the log-normal prior in the Correlated Topic Model [19]. LDA models and their derivatives have been successful in achieving good empirical performance on a wide range of problems [1], [13].

The prevailing approaches for estimation and inference problems in topic modeling are based on MAP or ML estimation [1]. However, the computation of posterior distributions conditioned on observations $X$ is intractable [6]. Moreover, the MAP estimation objective is non-convex and has been shown to be $NP$-hard [9], [10]. Therefore, various approximation and heuristic strategies have been employed. These approaches fall into two major categories – sampling approaches and optimization approaches. Most sampling approaches are based on MCMC algorithms that are carefully designed to generate (approximately) independent samples from a Markov Chain such that the sample distribution converges to the true posterior [11], [20]. Optimization approaches are typically based on
the so called Variational-Bayes methods. These methods optimize the parameters of a simpler parametric distribution so that it is close to the true posterior in terms of KL divergence [6], [12]. Expectation-Maximization-type algorithms are typically used in these methods. In practice, while both Variational Bayes and MCMC algorithms have similar performance, Variational Bayes is typically faster than MCMC [1], [21].

Non-negative Matrix Factorization (NMF) is an alternative approach for topic estimation. NMF-based methods exploit the fact that both the topic matrix $\beta$ and the mixing weights are non-negative and attempt to decompose the empirical observation matrix $X$ into a product of a nonnegative topic matrix $\beta$ and the matrix of mixing weights by minimizing a cost function of the form [21]–[24], $\sum_{m=1}^{M} d(X^m, \beta^m) + \lambda \psi(\beta, \theta_1, \ldots, \theta_M)$, where $d(,)$ is some measure of closeness and $\psi$ is a regularization term which enforces desirable properties, e.g., sparsity, on $\beta$ and the mixing weights. The NMF problem, however, is also known to be non-convex and $\mathcal{NP}$-hard [25] in general. Sub-optimal strategies such as alternating minimization, greedy gradient descent, and heuristics are used in practice [23].

In contrast to the above approaches, a new approach has recently emerged which is based on imposing additional structure on the model parameters [3], [5], [7], [9], [14], [15]. These approaches show that the topic discovery problem lends itself to provably consistent and polynomial-time solutions by making assumptions about the structure of topic matrix $\beta$ and the distribution of the mixing weights. In this category of approaches are methods based on a tensor decomposition of the moments of $X$ [14], [26]. The algorithm in [26] uses second order empirical moments and is shown to be asymptotically consistent when the topic matrix $\beta$ has a special sparsity structure. The algorithm in [14] uses the third order tensor of observations. It is, however, strongly tied to the specific structure of the Dirichlet prior on the mixing weights and requires knowledge of the concentration parameters of the Dirichlet distribution. Furthermore, in practice these approaches are computationally intensive and require some initial coarse dimensionality reduction, gradient descent speedups, and GPU acceleration to process large-scale text corpora like the NYT dataset [14].

Our work falls into the family of approaches that exploit the separability property of $\beta$ and its geometric implications [3]–[5], [7], [9], [15], [27], [28]. A provably asymptotically consistent topic estimation algorithm with polynomial statistical and computational efficiency guarantees was first proposed in [9] based on a simplicial condition on the topic weights. However, this method requires solving $W$ linear programs, each with $W$ variables and is computationally impractical for large text corpora [7, p.5]. Subsequent work improves the computational efficiency [15], [24], but theoretical guarantees of asymptotic consistency (when $N$ fixed, and the number of documents $M \to \infty$) are not available. Algorithms in [7] and [3] are both practical and provably consistent and have polynomial statistical and computational efficiency guarantees. Both algorithms require a stronger, but slightly different, technical condition on the topic mixing weights than the simplicial condition in [9]. Specifically, [7] imposes a full-rank condition on the second-order correlation matrix of the mixing weights and proposes a Gram-Schmidt procedure to identify the extreme points. Similarly, [3] imposes a diagonal-dominance condition on the same second-order correlation matrix and proposes a random projections based procedure. These procedures are tied to the specific conditions imposed and may fail to both detect the novel words of some topics and estimate the topics when the imposed conditions (which are sufficient, but not necessary for novel word detection or topic estimation) fail to hold [5, Sec.6.2].

The present work is most closely related to [7], [9], and [4], but has major differences. While the simplicial condition used in [4] and this work is equivalent to that used in [9], neither [9] nor [7] nor any other works discuss the necessity of the simplicial condition for either novel word detection or topic estimation in separable topic models. The necessity of the simplicial condition for novel word detection in separable topic models is first investigated in [4]. In this paper, we go beyond this and show that the simplicial condition is not sufficient for topic estimation in separable topic models (even though it is necessary for novel word detection). In [4] a full-rank condition is required for consistent topic estimation. In contrast, here we identify the affine independence condition as the minimal necessary and sufficient condition for asymptotically consistent topic estimation in separable topic models. The present work is the first to provide a complete picture of the necessary and sufficient conditions for detecting novel words as well as estimating topics in separable topic models. The investigation of necessary and sufficient conditions helps us to understand the limitations of existing algorithms and develop new ones that require only the minimal assumptions and have the broadest practical applicability. In fact, the motivation for the present work came from an empirical observation that the algorithm in [7] fails in some datasets where the simplicial condition is not satisfied whereas the random projections based algorithm does not.

Algorithmically, random projections are used for very different purposes in [7] and in our work. In [7], random projections are used as a dimensionality-reduction device for purely computational benefits: the algorithm in [7] is distance-based and high-dimensional pairwise distances are preserved with high probability under random projections (Johnson-Lindenstrauss). In contrast, our algorithm is counts-based and operates in a separable manner with respect to the projection directions, meaning that the processing is done one projection coordinate at a time in contrast to a distance based method. Random projections are integral to our entire algorithmic approach. As explained towards the end of Section IV-C, random projections also enable an efficient parallel distributed/online implementation of our algorithm. In contrast, it is not straightforward to parallelize the algorithm in [7] because it is based on a Gram-Schmidt-like sequential procedure. Table I summarizes the major differences between these closely related works.

We note that the separability property has been exploited in other recent work as well [27], [28]. In [28], a singular value decomposition based approach is proposed for topic estimation.
In [27], it is shown that the standard Variational-Bayes approximation can be asymptotically consistent if $\beta$ is separable. However, the additional constraints proposed essentially boil down to the requirement that each document contain predominantly only one topic. In addition to assuming the existence of such “pure” documents, [27] also requires a strict initialization. It is thus unclear how this can be achieved using only the observations $X$.

The separability property has been re-discovered in different fields and has found application in several problems. To the best of our knowledge, this concept was first introduced as the Pure Pixel Index assumption (pixels contain predominantly one species) in the Hyperspectral Image unmixing problem [29]. Separability has also been studied in the NMF literature in the context of ensuring the uniqueness of NMF [30]. Subsequent works have developed NMF algorithms that exploit separability [24], [31]. The uniqueness and correctness results in this line of work has primarily focused on the noiseless case. We finally note that separability has also been recently exploited in the problem of learning multiple ranking preferences from pairwise comparisons for personal recommendation systems and information retrieval [32], [33] and has led to provably consistent and efficient estimation algorithms.

III. Topic Separability, Necessary and Sufficient Conditions, and the Geometric Intuitions

In this section, we unravel the key ideas that motivate our algorithmic approach by focusing on the ideal case where there is no “sampling-noise”, i.e., each document is infinitely long ($N = \infty$). In the next section, we will turn to the finite $N$ case. We recall that $\beta$ and $X$ denote the $W \times K$ topic matrix and the $W \times M$ empirical word counts/frequency matrix respectively. Also, $M, W,$ and $K$ denote, respectively, the number of documents, the vocabulary size, and the number of topics. For convenience, we group the document-specific mixing weights, the $\theta^{\text{mix}}$’s, into a $K \times M$ weight matrix $\theta = [\theta^{1}, \ldots, \theta^{M}]$ and the document-specific distributions, the $A^{\text{mix}}$’s, into a $W \times M$ document distribution matrix $A = [A^{1}, \ldots, A^{M}]$. The generative procedure that describes a topic model then implies that $A = \beta \theta$. In the ideal case considered in this section ($N = \infty$), the empirical word frequency matrix $X = A$. Notation: A vector $a$ without specification will denote a column-vector, $1$ the all-ones column vector of suitable size, $X_i$ the $i$-th column vector and $X_j$ the $j$-th row vector of matrix $X$, and $\bar{B}$ an $\ell_1$-row-normalized version (described later) of a non-negative matrix $B$. Also, $[n] := \{1, \ldots, n\}$.

