Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis

Jian Tang[1] TANGJIAN@NET.PKU.EDU.CN
Zhaoshi Meng[2] MENGZS@UMICH.EDU
XuanLong Nguyen[3,2] XUANLONG@UMICH.EDU
Qiaozhu Mei[4,2] QMEI@UMICH.EDU
Ming Zhang[1] MZHANG@NET.PKU.EDU.CN
[1]School of EECS, Peking University, [2]Department of EECS, University of Michigan
[3]Department of Statistics, University of Michigan, [4]School of Information, University of Michigan

Abstract

Topic models such as the latent Dirichlet allocation (LDA) have become a standard staple in the modeling toolbox of machine learning. They have been applied to a vast variety of data sets, contexts, and tasks to varying degrees of success. However, to date there is almost no formal theory explicating the LDA's behavior, and despite its familiarity there is very little systematic analysis of and guidance on the properties of the data that affect the inferential performance of the model. This paper seeks to address this gap, by providing a systematic analysis of factors which characterize the LDA's performance. We present theorems elucidating the posterior contraction rates of the topics as the amount of data increases, and a thorough supporting empirical study using synthetic and real data sets, including news and web-based articles and tweet messages. Based on these results we provide practical guidance on how to identify suitable data sets for topic models, and how to specify particular model parameters.

1. Introduction

Topic models such as the latent Dirichlet allocation (LDA) have become a familiar tool in machine learning (Blei et al., 2003; Pritchard et al., 2000). They have played an important role in a variety of data mining tasks, both within the scope of computer science (Blei et al., 2003; Griffiths & Steyvers, 2004; Ramage et al., 2010; Liu et al., 2009; Wang & Blei, 2011) and reaching out to other fields, such as biomedical informatics (Ling et al., 2007), scientometrics (Griffiths & Steyvers, 2004; McCallum et al., 2006), social and political science (Ramage et al., 2009; Grimmer, 2010), and digital humanities (Mimno, 2012). Despite these successes, a thorough understanding of the LDA's statistical behavior is lacking. Certain properties of the LDA have become part of the machine learning folklore, but never been formally proven; some other properties seem to be misunderstood, resulting in instances of the model being used but not truly applicable.

For instance, the LDA is known to achieve promising results in modeling text collections such as news articles (Blei et al., 2003; Newman et al., 2006), scientific papers (Griffiths & Steyvers, 2004; Wang & Blei, 2011), and blogs (Liu et al., 2009). However, the results are mixed when it is applied to unconventional data sets such as tweets (i.e., short messages posted on Twitter.com) (Hong & Davison, 2010; Zhao et al., 2011), query logs (Bing et al., 2011; Carman et al., 2010), digital books (Mimno & McCallum, 2007), and metadata records (Newman et al., 2010). These data sets, harvested from different real tasks, usually fall into the situations where there are too few documents (e.g., Shakespeare's plays), the documents are too short (e.g., tweets, queries), or a document contains many topics (e.g., books). As a result, ad hoc heuristics have to be employed to preprocess the documents, such as to break books into pages so that each page becomes a shorter "document" (Mimno & McCallum, 2007), or to aggregate tweets with the same attribute (e.g., written by the same user) into longer "documents" (Hong & Davison, 2010; Bing et al., 2011). Alternatively, arbitrary manipulations of the LDA model have to be introduced in order to adapt the topic model to a particular context (e.g., Mimno & McCallum (2007); Zhao et al. (2011); Carman et al. (2010)).

We are motivated by eager non-expert consumers of topic modeling, who often ask questions such as: *Is my data*

Proceedings of the 31^{st} *International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

topic-model "friendly?" Why did the LDA fail on my data? How many documents do I need to learn 100 topics? It is difficult even for experts to provide quick and satisfactory answers to those questions. As noted above some behaviors of the LDA are known intuitively in the machine learning folklore, but they are never theoretically justified, such as the LDA's deficiency in handling short documents. Other situations remain quite mysterious, such as how the LDA performs on a few lengthy documents, and what happens when it overfits with a larger number of topics than actually present in the data. We will show in this paper, like Liebig's barrel, the performance of LDA is limited by the length of the shortest stave, making it challenging to judge its effectiveness without thorough empirical experiments.

In this paper we aim to provide a systematic analysis of the behavior of the LDA in various settings. We identify several limiting factors whose interactions play the crucial role in determining the LDA's performance. These factors include the number of documents, the length of individual documents, the number of topics, and the Dirichlet (hyper)parameters. The main contributions include (i) theoretical results explicating the convergence behavior of the posterior distribution of latent topics as the amount of training data increases – as shown in Section 2, the convergence behavior is found to depend on the interactions of the limiting factors mentioned above; (ii) a thorough empirical study reported in Section 3 that provides support for the theory, by varying the settings of the limiting factors on synthetic and real data sets; (iii) these findings are translated to a number of concrete guidelines regarding the practical use of the LDA model – these are discussed in details in Section 4.

