Most Large Topic Models are Approximately Separable

Weicong Ding  
Department of ECE  
Boston University  
Boston, MA 02215, USA  
Email: dingwc@bu.edu

Prakash Ishwar  
Department of ECE  
Boston University  
Boston, MA 02215, USA  
Email: pi@bu.edu

Venkatesh Saligrama  
Department of ECE  
Boston University  
Boston, MA 02215, USA  
Email: srv@bu.edu

Abstract—Separability has recently been leveraged as a key structural condition in topic models to develop asymptotically consistent algorithms with polynomial statistical and computational efficiency guarantees. Separability corresponds to the presence of at least one novel word for each topic. Empirical estimates of topic matrices for Latent Dirichlet Allocation models have been observed to be approximately separable. Separability may be a convenient structural property, but it appears to be too restrictive a condition. In this paper we explicitly demonstrate that separability is, in fact, an inevitable consequence of high-dimensionality. In particular, we prove that when the columns of the topic matrix are independently sampled from a Dirichlet distribution, the resulting topic matrix will be approximately separable with probability tending to one as the number of rows (vocabulary size) scales to infinity sufficiently faster than the number of columns (topics). This is based on combining concentration of measure results with properties of the Dirichlet distribution and union bounding arguments. Our proof techniques can be extended to other priors for general nonnegative matrices.

I. INTRODUCTION

Topic models such as Latent Dirichlet Allocation (LDA) are an important class of Mixed Membership Latent Variable Models that have been extensively studied over the last decade [1, 2]. They consider a corpus of text documents composed of words from a fixed vocabulary and view each document as a probabilistic mixture of a few latent “topics” that are shared across the corpus. Each topic is modeled as a distribution over the vocabulary. The primary learning problem here is to estimate the latent topics given the observations.

In its full generality, this topic discovery problem is intractable and \( \mathcal{NP} \)-hard [10]. The popular prevailing approaches to this problem make use of non-parametric Bayes approximation methods such as variational Bayes and MCMC [3, 4, 11, 12]. Despite their “satisfactory” empirical performance in several real-world datasets, the lack of asymptotic consistency and sample/algorithmic efficiency guarantees makes it difficult to evaluate model fidelity: failure to produce satisfactory results could be due to the use of approximations or due to model mis-specification which is more fundamental. Furthermore, they are known to be computation-intensive for large problem sizes [3, 8].

While the general topic estimation problem is hard, recent work has demonstrated that the topic discovery problem can lend itself to provably efficient solutions under additional structural conditions [e.g., 7, 10, 13, 15]. One key condition that has been successfully leveraged in recent work is topic separability: each topic has at least one novel word that is primarily unique to that topic [e.g., 5, 6, 8, 14, 17]. This is, in essence, a property of the support of the topic matrix. Recent work has shown that the latent topics in separable topic models can be learned consistently with polynomial sample and computational complexity. In addition, empirical topic estimates produced by nonparametric Bayes methods have been observed to be approximately separable. Despite these appealing properties, the separability condition appears to be rather restrictive and somewhat artificial. Is it merely an assumption of convenience or is it more fundamental?

In this paper we explicitly demonstrate that the separability condition is a natural and inevitable consequence of the high dimensionality of the topic modeling problem. Specifically, we prove that when the columns of the topic matrix are independently sampled from a Dirichlet distribution, the resulting topic matrix will be approximately separable with probability tending to one as the number of rows (vocabulary size) scales to infinity sufficiently faster than the number of columns (topics). This is based on combining concentration of measure results with properties of the Dirichlet distribution and union bounding arguments. Our results formally validate that separability is a good approximation for most large topic models.

The rest of this paper is organized as follows. We first briefly overview the topic modeling problem in Section II and formally define the separability condition. We then summarize our main results on separability and outline the key proof steps in Section III. In Section IV, we discuss the implications of our results in detail. Specifically, we show that the results of our analysis agree with and provide a justification for some of the practical guidelines that have been used in the literature.