**TABLE I**

| Condition for novel word detection | Condition for topic estimation | Necessity proved? | Statistical & Computational guarantees | Algorithm based on | Random Projections |
|------------------------------------|--------------------------------|-------------------|---------------------------------------|-------------------|-------------------|
| [9] Simplicial                      | Full-rank                      | No                | Both Polynomial                       | Linear Programming| Not used          |
| [7] Full-rank                      | Full-rank                      | No                | Both Polynomial                       | Gram-Schmidt Orthogonalization | For dimensionality reduction |
| [4] Simplicial                      | Full-rank                      | Yes, for novel word detection | Both Polynomial                       | Random Projections | Integral to novel word detection |
| This work                          | Affine-independence            | Yes, for topic estimation | Both Polynomial                       | Random Projections | Integral to novel word detection |

![Fig. 1. An example of a separable topic matrix $\beta$ (left) and the underlying geometric structure (right) of the row space of the normalized document distribution matrix $A$. Note: the word ordering is only for visualization and has no bearing on separability. Solid circles represent rows of $A$. Empty circles represent rows of $\bar{X}$ when $N$ is finite (in the ideal case, $A = \bar{X}$). Projections of $A_{\omega}$’s (resp. $X_{\omega}$’s) along a random isotropic direction $d$ can be used to identify novel words.]

A. **Key Structural Property: Topic Separability**

We first introduce separability as a key structural property of a topic matrix $\beta$. Formally,

**Definition 1:** (Separability) A topic matrix $\beta \in \mathbb{R}^{W \times K}$ is separable if for each topic $k$, there is some word $i$ such that $\beta_{i,k} > 0$ and $\beta_{i,l} = 0, \forall l \neq k$.

Topic separability implies that each topic contains word(s) which appear only in that topic. We refer to these words as the **novel words** of the $K$ topics. Figure 1 shows an example of a separable $\beta$ with $K = 3$ topics. Words 1 and 2 are novel to topic 1, words 3 and 4 to topic 2, and word 5 to topic 3. Other words that appear in multiple topics are called non-novel words (e.g., word 6). Identifying the novel words for $K$ distinct topics is the key step of our proposed approach.

We note that separability has been empirically observed to be approximately satisfied by topic estimates produced by variational and MCMC based algorithms [5], [7], [27]. More fundamentally, in very recent work [16], it has been shown that topic separability is an inevitable consequence of having a relatively small number of topics in a very large vocabulary (high-dimensionality). In particular, when the $K$ columns (topics) of $\beta$ are independently sampled from a Dirichlet distribution (on a $(W - 1)$-dimensional probability simplex), the resulting topic matrix $\beta$ will be (approximately) separable with probability tending to 1 as $W$ scales to infinity sufficiently faster than $K$. A Dirichlet prior on $\beta$ is widely-used in smoothed settings of topic modeling [1].
As we will discuss next in Sec. III-C, the topic separability property combined with additional conditions on the second-order statistics of the mixing weights leads to an intuitively appealing geometric property that can be exploited to develop a provably consistent and efficient topic estimation algorithm.

B. Conditions on the Topic Mixing Weights

Topic separability alone does not guarantee that there will be a unique \( \beta \) that is consistent with all the observations \( X \). This is illustrated in Fig. 2, [4]. Therefore, in an effort to develop provably consistent topic estimation algorithms, a number of different conditions have been imposed on the topic mixing weights \( \theta \) in the literature [3], [5], [7], [9], [15]. Complementing the work in [4] which identifies necessary and sufficient conditions for consistent detection of novel words, in this paper we identify necessary and sufficient conditions for consistent estimation of a separable topic matrix. Our necessity results are information-theoretic and algorithm-independent in nature, meaning that they are independent of any specific statistics of the observations and the algorithms used. The novel words and the topics can only be identified up to a permutation and is accounted for in our results.

Let \( a := E(\theta^m) \) and \( R := E(\theta^m\theta^m^\top) \) be the \( K \times 1 \) expectation vector and the \( K \times K \) correlation matrix of the weight prior \( P(\alpha) \). Without loss of generality, we can assume that the elements of \( a \) are strictly positive since otherwise some topic(s) will not appear in the corpus. A key quantity is \( R := \text{diag}(a)^{-1}R\text{diag}(a)^{-1} \) which may be viewed as a "normalized" second-moment matrix of the weight vector. The following conditions are central to our results.

**Condition 1: (Simplicial Condition)** A matrix \( B \) is (row-wise) \( \gamma_a \)-simplicial if any row-vector of \( B \) is at a Euclidean distance of at least \( \gamma_a > 0 \) from the convex hull of the remaining row-vectors. A topic model is \( \gamma_a \)-simplicial if its normalized second-moment \( R \) is \( \gamma_a \)-simplicial.

**Condition 2: (Affine-Independence)** A matrix \( B \) is (row-wise) \( \gamma_a \)-affine-independent if \( \min_\Lambda \| \sum_{k=1}^{K} \lambda_k B_k \|_2 / \| \lambda \|_2 \geq \gamma_a > 0 \), where \( B_k \) is the \( k \)-th row of \( B \) and the minimum is over all \( \Lambda \in \mathbb{R}^K \) such that \( \Lambda \neq 0 \) and \( \sum_{k=1}^{K} \lambda_k = 0 \). A topic model is \( \gamma_a \)-affine-independent if its normalized second-moment \( R \) is \( \gamma_a \)-affine-independent.

Here, \( \gamma_a \) and \( \gamma_a \) are called the simplicial and affine-independence constants respectively. They are condition numbers which measure the degree to which the conditions that they are respectively associated with hold. The larger that these condition numbers are, the easier it is to estimate the topic matrix. Going forward, we will say that a matrix is simplicial (resp. affine independent) if it is \( \gamma_a \)-simplicial (resp. \( \gamma_a \)-affine-independent) for some \( \gamma_a > 0 \) (resp. \( \gamma_a > 0 \)). The simplicial condition was first proposed in [9] and then further investigated in [4]. This paper is the first to identify affine-independence as both necessary and sufficient for consistent separable topic estimation. Before we discuss their geometric implications, we point out that affine-independence is stronger than the simplicial condition:

**Proposition 1:** \( R \) is \( \gamma_a \)-affine-independent \( \Rightarrow \) \( R \) is at least \( \gamma_a \)-simplicial. The reverse implication is false in general.

**The Simplicial Condition is both Necessary and Sufficient for Novel Word Detection:** We first focus on detecting all the novel words of the \( K \) distinct topics. For this task, the simplicial condition is an algorithm-independent, information-theoretic necessary condition. Formally,

**Lemma 1:** (Simplicial Condition is Necessary for Novel Word Detection) Let \( B \) be separable and \( W > K \). If there exists an algorithm that can consistently identify all novel words of all \( K \) topics from \( X \), then \( R \) is simplicial.

The key insight behind this result is that when \( R \) is non-simplicial, we can construct two distinct separable topic matrices with different sets of novel words which induce the same distribution on the empirical observations \( X \). Geometrically, the simplicial condition guarantees that the \( K \) rows of \( R \) will be extreme points of the convex hull that they themselves form. Therefore, if \( R \) is not simplicial, there will exist at least one redundant topic which is just a convex combination of the other topics.

It turns out that \( R \) being simplicial is also sufficient for consistent novel word detection. This is a direct consequence of the consistency guarantees of our approach as outlined in Theorem 3.

**Affine-Independence is Necessary and Sufficient for Separable Topic Estimation:** We now focus on estimating a separable topic matrix \( \beta \), which is a stronger requirement than detecting novel words. It naturally requires conditions that are stronger than the simplicial condition. Affine-independence turns out to be an algorithm-independent, information-theoretic necessary condition. Formally,

**Lemma 2:** (Affine-Independence is Necessary for Separable Topic Estimation) Let \( B \) be separable with \( W \geq 2 + K \). If there exists an algorithm that can consistently estimate \( \beta \) from \( X \), then its normalized second-moment \( R \) is affine-independent.

Similar to Lemma 1, if \( R \) is not affine-independent, we can construct two distinct and separable topic matrices that induce the same distribution on the observation which makes consistent topic estimation impossible. Geometrically, every point in a convex set can be decomposed uniquely as a
convex combination of its extreme points, if, and only if, the extreme points are affine-independent. Hence, if $\mathbf{R}$ is not affine-independent, a non-novel word can be assigned to different subsets of topics.