2. Posterior Contraction Analysis of Topic Polytope in the LDA

The latent Dirichlet allocation (LDA) model explains the generation of text documents, which can be viewed as samples from a mixture of multinomial distributions over a vocabulary of words. Each multinomial mixture component is called a "topic". We refer the reader to Blei et al. (2003) and Pritchard et al. (2000) for a background on the model. Here we provide a *geometric reformulation* of the LDA model and present several theoretical results that explain the behavior of the posterior distribution of the latent topic variables, as the number of training data increases.

2.1. Latent Topic Polytope in the LDA

Let V denote the vocabulary (set of words). Each topic ϕ_k is a vector in $\Delta^{|V|-1}$ – the $(|V|-1)$-dimensional probability simplex. We assume that the documents are generated by K topics $\Phi = (\phi_1 \ldots, \phi_K)$. Each document in the corpus $d \in \{1, \ldots, D\}$ is then associated with a topic proportion vector $\theta_d = (\theta_{d,1}, \ldots, \theta_{d,K}) \in \Delta^{K-1}$, where $\theta_{d,k}$ is the probability that each word in document d is assigned to topic k. Equivalently, document d uniquely corresponds to a word probability vector $\eta_d = \sum_{k=1}^{K} \theta_{d,k}\phi_k \in \Delta^{|V|-1}$. The words of document d, i.e. $W^d_{[N_d]} := (w_{dn})_{n=1}^{N_d}$, are independent and identically distributed samples from a multinomial distribution parameterized by this probability vector η_d. To simplify the analysis, we assume all documents have the same number of words $N_d = N$. The joint distribution of the full data set $W^{[D]}_{[N]} := (W^d_{[N]})_{d=1}^{D}$, denoted by $P^D_{W_{[N]}}$, is the product distribution of all single document distributions: $P^D_{W_{[N]}}(W^{[D]}_{[N]}) := \prod_{d=1}^{D} P_{W_{[N]}}(W^d_{[N]})$. Our primary interest is the inference of the topic parameters Φ on the basis of the sampled $D \times N$ words $W^{[D]}_{[N]}$, and how this inference is affected by the way that the words form into the documents. Note that compared to the original definition of Blei et al. (2003), this alternative representation does not involve latent assignment variables z_{dn}'s – they are simply marginalized out.

In a Bayesian estimation setting of the LDA model, the document-topic and topic-word proportion vectors (θ_d and ϕ_k's) are assumed to be random and endowed with Dirichlet prior distributions, parameterized by hyperparameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\beta = (\beta_1, \ldots, \beta_{|V|})$, respectively. Accordingly one is interested in the behavior of the *posterior* distribution of the topic parameters Φ given the observed documents. In particular, we want to understand the convergence behavior of the posterior topic distribution when the total amount of data $D \times N$ increases to infinity. It is expected that the posterior distribution of the topic variables should contract toward their true values as one has more data. A natural question to ask is the rates at which this posterior contraction phenomenon occurs.

In order to establish such an analysis, we shall introduce a metric which describes the (contracting) neighborhood centered at the true topic values, where the posterior distribution will be shown to place most its probability mass on. The faster the contraction, the more efficient the statistical inference. Although it would be ideal to examine the convergence for each individual topic parameter, it is challenging due to the issue of identifiability: one relatively minor problem is the "label-switching" issue, which means that one can only identify the collection of Φ up to a permutation. A more difficult identification problem is that, any vector that can be expressed as a convex combination of the topic parameters would be hard to identify and analyze. To address these theoretical difficulties, instead of investigating the convergence behavior of individual topics, we study the convergence of the latent topic structure through its convex hull:

$$G(\Phi) = \text{conv}(\phi_1, \ldots, \phi_K),$$

which is referred to as the *topic polytope* (cf. Nguyen

(2012)). By definition, all vectors that can be written as a convex combination of the collection of topics will lie in the polytope they span. To characterize the distance between two polytopes G and G', we use the "minimum-matching" Euclidean distance metric, defined as:

$$d_{\mathcal{M}}(G, G') = \max\{d(G, G'), d(G', G)\}, \tag{1}$$

where

$$d(G, G') := \max_{\phi \in \text{extr}(G)} \min_{\phi' \in \text{extr}(G')} \|\phi - \phi'\|_2, \tag{2}$$

is the maximum value of the best matching distances between the *extreme points* of G and G'. Intuitively this metric is a stable measure of the dissimilarity between two topic polytopes. Under very mild conditions, this metric can be proven to be equivalent to the well-known Hausdorff metric in convex geometry (Nguyen, 2012). This metric also has a very intuitive interpretation in the information retrieval context, as will be discussed later in Section 3.