II. SEPARABLE TOPIC MODELS

In topic models, we have a collection of \( M \) documents, each composed of \( N \geq 2 \) words from a fixed vocabulary of size \( W \). Topic models are based on the classic “bags of words” modeling paradigm where the \( N \) words are modeled
as i.i.d. drawings from an unknown $W \times 1$ document word-distribution vector. Each document word-distribution vector is itself modeled as an unknown probabilistic mixture of $K \ll W$ unknown topics that are shared among the $M$ documents. Each latent topic is a $W \times 1$ distribution vector over the vocabulary.

We denote the $W \times K$ column-stochastic topic matrix whose $K$ columns are the $K$ latent topics by $\beta$. We denote by $\theta_m$ the probabilistic weight vector over the $K$ latent topics for document $m$. Each $\theta_m$ is independently sampled from a prior distribution such as Dirichlet in LDA [1] or Log-Normal in the Correlated Topic Model [18]. The generative process and the corresponding graphical representation are summarized in Figure 1. The primary estimation problem is to learn the topic matrix $\beta$ given the words in the $M$ documents.

![Graphical plate representation of a standard topic model](image)

**Definition 1:** ($\lambda$-approximate separability) A $W \times K$ non-negative matrix $\beta$ is $\lambda$-approximately separable for some constant $\lambda \in [0, 1)$, if $\forall k = 1, \ldots, K$, there exists at least one row (word) $i$ such that $\beta_{i,k} > 0$ and $\beta_{i,l} \leq \lambda \beta_{i,k}$, $\forall l \neq k$.

The $\lambda$-approximate separability condition requires the existence of words (rows of $\beta$) that occur *predominantly* in one topic (column of $\beta$) and have relatively negligible occurrences in the other topics, i.e., the row-weight concentrates predominantly in one column. We will refer to such words (rows of $\beta$) as $\lambda$-approximately novel words (rows). The smaller the value of $\lambda$, the sharper the concentration within a single topic and higher the novelty of the word and separability of the topic. When $\lambda = 0$, we will say that $\beta$ is exactly separable [3, 4, 8].

Table I illustrates the probability of generating a 0.01-separable topic matrix $\beta$ for different values of $K$ and $W$ that are encountered in some real-world benchmark datasets in the Topic Modeling literature. In these practical settings, the size of the vocabulary $W$ is much larger than the number of latent topics $K$. The $K$ columns of $\beta$ are i.i.d. samples from a symmetric Dirichlet prior $\text{Dir}(\beta_0)$ with concentration parameter $\beta_0 = 0.01$. The probability is estimated using 1000 Monte Carlo runs. The $3\sigma$ confidence intervals for the probability estimates are also indicated. This table demonstrates that approximate separability is a highly likely occurrence in real-world-sized datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Vocab. size</th>
<th># Topics $K$</th>
<th>Prob. 0.01-separable</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS [3, 19]</td>
<td>12,419</td>
<td>50</td>
<td>100 ± 0.0%</td>
</tr>
<tr>
<td>Wikipedia [3]</td>
<td>109,641</td>
<td>50</td>
<td>99.9 ± 0.3%</td>
</tr>
<tr>
<td>Twitter [9]</td>
<td>122,035</td>
<td>50</td>
<td>100 ± 0.1%</td>
</tr>
<tr>
<td>NYT [8, 19]</td>
<td>102,660</td>
<td>100</td>
<td>99.6 ± 0.6%</td>
</tr>
<tr>
<td>PubMed [3, 19]</td>
<td>141,043</td>
<td>100</td>
<td>99.9 ± 0.3%</td>
</tr>
</tbody>
</table>

### III. Inevitability of Separability

In this section, we analyze the probability that $\beta$ is $\lambda$-approximately separable for any small constant $\lambda \ll 1$. We provide an analytical framework to derive an upper bound for the probability that $\beta$ is not $\lambda$-approximately separable. We focus on the Dirichlet prior, but this framework can be extended to handle other priors for $\beta$. Recalling that $W$ is the size of vocabulary, $K$ is the number of latent topics, and $\beta_0$ is the concentration parameter of the Dirichlet prior on the columns of $\beta$, we have the following result:

**Lemma 1:** Let the $K$ columns of the topic matrix $\beta$ be generated independently from $\text{Dir}(\beta_0)$ for $\beta_0 \in (0, 1)$. Then, the probability that $\beta$ is $\lambda$-approximately separable is at least

$$1 - K c_1 \exp(-c_2 W \beta_0) - K \exp(-W p_1(\beta_0, \lambda/4, K))$$

(1)

where $c_1, c_2$ are some absolute constants and $p_1(\beta_0, \lambda/4, K)$ is the probability that a $1 \times K$ row vector with independent gamma($\beta_0, 1$)-distributed entries is a $\lambda/4$-approximately novel row for the first topic. This can be lower bounded as follows:

$$p_1(\beta_0, c, K) \geq \frac{c_3}{K} \left(\frac{c}{cK + 1 - c}\right)^{\beta_0 K}$$

(2)

where $c_3$ is some absolute constant.

The key idea underlying our approach is to reduce the analysis of the separability properties of $\beta$ to that of a related $W \times K$ dimensional random matrix whose rows are independent. Then computing the probability that $\beta$ is approximately separable reduces to examining of the probability that each independent
row vector in the related matrix is approximately novel to one of the $K$ topics.

**Proof:** Each Dir($\beta_0$)-distributed column $\beta_k$ can be generated by first sampling each of its $W$ entries independently from a gamma distribution with parameter $\beta_0$, and then dividing all the column entries by their sum in order to make the column-sum equal to one (column-normalization). We will refer to the un-normalized $W \times K$ random matrix with independent gamma($\beta_0, 1$)-distributed entries as the "gamma random matrix".

Our overall analysis approach is to (a) first calculate the probability that a row of the gamma random matrix is $\lambda/4$-approximately novel for a topic, i.e., $p_1(\beta_0, \lambda/4, K)$ as defined in Lemma 1 and (b) then show that all the column-normalization factors will concentrate around their means when $W$ is large and will therefore not impact the approximate-separability property of the gamma random matrix.

To formalize the above ideas, let $\mu_{w,k}, w = 1, \ldots, W, k = 1, \ldots, K$ be i.i.d samples from the gamma($\beta_0, 1$) distribution. We denote by $b_k = \sum_{w=1}^{W} \mu_{w,k}$ the column-normalization factor for the $k$-th column. Let $A$ denote the event that all the normalization factors $b_k, k = 1, \ldots, K$, are within a $W\beta_0/2$ radius of their means $W\beta_0$. Let $B$ denote the event that the gamma random matrix has at least one $\lambda/4$-approximately novel word for each topic. When event $A$ occurs, then $\forall i, j, i \neq j, \beta_i/b_j \in (1/4, 4)$. Then the $\lambda/4$-approximate novel words of the gamma random matrix will become at most $\lambda$-approximate novel words after column-normalization. Thus, for the event that $\beta$ is $\lambda$-approximately separable to occur it is sufficient that the intersection of events $A \cap B$ occurs.

For event $B$, we define $p_1(\beta_0, \lambda/4, K)$ to be the probability that the first row of the gamma random matrix is $\lambda/4$ approximately novel for the first column (topic 1). Since all entries in the gamma random matrix are i.i.d., the probability that any row of the gamma random matrix is approximately novel for any column would be exactly the same for all rows and columns (by symmetry). Next, note that

$$B^c = \bigcup_{k=1}^{K} B_k,$$

where $B_k$ is the event that none of the $W$ rows in the gamma random matrix is $\lambda/4$ approximately novel for the $k$-th topic. Since the rows of the gamma random matrix are independent, we have

$$\Pr(B_k) = (1 - p_1(\beta_0, \lambda/4, K))^W \leq \exp(-Wp_1).$$

Therefore, using the union bound, we get $\Pr(B^c) \leq K \exp(-Wp_1)$.

We then consider $A = \{\forall k, |b_k - W\beta_0| \leq W\beta_0/2\}$. Note that by law of large numbers for sub-Gaussian random variables, we have $\Pr(|b_k - W\beta_0| > \frac{1}{2}W\beta_0) \leq c_1 \exp(-c_2W\beta_0)$ for some absolute constants $c_1$ and $c_2$.