The sufficiency of the affine-independence condition in separable topic estimation is again a direct consequence of the consistency guarantees of our approach as in Theorems 3 and 4. We note that since affine-independence implies the simplicial condition (Proposition 1), affine-independence is sufficient for novel word detection as well.

**Connection to Other Conditions on the Mixing Weights:**
We briefly discuss other conditions on the mixing weights $\theta$ that have been exploited in the literature. In [7], [15], $\mathbf{R}$ (equivalently $\bar{\mathbf{R}}$) is assumed to have full-rank (with minimum eigenvalue $\gamma_r > 0$). In [3], $\mathbf{R}$ is assumed to be diagonal dominant, i.e., $\forall i, j, i \neq j, \bar{R}_{i,i} - \bar{R}_{i,j} \geq \gamma_d > 0$. They are both sufficient conditions for detecting all the novel words of all distinct topics. The constants $\gamma_r$ and $\gamma_d$ are condition numbers which measure the degree to which the full-rank and diagonal dominance conditions hold respectively. They are counterparts of $\gamma_s$ and $\gamma_a$ and like them, the larger they are, the easier it is to consistently detect the novel words and estimate $\beta$. The relationships between these conditions are summarized in Proposition 2 and illustrated in Fig. 3.

**Proposition 2:** Let $\mathbf{R}$ be the normalized second-moment of the topic prior. Then,

1. $\mathbf{R}$ is full rank with minimum eigenvalue $\gamma_r \Rightarrow \mathbf{R}$ is at least $\gamma_r$-affine-independent $\Rightarrow \mathbf{R}$ is at least $\gamma_r$-simplicial.
2. $\mathbf{R}$ is $\gamma_d$-diagonal dominant $\Rightarrow \mathbf{R}$ is at least $\gamma_d$-simplicial.
3. $\mathbf{R}$ being diagonal dominant neither implies nor is implied by $\mathbf{R}$ being affine-independent (or full-rank).

We note that in our earlier work [5], the provable guarantees for estimating the separable topic matrix require $\mathbf{R}$ to have full rank. The analysis in this paper provably extends the guarantees to the affine-independence condition.

**C. Geometric Implications and Random Projections Based Algorithm**

We now demonstrate the geometric implications of topic separability combined with the simplicial/affine-independence condition on the topic mixing weights. To highlight the key ideas we focus on the ideal case where $N = \infty$. Then, the empirical document word-frequency matrix $\mathbf{X} = \mathbf{A} = \beta \theta$.

**Novel Words are Extreme Points:** To expose the underlying geometry, we $\ell_1$-normalize the rows of $\mathbf{A}$ and $\theta$ to obtain row-stochastic matrices $\bar{\mathbf{A}} := \text{diag}(\mathbf{A}1)^{-1} \mathbf{A}$ and $\bar{\boldsymbol{\theta}} := \text{diag}(\bar{\boldsymbol{\theta}}1)^{-1} \bar{\boldsymbol{\theta}}$. Then since $\mathbf{A} = \beta \bar{\mathbf{A}}$, we have $\bar{\mathbf{A}} = \beta \bar{\boldsymbol{\theta}}$ where $\beta := \text{diag}(\mathbf{A}1)^{-1} \beta \text{diag}(\theta1)$ is a row-normalized “topic matrix” which is both row-stochastic and separable with the same sets of novel words as $\beta$.

Now consider the row vectors of $\bar{\mathbf{A}}$ and $\bar{\boldsymbol{\theta}}$. First, it can be shown that if $\mathbf{R}$ is simplicial (cf. Condition 1) then, with probability $\to 1$ as $M \to \infty$, no row of $\bar{\boldsymbol{\theta}}$ will be in the convex hull of the others [4]. Next, the separability property ensures that if $w$ is a novel word of topic $k$, then $\bar{\beta}_{wk} = 1$ and $\bar{\beta}_{w,j} = 0 \forall j \neq k$ so that $\bar{\mathbf{A}}_w = \bar{\boldsymbol{\theta}}_k$. Revisiting the example in Fig. 1, the rows of $\bar{\mathbf{A}}$ which correspond to novel words, e.g., words 1 through 5, are all row-vectors of $\bar{\boldsymbol{\theta}}$ and together form a convex hull of $K$ extreme points. For example, $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$ and $\bar{\mathbf{A}}_3 = \bar{\mathbf{A}}_4 = \bar{\boldsymbol{\theta}}_2$. If, however, $w$ is a non-novel word, then $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$ lives inside the convex hull of the rows of $\bar{\boldsymbol{\theta}}$. In Fig. 1, row $\bar{\mathbf{A}}_6$ which corresponds to non-novel word 6, is inside the convex hull of $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2, \bar{\boldsymbol{\theta}}_3$. In summary, the novel words can be detected as extreme points of all the row-vectors of $\bar{\mathbf{A}}$. Also, multiple novel words of the same topic correspond to the same extreme point (e.g., $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$). Formally,

**Lemma 3:** Let $\bar{\mathbf{R}}$ be $\gamma_s$ simplicial and $\beta$ be separable. Then, with probability at least $1 - 2K \exp(-c_1 M) - \exp(-c_2 M)$, the $i$-th row of $\bar{\mathbf{A}}$ is an extreme point of the convex hull spanned by all the rows of $\bar{\mathbf{A}}$ if, and only if, word $i$ is novel. Here, $c_1 := \gamma_a^2 a_{\text{max}}^4/\lambda_{\text{max}}$, $c_2 := \gamma_a^2 a_{\text{min}}^4/\lambda_{\text{max}}^2$, $a_{\text{min}}$ is the minimum element of $\mathbf{a}$, and $\lambda_{\text{max}}$ is the maximum singular-value of $\mathbf{R}$.

To see how identifying novel words can help us estimate $\beta$, recall that the row-vectors of $\bar{\mathbf{A}}$ corresponding to novel words coincide with the rows of $\bar{\boldsymbol{\theta}}$. Thus $\bar{\boldsymbol{\theta}}$ is known once one novel word for each topic is known. Also, for all words $w$, $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$. Thus, if we can uniquely decompose $\bar{\mathbf{A}}_w$ as a convex combination of the extreme points, then the coefficients of the decomposition will give us the $w$-th row of $\bar{\boldsymbol{\theta}}$. A unique decomposition exists with probability $\to 1$ as $M \to \infty$, when $\mathbf{R}$ is affine-independent and can be found by solving a constrained linear regression problem. This gives us $\bar{\boldsymbol{\theta}}$. Finally, noting that $\text{diag}(\mathbf{A}1) \beta = \beta \text{diag}(\theta1)$, $\beta$ can be recovered by suitably renormalizing rows and then columns of $\bar{\boldsymbol{\theta}}$.

**Lemma 4:** Let $\mathbf{A}$ and one novel word per distinct topic be given. If $\mathbf{R}$ is $\gamma_a$ affine-independent, then, with probability at least $1 - 2K \exp(-c_1 M) - \exp(-c_2 M)$, $\beta$ can be recovered uniquely via constrained linear regression. Here, $c_1 := \gamma_a^2 a_{\text{max}}^4/\lambda_{\text{max}}$, $c_2 := \gamma_a^2 a_{\text{min}}^4/2\lambda_{\text{max}}^2$, $a_{\text{min}}$ is the minimum element of $\mathbf{a}$, and $\lambda_{\text{max}}$ is the maximum singular-value of $\mathbf{R}$.

Lemmas 3 and 4 together provide a geometric approach for learning $\beta$ from $\mathbf{A}$ (equivalently $\bar{\mathbf{A}}$): (1) Find extreme points of rows of $\bar{\mathbf{A}}$. Cluster the rows of $\bar{\mathbf{A}}$ that correspond to the same extreme point into the same group. (2) Express the remaining rows of $\bar{\mathbf{A}}$ as convex combinations of the $K$ distinct extreme points. (3) Renormalize $\beta$ to obtain $\bar{\boldsymbol{\theta}}$.

**Detecting Extreme Points using Random Projections:** A key contribution of our approach is an efficient random projections based algorithm to detect novel words as extreme points. The idea is illustrated in Fig. 1: if we project every point of a convex body onto an isotropically distributed random direction.
d, the maximum (or minimum) projection value must correspond to one of the extreme points with probability 1. On the other hand, the non-novel words will not have the maximum projection value along any random direction. Therefore, by repeatedly projecting all the points onto a few isotropically distributed random directions, we can detect all the extreme points with very high probability as the number of random directions increases. An explicit bound on the number of projections needed appears in Theorem 3.

Finite N in Practice: The geometric intuition discussed above was based on the row-vectors of A. When N = ∞, A = X the matrix of ℓ1-row-normalized empirical word-frequencies of all documents. If N is finite but very large, A can be well-approximated by X thanks to the law of large numbers. However, in real-word text corpora, N ≪ W (e.g., N = 298 while W = 14,943 in the NYT dataset). Therefore, the row-vectors of X are significantly perturbed away from the ideal rows of A as illustrated in Fig. 1. We discuss the effect of small N and how we address the accompanying issues next.

IV. Topic Geometry With Finite Samples: Word Co-occurrence Matrix Representation, Solid Angle, and Random Projections Based Approach

The extreme point geometry sketched in Sec. III-C is perturbed when N is small as highlighted in Fig. 1. Specifically, the rows of the empirical word-frequency matrix X deviate from the rows of A. This creates several problems: (1) points in the convex hull corresponding to non-novel words may also become “outlier” extreme points (e.g., X_6 in Fig. 1); (2) some extreme points that correspond to novel words may no longer be extreme (e.g., X_3 in Fig. 1); (3) multiple novel words corresponding to the same extreme point may become multiple distinct extreme points (e.g., X_1 and X_2 in Fig. 1). Unfortunately, these issues do not vanish as M increases with N fixed—a regime which captures the characteristics of typical benchmark datasets—because the dimensionality of the rows (equal to M) also increases. There is no “averaging” effect to smoothen-out the sampling noise.

Our solution is to seek a new representation, a statistic of X, which can not only smoothen out the sampling noise of individual documents, but also preserve the same extreme point geometry induced by the separability and affine independence conditions. In addition, we also develop an extreme point robustness measure that naturally arises within our random projections based framework. This robustness measure can be used to detect and exclude the “outlier” extreme points.

A. Normalized Word Co-occurrence Matrix Representation

We construct a suitably normalized word co-occurrence matrix from X as our new representation. The co-occurrence matrix converges almost surely to an ideal statistic as M → ∞ for any fixed N ≥ 2. Simultaneously, in the asymptotic limit, the original novel words continue to correspond to extreme points in the new representation and overall extreme point geometry is preserved.

The new representation is (conceptually) constructed as follows. First randomly divide all the words in each document into two equal-sized independent halves and obtain two W × K empirical word-frequency matrices X and X’ each containing N/2 words. Then ℓ_1-normalize their rows like in Sec. III-C to obtain X and X’ which are row-stochastic. The empirical word co-occurrence matrix of size W × W is then given by

\[ \hat{E} := M \hat{X}' \hat{X}^\top \]  

(1)

We note that in our random projection based approach, \( \hat{E} \) is not explicitly constructed by multiplying \( \hat{X}' \) and X. Instead, we keep \( \hat{X}' \) and X and exploit their sparsity properties to reduce the computational complexity of all subsequent processing.

Asymptotic Consistency: The first nice property of the word co-occurrence representation is its asymptotic consistency when N is fixed. As the number of documents M → ∞, the empirical \( \hat{E} \) converges, almost surely, to an ideal word co-occurrence matrix E of size W × W. Formally,

Lemma 5: ([3]; also see Lemma 3.7 in [9]) Let \( \hat{E} \) be the empirical word co-occurrence matrix defined in Eq. (1). Then,

\[ \hat{E} \xrightarrow{M \to \infty} \text{almost surely} \beta \hat{R} \hat{R}^\top =: E \]  

(2)

where \( \hat{\beta} := \text{diag}^{-1}(\beta a)\text{diag}(a) \) and \( \hat{R} := \text{diag}^{-1}(a)\text{diag}^{-1}(a) \). Furthermore, if \( \eta := \min_{1 \leq i \leq W} (\beta a)_i > 0 \), then \( \Pr(||\hat{E} - E||_{\infty} \geq \epsilon) \leq 8W^2 \exp(-\epsilon^2/4MN/20) \).

Here \( \hat{E} \) is the same normalized second-moment of the topic priors as defined in Sec. III and \( \beta \) is an ℓ_1-row-normalized version of \( \beta \). We make note of the abuse of notion for \( \beta \) which was defined in Sec. III-C. It can be shown that the \( \beta \) defined in Lemma 5 is the limit of the one defined in Sec. III-C as M → ∞. The convergence result in Lemma 5 shows that the word co-occurrence representation \( \hat{E} \) can be consistently estimated by \( E \) as M → ∞ and the deviation vanishes exponentially in M which is large in typical benchmark datasets.

Novel Words are Extreme Points: Another reason for using this word co-occurrence representation is that it preserves the extreme point geometry. Consider the ideal word co-occurrence matrix \( E = \beta (R \beta^\top) \). It is straightforward to show that if \( \beta \) is separable and \( R \) is simplicial then \( (R \beta^\top) \) is also simplicial. Using these facts it is possible to establish the following counterpart of Lemma 3 for \( E \):

Lemma 6: (Novel Words are Extreme Points [5, Lemma 1], [9]) Let \( R \) be simplicial and \( \beta \) be separable. Then, a word \( i \) is novel if, and only if, the \( i \)-th row of \( E \) is an extreme point of the convex hull spanned by all the rows of \( E \).

In another words, the novel words correspond to the extreme points of all the row-vectors of the ideal word co-occurrence matrix \( E \). Consider the example in Fig. 4 which is based on the same topic matrix \( \beta \) as in Fig. 1. Here, \( E_1 = E_2, E_3 = E_4, \) and \( E_5 \) are K = 3 distinct extreme points of all row-vectors of \( E \) and \( E_6 \) which corresponds to a non-novel word, is inside the convex hull.

Once the novel words are detected as extreme points, we can follow the same procedure as in Lemma 4 and express each row \( E_{iw} \) of \( E \) as a unique convex combination of the \( K \) extreme rows of \( E \) or equivalently the rows of \( (R \beta^\top) \). The weights of the
convex combination are the $\hat{\beta}_{wk}$'s. We can then apply the same row and column renormalization to obtain $\beta$. The following result is the counterpart of Lemma 4 for $E$:

**Lemma 7:** Let $E$ and one novel word for each distinct topic be given. If $\bar{R}$ is affine-independent, then $\beta$ can be recovered uniquely via constrained linear regression.

One can follow the same steps as in the proof of Lemma 4. The only additional step is to check that $\bar{R}\beta^T = [\bar{R}, \bar{R}B]$ is affine-independent if $\bar{R}$ is affine-independent.

We note that the finite sampling noise perturbation $\bar{E} - E$ is still not 0 but vanishes as $M \to \infty$ (in contrast to the $X$ representation in Sec. III-C). However, there is still a possibility of observing “outlier” extreme points if a non-novel word lies on the facet of the convex hull of the rows of $E$. We next introduce an extreme point robustness measure based on a certain solid angle that naturally arises in our random projections based approach, and discuss how it can be used to detect and distinguish between “true” novel words and such “outlier” extreme points.

### B. Solid Angle Extreme Point Robustness Measure

To handle the impact of a small but nonzero perturbation $\|\bar{E} - E\|_\infty$, we develop an extreme point “robustness” measure. This is necessary for not only applying our approach to real-world data but also to establish finite sample complexity bounds. Intuitively, a robustness measure should be able to distinguish between the “true” extreme points (row vectors that are novel words) and the “outlier” extreme points (row vectors of non-novel words that become extreme points due to the nonzero perturbation). Towards this goal, we leverage a key geometric quantity, namely, the Normalized Solid Angle subtended by the convex hull of the rows of $E$ at an extreme point. To visualize this quantity, we revisit our running example in Fig. 4 and indicate the solid angles attached to each extreme point by the shaded regions. It turns out that this geometric quantity naturally arises in the context of random projections that was discussed earlier. To see this connection, in Fig. 4 observe that the shaded region attached to any extreme point coincides precisely with the set of directions along which its projection is larger (taking sign into account) than that of any other point (whether extreme or not). For example, in Fig. 4 the projection of $E_1 = E_2$ along $d_1$ is larger than that of any other point. Thus, the solid angle attached to a point $E_i$ (whether extreme or not) can be formally defined as the set of directions $\{d : \forall j : E_j \neq E_i, \langle E_i, d \rangle > \langle E_j, d \rangle\}$. This set is nonempty only for extreme points. The solid angle defined above is a set. To derive a scalar robustness measure from this set and tie it to the idea of random projections, we adopt a statistical perspective and define the normalized solid angle of a point as the probability that the point will have the maximum projection value along an isotropically distributed random direction. Concretely, for the $i$-th word (row vector), the normalized solid angle $q_i$ is defined as

$$q_i := \Pr(\forall j : E_j \neq E_i, \langle E_i, d \rangle > \langle E_j, d \rangle)$$  (3)

where $d$ is drawn from an isotropic distribution in $\mathbb{R}^W$ such as the spherical Gaussian. The condition $E_i \neq E_j$ in Eq. (3) is introduced to exclude the multiple novel words of the same topic that correspond to the same extreme point. For instance, in Fig. 4 $E_1 = E_2$, Hence, for $q_1, j = 2$ is excluded. To make it practical to handle finite sample estimation noise we replace the condition $E_i \neq E_j$ by the condition $\|E_i - E_j\| \geq \zeta$ for some suitably defined $\zeta$.