Therefore, $\Pr(A^c) \leq Kc_1 \exp(-c_2W\beta_0)$. Putting it all together, the probability that $\beta$ is $\lambda$-approximately separable is lower bounded by the probability of the intersection of $A$ and $B$, which is lower bounded by

$$c_1K \exp(-c_2W\beta_0) + K \exp(-Wp_1)$$

It remains to derive an explicit formula or bound for $p_1$. This is summarized in Lemma 2.

**Lemma 2:** Let $\mu = [\mu_1, \ldots, \mu_K]$ be a $1 \times K$ row vector where the $\mu_k$’s are i.i.d samples from the gamma($\beta_0, 1$) distribution. Then, the probability that $\mu$ is a $c$-approximately row novel for topic 1, $p_1(\beta_0, c, K)$, can be lower bounded as follows:

$$p_1(\beta_0, c, K) \geq \frac{c_3}{K} (\frac{c}{cK + 1 - c})^\beta_0K $$

**Proof:** Note that by definition of separability in Def. 1,

$$p_1(\beta_0, c, K) = \Pr(\mu_2 \leq \mu_1, \ldots, \mu_K \leq c\mu_1)$$

$$= \int_0^{\infty} \Pr(\mu_2 \leq c\mu_1, \ldots, \mu_K \leq c\mu_1|\mu_1)p(\mu_1)d\mu_1$$

$$= \int_0^{\infty} \gamma(\beta_0, c\mu_1)^{K-1}p(\mu_1)d\mu_1$$

where $\gamma(\beta_0, c\mu_1) = \int_0^{c\mu_1} x^{\beta_0-1} \exp(-x)dx$ is the incomplete gamma function (i.e., the CDF of the gamma distribution). We first consider a lower bound for the incomplete gamma function in closed-form,

$$\gamma(\beta_0, c\mu_1) = \int_0^{c\mu_1} x^{\beta_0-1} \exp(-x)dx$$

$$\geq \frac{\exp(-c\mu_1)}{\Gamma(\beta_0)} \int_0^{c\mu_1} x^{\beta_0-1}dx$$

$$= \frac{\exp(-c\mu_1)(c\mu_1)^{\beta_0}}{\Gamma(\beta_0) \beta_0}.$$
Therefore, for large $K$, $\Gamma(\beta_0 K) > \Gamma(\beta_0)$. In the region where $\beta_0 K < 1$, one can show that $\Gamma(K\beta_0)/\Gamma(\beta_0) = O(1/K)$. We also note that $c < 1$ and $\beta_0 < 1$ so that $c^{\beta_0} < 1$. Hence for $p_1(\beta_0, c, K)$, we have,

$$p_1(\beta_0, c, K) \geq \frac{c^\beta}{K} \left( \frac{c^K}{cK + 1 - c} \right)^{\beta_0 K}.$$ 

IV. Discussion and Implications

In this section we discuss some insights and implications that follow from Lemma 1. We first note that from Eq. (1), the upper bound on the probability that $\beta$ is not $\lambda$-approximately separable decays exponentially in $W$, the size of vocabulary, which is typically very large.

If we require that the probability that $\beta$ is not $\lambda$-approximately separable should decay at a polynomial rate with respect to $W$ (with $K$ held fixed), i.e., $\frac{W}{\log(W)}$ for some positive degree $a > 0$, then by Eq. (1), it suffices to require that

$$\frac{W}{\log(W)} \geq (a + 1)/\min\{c_2\beta_0, p_1\}$$

(4)

If the number of latent topics $K$ also scales, noting that $p_1$ is a function of $K$, we need to require that $W$ scale as

$$\frac{W}{\log(W)} \geq (a + 1) \max\left\{ \frac{1}{c_2\beta_0}, \frac{K}{c_3} \left( K - 1 + \frac{1}{\lambda} \right)^{\beta_0 K} \right\}$$

(5)

A. Role of hyper-parameter $\beta_0$

Equation (5) indicates that if $\beta_0$ is moderately small, the topic matrix is more likely to be separable and can be estimated using algorithms, such as those in [5, 6, 8], that come with provable guarantees.