As illustrated in Fig. 4, the solid angle for all the extreme points are strictly positive given $\bar{R}$ is $\gamma_s$-simplicial. On the other hand, for $i$ that is non-novel, the corresponding solid angle $q_i$ is zero by definition. Hence the extreme point geometry in Lemma 6 can be re-expressed in term of solid angles as follows:

**Lemma 8:** (Novel Words Have Positive Solid Angles) Let $\bar{R}$ be simplicial and $\beta$ be separable. Then, word $i$ is a novel word if, and only if, $q_i > 0$.

We denote the smallest solid angle among the $K$ distinct extreme points by $q_{\alpha} > 0$. This is a robust condition number of the convex hull formed by the rows of $E$ and is related to the simplicial constant $\gamma_s$ of $\bar{R}$.

In a real-world dataset we have access to only an empirical estimate $\bar{E}$ of the ideal word co-occurrence matrix $E$. If we replace $E$ with $\bar{E}$, then the resulting empirical solid angle estimate $\hat{q}_i$ will be very close to the ideal $q_i$ if $\bar{E}$ is close enough to $E$. Then, the solid angles of “outlier” extreme points will be close to 0 while they will be bounded away from zero for the “true” extreme points. One can then hope to correctly identify all $K$ extreme points by rank-ordering all empirical solid angle estimates and selecting the $K$ distinct row-vectors that have the largest solid angles. This forms the basis of our proposed algorithm. The problem now boils down to efficiently estimating the solid angles and establishing the asymptotic convergence of the estimates as $M \to \infty$. We next discuss how random projections can be used to achieve these goals.

### C. Efficient Solid Angle Estimation via Random Projections

The definition of the normalized solid angle in Eq. (3) motivates an efficient algorithm based on random projections to
estimate it. For convenience, we first rewrite Eq. (3) as
\[ q_i = \mathbb{E} \left[ \sum_{j} \mathbbm{1}\{ \| \mathbf{E}_j - \mathbf{E}_i \| \geq \zeta, \mathbf{E}_i \mathbf{d} \geq \mathbf{E}_j \mathbf{d} \} \right] \] (4)
and then propose to estimate it by
\[ \hat{q}_i = \frac{1}{P} \sum_{r=1}^{P} \mathbbm{1}\{ \sum_{j} \mathbb{E} \left[ \tilde{E}_{i,j} + \tilde{E}_{j,i} - 2\tilde{E}_{i,j} \geq \zeta/2, \tilde{E}_i \mathbf{d}^r > \tilde{E}_j \mathbf{d}^r \right] \} \] (5)
where \( \mathbf{d}^1, \ldots, \mathbf{d}^P \in \mathbb{R}^{W \times 1} \) are \( P \) iid directions drawn from an isotropic distribution in \( \mathbb{R}^W \). Algorithmically, by Eq. (5), we approximate the solid angle \( q_i \) at the \( i \)-th word (row-vector) by first projecting all the row-vectors onto \( P \) iid isotropic random directions and then calculating the fraction of times each row-vector achieves the maximum projection value. It turns out that the condition \( \mathbb{E} \left[ \tilde{E}_{i,j} + \tilde{E}_{j,i} - 2\tilde{E}_{i,j} \geq \zeta/2 \right] \) is equivalent to \( \| \mathbf{E}_i - \mathbf{E}_j \| \geq \zeta \) in terms of its ability to exclude multiple novel words from the same topic and is adopted for its simplicity.\(^2\)

This procedure of taking random projections followed by calculating the number of times a word is a maximizer via Eq. (5) provides a consistent estimate of the solid angle in Eq. (3) as \( M \to \infty \) and the number of projections \( P \) increases. The high-level idea is simple: as \( P \) increases, the empirical average in Eq. (5) converges to the corresponding expectation. Simultaneously, as \( M \) increases, \( \zeta \to 0 \) and \( \mathbb{E} \). Overall, the approximation \( \hat{q}_i \) proposed in Eq. (5) using random projections converges to \( q_i \).

This random projections based approach is also computationally efficient for the following reasons. First, it enables us to avoid the explicit construction of the \( W \times W \) dimensional matrix \( \mathbb{E} \); Recall that each column of \( \mathbf{X} \) and \( \mathbf{X}' \) has no more than \( N \ll W \) non-zero entries. Hence \( \mathbf{X} \) and \( \mathbf{X}' \) are both sparse. Since \( \tilde{\mathbf{E}} \mathbf{d} = \mathbb{E} \mathbf{X}' (\mathbf{X} \mathbf{d}) \), the projection can be calculated using two sparse matrix-vector multiplications. Second, it turns out that the number of projections \( P \) needed to guarantee consistency is small. In fact in Theorem 3 we provide a sufficient upper bound for \( P \) which is a polynomial function of \( \log(W) \), \( \log(1/\delta) \) and other model parameters, where \( \delta \) is the probability that the algorithm fails to detect all the distinct novel words.

**Parallelization, Distributed and Online Settings:** Another advantage of the proposed random projections based approach is that it can be parallelized and is naturally amenable to online or distributed settings. This is based on the following observation that each projection has an additive structure:
\[ \tilde{\mathbf{E}} \mathbf{d}^r = \mathbb{E} \mathbf{X}' \mathbf{X}^{\top} \mathbf{d}^r = \mathbb{E} \sum_{m}^{M} \mathbf{X}^{m'} \mathbf{X}^{m\top} \mathbf{d}^r. \]
The \( P \) projections can also be computed independently. Therefore,
- In a distributed setting in which the documents are stored on distributed servers, we can first share the same random directions across servers and then aggregate the projection values. The communication cost is only the “partial” projection values and is therefore insignificant [5] and does not scale as the number of observations \( N, M \) increases.
- In an online setting in which the documents are streamed in an online fashion [21], we only need to keep all the projection values and update the projection values (hence the empirical solid angle estimates) when new documents arrive.

The additive and independent structure guarantees that the statistical efficiency of these variations are the same as the centralized “batch” implementation. For the rest of this paper, we only focus on the centralized version.

**Outline of Overall Approach:** Our overall approach can be summarized as follows. (1) Estimate the empirical solid angles using \( P \) iid isotropic random directions as in Eq. (5). (2) Select the \( K \) words with distinct word co-occurrence patterns (rows) that have the largest empirical solid angles. (3) Estimate the topic matrix using constrained linear regression as in Lemma 4. We will discuss the details of our overall approach in the next section and establish guarantees for its computational and statistical efficiency.

V. ALGORITHM AND ANALYSIS

Algorithm 1 describes the main steps of our overall random projections based algorithm which we call RP. The two main steps, novel word detection and topic matrix estimation are outlined in Algorithms 2 and 3 respectively. Algorithm 2 outlines the random projection and rank-ordering steps. Algorithm 3 describes the constrained linear regression and the renormalization steps in a combined way.

Algorithm 1. RP

**Input:** Text document \( \mathbf{X}, \mathbf{X}' (W \times M) \); Number of topics \( K \); Number of iid random projections \( P \); Tolerance parameters \( \zeta, \epsilon > 0 \).

**Output:** Estimate of the topic matrix \( \hat{\beta}(W \times K) \).
1: Set of Novel Words \( I \leftarrow \) NovelWordDetect(\( \mathbf{X}, \mathbf{X}', K, P, \zeta \))
2: \( \hat{\beta} \leftarrow \) EstimateTopics(\( I, \mathbf{X}, \mathbf{X}', \epsilon \))

**Computational Efficiency:** We first summarize the computational efficiency of Algorithm 1:

**Theorem 2:** Let the number of novel words for each topic be a constant relative to \( M, W, N \). Then, the running time of Algorithm 1 is \( \mathcal{O}(MNP + WP + WK^3) \).

This efficiency is achieved by exploiting the sparsity of \( \mathbf{X} \) and the property that there are only a small number of novel words in a typical vocabulary. A detailed analysis of the computational complexity is presented in [17]. Here we point out that in order to upper bound the computation time of the linear regression in Algorithm 3 we used \( \mathcal{O}(WK^3) \) for \( W \) matrix inversions, one for each of the words in the vocabulary. In practice, a gradient descent implementation can be used for the constrained linear regression which is much more efficient. We also note that these \( W \) optimization problems are decoupled given the set of detected novel words. Therefore, they can be parallelized in a straightforward manner [5].
Algorithm 2. NovelWordDetect (via Random Projections)

Input: $X, X'$; Number of topics $K$; Number of projections $P$; Tolerance $\zeta$;
Output: The set of all novel words of $K$ distinct topics $I$.
1: $q_i \leftarrow 0, \forall i = 1, \ldots, W, E \leftarrow MXX^\top$.
2: for all $r = 1, \ldots, P$ do
3: Sample $d^r \in \mathbb{R}^W$ from an isotropic prior.
4: $v \leftarrow MXX^\top d^r$.
5: $i^* \leftarrow \arg \max_{1 \leq i \leq W} v_i, \hat{q}_i \leftarrow \hat{q}_i + 1/P$.
6: $J_{i^*} \leftarrow \{ j : \hat{E}_{i^*,j} + \hat{E}_{j,j} - 2\hat{E}_{i^*,j} \geq \zeta/2 \}$.
7: for all $k \in J_{i^*}$ do
8: $\hat{J}_k \leftarrow \{ j : \hat{E}_{k,j} + \hat{E}_{j,j} - 2\hat{E}_{k,j} \geq \zeta/2 \}$.
9: if $\forall j \in \hat{J}_k, v_k > v_j$ then
10: $\hat{q}_k \leftarrow \hat{q}_k + 1/P$.
11: end if
12: end for
13: end for
14: $I \leftarrow 0, k \leftarrow 0, j \leftarrow 1$.
15: while $k < K$ do
16: $i \leftarrow$ index of the $j^{th}$ largest value of $\{ \hat{q}_1, \ldots, \hat{q}_W \}$.
17: if $\forall p \in I : \hat{E}_{p,i} + \hat{E}_{i,p} - 2\hat{E}_{i,i} \geq \zeta/2$ then
18: $I \leftarrow I \cup \{i\}, k \leftarrow k + 1$.
19: end if
20: $j \leftarrow j + 1$.
21: end while
22: Return $I$.

Algorithm 3. EstimateTopics

Input: $I = \{i_1, \ldots, i_K\}$ set of novel words, one for each of the $K$ topics; $E$; precision parameter $\epsilon$.
Output: $\beta$, which is the estimate of the $\beta$ matrix.
1: $E^t_w = [\hat{E}^t_{w,i_1}, \ldots, \hat{E}^t_{w,i_K}]$.
2: $Y = (E^t_{11}, \ldots, E^t_{K1})^\top$.
3: for all $i = 1, \ldots, W$ do
4: Solve $b^* := \arg \min_j \| E^*_i - bY \|^2$
5: subject to $b_j \geq 0, \sum_{j=1}^K b_j = 1$.
6: using precision $\epsilon$ for the stopping-criterion.
7: $\hat{\beta}_i \leftarrow \frac{1}{M} X_i b^*$.
8: end for
9: $\hat{\beta} \leftarrow$ column normalize $\hat{\beta}$.

Asymptotic Consistency and Statistical Efficiency: We now summarize the asymptotic consistency and sample complexity bounds for Algorithm 1. The analysis is a combination of the consistency of the novel word detection step (Algorithm 2) and the topic estimation step (Algorithm 3). We state the results for both of these steps. First, for detecting all the novel words of the $K$ distinct topics, we have the following result:

**Theorem 3:** Let topic matrix $\beta$ be separable and $\hat{R}$ be $\gamma_{\alpha}$-affine-independent. If the projection directions are iid sampled from any isotropic distribution, then Algorithm 2 can identify all the novel words of the $K$ distinct topics as $M, P \to \infty$.

Furthermore, $\forall \delta \geq 0$, if

$$M \geq 20 \log (2W/\delta) Np^2 \eta^4$$

then Algorithm 2 fails with probability at most $\delta$. The model parameters are defined as follows. $\rho = \min \{ \frac{1}{2}, \frac{d_2}{d_2+1} \}$ where $d = (1-b)^2\gamma_{\alpha}^{2}/\lambda_{\alpha}$, $d_2 \triangleq (1-b)\gamma_{\alpha}$, $\lambda_{\alpha}$ is the maximum eigenvalue of $\hat{R}$, $b = \max \{ \epsilon \}$, and $C_0$ is the set of non-novel words. Finally, $q_{\lambda}$ is the minimum solid angle of the extreme points of the convex hull of the rows of $E$.

The detailed proof is presented in [17]. The results in Eq. (6) provide a sufficient finite sample complexity bound for novel word detection. The bound is polynomial with respect to $M, W, K, N, \log (\delta)$ and other model parameters. The number of projections $P$ that impacts the computational complexity scales as $\log(W)/q_{\lambda}^2$ in this sufficient bound where $q_{\lambda}$ can be upper bounded by $1/K$ (since $K$ nonnegative normalized solid angles for $K$ distinct topics sum up to 1, the minimum cannot exceed $1/K$). In practice, we have found that setting $P = O(K)$ is a good choice [5].

We note that the result in Theorem 3 only requires the simplicial condition which is the minimum condition required for consistent novel word detection (Lemma 1). This theorem holds true if the topic prior $\hat{R}$ satisfies stronger conditions such as affine-independence. We also point out that our proof in this paper holds for any isotropic distribution on the random projection directions $d_1, \ldots, d_P$. The previous result in [5], however, only applies to some specific isotropic distributions such as the Spherical Gaussian or the uniform distribution in a unit ball. In practice, we use Spherical Gaussian since sampling from such prior is simple and requires only $O(W)$ time for generating each random direction.

Next, given the successful detection of the set of novel words for all topics, we have the following result for the accurate estimation of the separable topic matrix $\beta$:

**Theorem 4:** Let topic matrix $\beta$ be separable and $\hat{R}$ be $\gamma_{\alpha}$-affine-independent. Given the successful detection of novel words for all $K$ distinct topics, the output of Algorithm 3 $\hat{\beta}$ $\xrightarrow{a.s.}$ $\beta$ element-wise (up to a column permutation). Specifically, if

$$M \geq 2560W^2K \log(W^4K/\delta) N\gamma_{\alpha}^2 \alpha_{\min}^2 \eta^4$$

then $\forall i, k$, $\hat{\beta}_{i,k}$ will be $\epsilon$ close to $\beta_{i,k}$ with probability at least $1 - \delta$, for any $0 < \epsilon < 1$. $\eta$ is the same as in Theorem 3. $\alpha_{\min}$ is the minimum value in $\alpha$.

We note that the sufficient sample complexity bound in Eq. (7) is again polynomial in terms of all the model parameters. Here we only require $\hat{R}$ to be affine-independent. Combining Theorem 3 and Theorem 4 gives the consistency and sample complexity bounds of our overall approach in Algorithm 1.

VI. EXPERIMENTAL RESULTS

In this section, we present experimental results on both synthetic and real world datasets. We report different performance measures that have been commonly used in the topic modeling literature. When the ground truth is available (Sec. VI-A),
we use the $\ell_1$ reconstruction error between the ground truth topics and the estimates after proper topic alignment. For the real-world text corpora in Sec. VI-B, we report the held-out probability, which is a standard measure used in the topic modeling literature. We also qualitatively (semantically) compare the topics extracted by the different approaches using the top probable words for each topic.

### A. Semi-Synthetic Text Corpus

In order to validate our proposed algorithm, we generate “semi-synthetic” text corpora by sampling from a synthetic, yet realistic, ground truth topic model. To ensure that the semi-synthetic data is similar to real-world data, in terms of dimensionality, sparsity, and other characteristics, we use the following generative procedure adapted from [5], [7].

We first train an LDA model (with $K = 100$) on a real-world dataset using a standard Gibbs Sampling method with default parameters (as described in [11], [34]) to obtain a topic matrix $\beta_0$ of size $W \times K$. The real-world dataset that we use to generate our synthetic data is derived from the NYT dataset [8]. The original vocabulary is first pruned based on document frequencies. Specifically, as is standard practice, only words that appear in more than 500 documents are retained. Thereafter, again as per standard practice, the words in the so-called stopwords list are deleted as recommended in [35]. After these steps, $M = 300,000, W = 14,943$, and the average document length $N = 298$. We then generate semi-synthetic datasets, for various values of $M$, by fixing $N = 300$ and using $\beta_0$ and a Dirichlet topic prior. As suggested in [11] and used in [5], [7], we use symmetric hyper-parameters (0.03) for the Dirichlet topic prior.