In fact, this implication of our analysis agrees with the practical guidelines adopted in the topic modeling literature to set the hyper-parameters. We first note that it has been empirically observed that topic models with a moderately small $\beta_0$ can be more efficiently learned using approximation methods compared to those with a larger $\beta_0$ (especially $\beta_0$ close to 1) [e.g., 9]. In the literature, the hyper-parameter $\beta_0$ is often set to a moderately small positive number [e.g., 1, 2, 3, 4, 9, 11, 20]. This is in accordance with our alternative explanation using the separability condition.

Further, we note that a smaller $\beta_0$ can indeed compensate for the exponential dependency of $W$ on $K$ in Eq. (5). As reported in the literature, empirically satisfactory results are often obtained with $\beta_0 \approx 0.01$ and the number of latent topics ranging from $K = 50 \sim 200$ [2, 3, 9, 20] (also see Table I). For these values, the exponent $\beta_0 K$ in Eq. (5) would range from 0.5 to 2. Hence the requirement in Eq. (5) can be satisfied for moderate values of $W$.

Finally, we note that in popular topic modeling packages such as [21], the default hyper-parameter setting is $\beta_0 = 0.01$. In other packages such as [11, 22], it is even suggested that the hyper-parameter be set according to the rule $\beta = c/W$, for some constant $c \approx 200$, in order to obtain good empirical results.

B. Role of $\lambda$

In terms of the degree of approximate separability, i.e., the small constant $\lambda$, a scenario of special interest is when the weight (entry in the topic matrix) of each novel word in its corresponding topic is much larger than its cumulative weight in all the remaining topics, e.g., $\sum_{k=2}^{K} \beta_{i,k} \ll \beta_{i,1}$ if word $i$ is a $\lambda$-approximately novel word for topic 1. This translates to $\lambda(K - 1) \ll 1$ or $\lambda \ll 1/K$. In this scenario, the expression in Eq. (5) can be further simplified. Note that when $\lambda \ll 1/K$, $(K - 1 + 1/\lambda)$, is dominated by $1/\lambda$. Therefore, Eq. (5) can be simplified to the following

$$\frac{W}{\log(W)} \geq (a + 1) \max\left\{ \frac{1}{c_2\beta_0}, \frac{K}{\lambda\beta_0 K} \right\}.$$

C. Validation using Parameters in Benchmark Datasets

As explained in Sec. I in order to validate the separability condition in real-world problems, we conducted the following simulations. We first obtained the parameters of some benchmark datasets that have been used in the topic modeling literature, specifically, the size of the vocabulary $W$ as well as the number of latent topics specified $K$. We then generated random realizations of the topic matrix $\beta$ and checked if the $\lambda$-approximate separability condition is satisfied.

As discussed in previous sections, we set $\beta_0 = 0.01$ and consider $\lambda = 0.01$-approximate separability. For each setting, we generated 1000 Monte Carlo runs to estimate the probability of generating a 0.01-approximately separable matrix. The results are summarized in Table I. We can observe that in most examples, the topic matrix is 0.01-approximately separable with very high probability.

D. Other Mixed Membership Latent Variable Models

Topic models such as LDA are an example of the general family of Mixed Membership Latent Variable Models [23]. Mixed membership latent variable models have been studied in a wide range of other problems including rank aggregation, community discovery, etc. [24, 25]. Although our analysis in this paper focused on topic models, one can show, using similar techniques, that the corresponding topic matrix in a topic-based ranking model [24, 26], in a mixture of Mallows model [27, 28], and in mixed membership stochastic blockmodels [25, 29] is separable with an overwhelmingly large probability when the number of rows grows much faster than the number of columns. We defer the derivation of analogous results for these problems to future publications.

ACKNOWLEDGMENT

This article is based upon work supported by the U.S. AFOSR and the U.S. NSF under award numbers # FA9550-10-1-0458 (subaward # A1795) and # 1218992 respectively. The views and conclusions contained in this article are those of the authors and should not be interpreted as necessarily
representing the official policies, either expressed or implied, of the agencies.

REFERENCES