The $W \times K$ topic matrix $\beta_0$ may not be separable. To enforce separability, we create a new separable $(W + K) \times K$ dimensional topic matrix $\beta_{sep}$ by inserting $K$ synthetic novel words (one per topic) having suitable probabilities in each topic. Specifically, $\beta_{sep}$ is constructed by transforming $\beta_0$ as follows. First, for each synthetic novel word in $\beta_{sep}$, the value of the sole nonzero entry in its row is set to the probability of the most probable word in the topic (column) of $\beta_0$ for which it is a novel word. Then the resulting $(W + K) \times K$ dimensional nonnegative matrix is renormalized column-wise to make it column-stochastic. Finally, we generate semi-synthetic datasets, for various values of $M$, by fixing $N = 300$ and using $\beta_{sep}$ and the same symmetric Dirichlet topic prior used for $\beta_0$.

We use the name Semi-Syn to refer to datasets that are generated using $\beta_0$ and the name Semi-Syn +Novel for datasets generated using $\beta_{sep}$.

In our proposed random projections based algorithm, which we call RP, we set $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$. We compare RP against the provably efficient algorithm RecoverL2 in [7] and the standard Gibbs Sampling based LDA algorithm (denoted by Gibbs) in [11], [34]. In order to measure the performance of different algorithms in our experiments based on semi-synthetic data, we compute the $\ell_1$ norm of the reconstruction error between $\hat{\beta}$ and $\beta$. Since all column permutations of a given topic matrix correspond to the same topic model (for a corresponding permutation of the topic mixing weights), we use a bipartite graph matching algorithm to optimally match the columns of $\hat{\beta}$ with those of $\beta$ (based on minimizing the sum of $\ell_1$ distances between all pairs of matching columns) before computing the $\ell_1$ norm of the reconstruction error between $\hat{\beta}$ and $\beta$.

The results on both Semi-Syn +Novel NYT and Semi-Syn NYT are summarized in Fig. 5 for all three algorithms for various choices of the number of documents $M$. We note that in these figures the $\ell_1$ norm of the error has been normalized by the number of topics ($K = 100$).

As Fig. 5 shows, when the separability condition is strictly satisfied (Semi-Syn +Novel), the reconstruction error of RP converges to 0 as $M$ becomes large and outperforms the approximation-based Gibbs. When the separability condition is not strictly satisfied (Semi-Syn), the reconstruction error of RP is comparable to Gibbs (a practical benchmark).

#### Solid Angle and Model Selection

In our proposed algorithm RP, the number of topics $K$ (the model-order) needs to be specified. When $K$ is unavailable, it needs to be estimated from the data. Although not the focus of this work, Algorithm 2, which identifies novel words by sorting and clustering the estimated solid angles of words, can be suitably modified to estimate $K$. Indeed, in the ideal scenario where there is no sampling noise ($M = \infty, \mathbf{E} = \mathbf{E}$, and $\forall i, \tilde{q}_i = q_i$), only novel words have positive solid angles ($\tilde{q}_i$’s) and the rows of $\mathbf{E}$ corresponding to the novel words of the same topic are identical, i.e., the distance between the rows is zero or, equivalently, they are within a neighborhood of size zero of each other. Thus, the number of distinct neighborhoods of size zero among the non-zero solid angle words equals $K$.

In the nonideal case $M$ is finite. If $M$ is sufficiently large, one can expect that the estimated solid angles of non-novel words will not all be zero. They are, however, likely to be much smaller than those of novel words. Thus to reliably estimate $K$...
one should not only exclude words with exactly zero solid angle estimates, but also those above some nonzero threshold. When $M$ is finite, the rows of $\hat{E}$ corresponding to the novel words of the same topic are unlikely to be identical, but if $M$ is sufficiently large they are likely to be close to each other. Thus, if the threshold $\zeta$ in Algorithm 2, which determines the size of the neighborhood for clustering all novel words belonging to the same topic, is made sufficiently small, then each neighborhood will have only novel words belonging to the same topic.

With the two modifications discussed above, the number of distinct neighborhoods of a suitably nonzero size (determined by $\zeta > 0$) among the words whose solid angle estimates are larger than some threshold $\tau > 0$ will provide an estimate of $K$. The values of $\tau$ and $\zeta$ should, in principle, decrease to zero as $M$ increases to infinity. Leaving the task of unraveling the dependence of $\tau$ and $\zeta$ on $M$ to future work, here we only provide a brief empirical validation on both the Semi-Syn + Novel and Semi-Syn NYT datasets. We set $M = 2, 000, 000$ so that the reconstruction error has essentially converged (see Fig. 5), and consider different choices of the threshold $\zeta$.

We run Algorithm 2 with $K = 100$, $P = 150 \times K$, and a new line of code: 16': (if $\{\hat{q}_i = 0\}$, break); inserted between lines 16 and 17. The values of $j$ and (little) $k$ when Algorithm 2 terminates are indicated, respectively, by the position of the vertical dashed line and the rectangular box next to it for different $\zeta$.

| $\zeta$ | Solid Angle | Words |
|--------|-------------|-------|
| 0.1    | 100         | 1200  |
| 1      | 100         | 1200  |
| 5      | 100         | 1200  |
| 10     | 100         | 1200  |

Fig. 6. Solid-angles (in descending order) of all 14943 + 100 words in the Semi-Syn + Sep NYT dataset (left) and all 14943 words in the Semi-Syn NYT dataset (right) estimated (for different values of $\zeta$) by Algorithm 2 with $K = 100$, $P = 150 \times K$, $M = 2, 000, 000$, and a new line of code: 16': (if $\{\hat{q}_i = 0\}$, break); inserted between lines 16 and 17. The values of $j$ and (little) $k$ when Algorithm 2 terminates are indicated, respectively, by the position of the vertical dashed line and the rectangular box next to it for different $\zeta$.

B. Real-World Data

We now describe results for two real-world text corpora: the NYT dataset that was used in Sec. VI-A to construct the semi-synthetic datasets ($M = 300, 000$; $W = 14, 943$; $N \approx 300$) and the NIPS dataset of articles from the NIPS conference ($M = 1, 500; W = 12, 419; N \approx 1266$) [8]. In the NYT dataset, $M$ is large and $N$ is relatively small. The situation is roughly the opposite in the NIPS dataset. Following standard practice, only words that appear in more than 500 documents in the NYT and more than 50 documents in the NIPS dataset are kept. We also delete a standard list of stop-words as recommended in [35]. For RecoverL2, as prescribed by the authors in [7, Suppl.], we further restrict the novel words to those whose document frequency exceeds a specified threshold.

Since ground truth topics are unavailable, we measure performance using the so-called predictive held-out log-probability. This is a standard measure which is typically used to evaluate how well a learned topic model fits real-world data. To calculate this for each of the three topic estimation methods (Gibbs [11], [34], RecoverL2 [7], and RP), we first randomly select 60,000 documents from NYT (resp. 300 from NIPS) to test the goodness of fit and use the remaining 240,000 NYT (resp. 1200 NIPS) documents to produce an estimate $\hat{\beta}$ of the topic matrix. Next we assume a Dirichlet prior on the topics and estimate its concentration hyper-parameter $\alpha$. In Gibbs, this estimate $\hat{\alpha}$ is a byproduct of the algorithm. In RecoverL2 and RP this can be estimated from $\hat{\beta}$ and $\hat{X}$. We then calculate the probability of observing the test documents given the learned topic model $\hat{\beta}$ and $\hat{X}$: $\log Pr(X_{\text{test}}|\hat{\beta}, \hat{X})$. Since an exact evaluation of this predictive log-likelihood is intractable in general, we calculate it using the MCMC based approximation proposed in [20] which is now a standard approximation tool [34]. For RP, we use $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$ as in Sec. VI-A. We report the held-out log probability, normalized by the total number of words in the test documents, averaged across 5 training/testing splits. The results are summarized in Table II. As shown in Table II, Gibbs has the best descriptive power for new documents. RP and RecoverL2 have similar, but somewhat lower values than Gibbs. This may be attributed to missing novel words that appear only in the test set and are crucial to the success of RecoverL2 and RP. Specifically, in real-world examples, there is a model-mismatch as a result of which the data likelihoods of RP and RecoverL2 suffer.

Finally, we qualitatively access the topics produced by our RP algorithm. Table III shows some example topics and all the novel words for those topics produced by RP trained on all the
computational complexity but also state-of-the-art performance on semi-synthetic and real-world datasets.

This work focused on the standard centralized batch implementation, but it turns out that our random projections algorithm is naturally amenable to an efficient distributed implementation which is of interest when the documents are stored on a network of distributed servers. The distributed implementation can provably match the polynomial computational and statistical efficiency guarantees of its centralized counterpart. It therefore provides a provably efficient alternative to the distributed topic estimation problem which has been tackled using variations of MCMC or Variational Bayes in the literature [21], [36]–[38] This is appealing for modern web-scale databases as evidenced by, for instance, Twitter Streaming. A comprehensive theoretical and empirical investigation of the distributed variation of our algorithm can be found in [5].

Topic models like LDA discussed in this paper belong to the larger family of Mixed Membership Latent Variable Models [13] which have been successfully employed in a variety of problems that include text analysis, genetic analysis, network community detection, and ranking and preference discovery. The structure-leveraging approach proposed in this paper can be potentially extended to this larger family of models. Some initial steps in this direction for rank and preference data are explored in [33].

Proofs of all results appear in [17].

### VII. Conclusion and Discussion

This paper proposed a provably consistent and efficient algorithm for topic discovery by leveraging geometric implications of topic separability—a natural structural property for topic matrices. We resolved the necessary and sufficient conditions that can guarantee both consistent novel words detection and topic estimation. We then proposed a random projections based algorithm that has not only provably polynomial statistical and computational complexity but also state-of-the-art performance on semi-synthetic and real-world datasets.

This work focused on the standard centralized batch implementation, but it turns out that our random projections algorithm is naturally amenable to an efficient distributed implementation which is of interest when the documents are stored on a network of distributed servers. The distributed implementation can provably match the polynomial computational and statistical efficiency guarantees of its centralized counterpart. It therefore provides a provably efficient alternative to the distributed topic estimation problem which has been tackled using variations of MCMC or Variational Bayes in the literature [21], [36]–[38] This is appealing for modern web-scale databases as evidenced by, for instance, Twitter Streaming. A comprehensive theoretical and empirical investigation of the distributed variation of our algorithm can be found in [5].

Topic models like LDA discussed in this paper belong to the larger family of Mixed Membership Latent Variable Models [13] which have been successfully employed in a variety of problems that include text analysis, genetic analysis, network community detection, and ranking and preference discovery. The structure-leveraging approach proposed in this paper can be potentially extended to this larger family of models. Some initial steps in this direction for rank and preference data are explored in [33].

Proofs of all results appear in [17].

### REFERENCES

[1] D. Blei, “Probabilistic topic models,” Commun. ACM, vol. 55, no. 4, pp. 77–84, 2012.
[2] W. Ding, M. Rohban, P. Ishwar, and V. Saligrama, “A new geometric approach to latent topic modeling and discovery,” in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., Vancouver, BC, Canada, May 2013, pp. 5568–5572.
[3] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama, “Topic discovery through data dependent and random projections,” in Proc. 30th Int. Conf. Mach. Learn., Atlanta, GA, USA, Jun. 2013, pp. 1202–1210.
[4] W. Ding, P. Ishwar, M. H. Rohban, and V. Saligrama, “Necessary and sufficient conditions for novel word detection in separable topic models,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS’13) Workshop Topic Models Comput. Appl. Eval., Lake Tahoe, NV, USA, Dec. 2013.
[5] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama, “Efficient distributed topic modeling with provable guarantees,” in Proc. 17th Int. Conf. Artif. Intell. Statist., Reykjavik, Iceland, Apr. 2014, pp. 167–175.
[6] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.
[7] S. Arora et al., “A practical algorithm for topic modeling with provable guarantees,” in Proc. 30th Int. Conf. Mach. Learn., Atlanta, GA, USA, Jun. 2013, pp. 280–288.
[8] M. Lichman. (2013). UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml
[9] S. Arora, R. Ge, and A. Moitra, “Learning topic models—Going beyond SVD,” in Proc. IEEE 53rd Annu. Symp. Found. Comput. Sci., New Brunswick, NJ, USA, Oct. 2012, pp. 1–10.
[10] D. Sontag and D. Roy, “Complexity of inference in latent Dirichlet allocation,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS’11), Dec. 2011, pp. 1008–1016.
[11] T. Griffiths and M. Steyvers, “Finding scientific topics,” Proc. Nat. Acad. Sci., vol. 101, pp. 5228–5235, 2004.
[12] M. Wainwright and M. Jordan, “Graphical models, exponential families, and variational inference,” Found. Trends Mach. Learn., vol. 1, nos. 1–2, pp. 1–305, 2008.
[13] E. Airoldi, D. Blei, E. Erosheva, and S. Fienberg, Handbook of Mixed Membership Models and Their Applications. Boca Raton, FL, USA: CRC Press, 2014.
DING et al.: PROVABLY EFFICIENT ALGORITHM FOR SEPARABLE TOPIC DISCOVERY

[11] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, “A tensor approach to learning mixed membership community models,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 2239–2312, 2014.

[12] A. Kumar, V. Sindhwani, and P. Kambadur, “Fast conical hull algorithms for near-separable non-negative matrix factorization,” in Proc. 30th Int. Conf. Mach. Learn., Atlanta, GA, USA, Jun. 2013, pp. 231–239.

[13] W. Ding, P. Ishwar, and V. Saligrama, “Most large topic models are approximately separable,” in Proc. 10th IEEE Int. Workshop Inf. Theory Appl., San Diego, CA, USA, Feb. 2015, pp. 199–203.

[14] W. Ding, P. Ishwar, and V. Saligrama, “Necessary and sufficient conditions and a provably efficient algorithm for separable topic discovery,” arXiv preprint arXiv:1508.05565 [cs.LG], 2015.

[15] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. 22nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Berkeley, CA, USA, Aug. 1999, pp. 50–57.

[16] D. Blei and J. Lafferty, “A correlated topic model of science,” Ann. Appl. Statist., vol. 1, no. 1, pp. 17–35, 2007.

[17] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in Proc. 26th Int. Conf. Mach. Learn., Montreal, QC, Canada, Jun. 2009, pp. 1105–1112.

[18] M. Hoffman, F. R. Bach, and D. M. Blei, “Online learning for latent Dirichlet allocation,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS'10), Lake Tahoe, NV, USA, Dec. 2010, pp. 856–864.

[19] D. Lee and H. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[20] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. Hoboken, NJ, USA: Wiley, 2009.

[21] B. Recht, C. Re, J. Tropp, and V. Bittorf, “Factoring nonnegative matrices with linear programs,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS'12), Lake Tahoe, NV, USA, Dec. 2012, pp. 1223–1231.

[22] S. Vavasis, “On the complexity of nonnegative matrix factorization,” SIAM J. Optim., vol. 20, no. 3, pp. 1364–1377, Oct. 2009.

[23] A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade, “Learning linear Bayesian networks with latent variables,” in Proc. 30th Int. Conf. Mach. Learn., Atlanta, GA, USA, Jun. 2013, pp. 249–257.

[24] D. Lee and H. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[25] T. Bansal, C. Bhattacharyya, and R. Kannan, “A provable SVD-based algorithm for learning topics in dominant admixture corpus,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS'14), Montral, QC, Canada, Dec. 2014, pp. 1997–2005.

[26] J. Boardman, “Automating spectral unmixing of AVIRIS data using convex geometry concepts,” in Proc. Ann. JPL Airborne Geosci. Workshop, 1993, pp. 11–14.

[27] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?” in Proc. Adv. Neural Inf. Process. Syst. (NIPS'03), Lake Tahoe, NV, USA, Dec. 2003, pp. 1141–1148.

[28] N. Gillis and S. A. Vavasis, “Fast and robust recursive algorithms for separable nonnegative matrix factorization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 4, pp. 698–714, Apr. 2014.

[29] V. Farias, S. Jagabathula, and D. Shah, “A data-driven approach to modeling choice,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS'09), Vancouver, BC, Canada, Dec. 2009, pp. 504–512.

[30] W. Ding, P. Ishwar, and V. Saligrama, “A topic modeling approach to ranking,” in Proc. 18th Int. Conf. Artif. Intell. Statist., San Diego, CA, USA, May 2015, pp. 214–222.

[31] A. McCallum. (2002). Mallet: A Machine Learning for Language Toolkit [Online]. Available: http://mallet.cs.umass.edu

[32] D. Lewis, Y. Yang, T. Rose, and F. Li, “RCV1: A new benchmark collection for text categorization research,” J. Mach. Learn. Res., vol. 5, pp. 361–397, Dec. 2004.