Using the minimum-matching Euclidean metric, the theorems in the sequel give explicit upper bounds for the asymptotic posterior contraction rate, i.e., the rate at which the posterior distribution of the topic variables contracts around their true values.

2.2. Contraction of the Posterior of Topic Polytope

In order to state the theorems, several mild regularity conditions are required on the true topic polytope and the induced prior distribution Π on the polytopes. At a high level, these assumptions are required to prevent the topic polytope from being degenerate or collapsing. They are also needed to ensure that the prior topic distribution is thick (dense) enough in the space of parameters to be estimated (cf. Assumptions (S0)-(S3) in Nguyen (2012)). [1] Let $G^* = G(\Phi^*)$ be the true topic polytope spanned by the K^* true topic parameters, based on which the $D \times N$ data set $W_{[N]}^{[D]}$ is generated according to the true distribution denoted by $P^D_{W_{[N]}|G^*}$. In general we allow for $K \geq K^*$. The case of $K = K^*$ means the number of true topics is known, while $K > K^*$ corresponds to learning with an overfitted LDA model. Now we state the main theorems.

Theorem 1 *Assume the regularity conditions hold for prior Π and G^* is in the support of Π. If the Dirichlet parameters for topic proportions $\alpha_k \in (0, 1]$, and either one of the following two conditions holds:*

(A1) $K = K^*$, *i.e. the true number of topics is known;*

(A2) *The Euclidean distance between every pair of topics is bounded from below by a known positive constant r_0;*

then as data sizes $D \to \infty$ and $N \to \infty$ such that

$$\log \log D \leq \log N = o(D), \tag{3}$$

for some sufficiently large constant C independent of N and D,

$$\Pi\left(d_{\mathcal{M}}(G, G^*) \leq C \cdot \delta_{D,N} \mid W_{[N]}^{[D]}\right) \longrightarrow 1 \tag{4}$$

in $P^D_{W_{[N]}|G^*}$-probability. Here the upper bound for the contraction rate is

$$\delta_{D,N} = \left(\frac{\log D}{D} + \frac{\log N}{N} + \frac{\log N}{D}\right)^{\frac{1}{2}}. \tag{5}$$

We make a number of observations:

1. Thm. 1 characterizes an upper bound for the learning rate of the topic polytope through studying its extreme points. In order to ensure the contraction rate to hold, D and N are required to grow in a controlled manner as described in (3). In particular, we should have $\log D \leq N$, which means the length of the documents should be at least on the order of $\log D$ (up to a constant factor).

2. This upper bound rate is dominated by the maximum one between $(\frac{\log N}{N})^{\frac{1}{2}}$, $(\frac{\log D}{D})^{\frac{1}{2}}$, as well as $(\frac{\log N}{D})^{\frac{1}{2}}$. Our empirical study confirms the dependence on the first two terms, while the last term does not appear to play a noticeable role. We conjecture that the third term in the expression (5) may well be an artifact due to the proof techniques.

3. It should be emphasized that in practice the actual contraction rate of the LDA could be faster than the upper bound given above. However, this looseness of the upper bound only occurs in the exponent, the dependence on $\frac{1}{D}$ and $\frac{1}{N}$ should remain due to a lower bound of the form $\Omega(\frac{1}{DN})$ (cf. Thm. 3(c) in Nguyen (2012)). In our empirical study in Sec. 3.1.4, we find that the actual rate of the LDA is well-approximated by $\left(\frac{1}{N} + \frac{1}{D}\right)^r$, where $r \geq 1/2$.

4. The condition (A2) specifies a well-separated prior distribution (with known constant r_0) on the topics, which is reasonable in practice. Roughly speaking, this condition is satisfied with high probability when the hyper-parameter of the topic-word distribution is very small ($\beta \ll 1$), in which case the topic vectors mostly concentrate on a few words. This is also supported by our empirical study in the sequel.

5. That the convergence rate does not depend on the number of topics K is quite surprising and encouraging. This suggests that once K is known, or the topics are well-separated, the LDA inference is statistically efficient.

Observe that condition (A1) is probably not met in practice, because the true number of topics K^* is usually unknown. While underfitting will result in a persistent error even with infinite amount of data, we are most likely to prefer the

[1] In the technical report (Nguyen, 2012), m, n are used instead of D, N, respectively. The latter choice of notation is more common within the topic modeling literature.

overfitted setting, by having $K \gg K^*$. When neither Condition (A1) nor (A2) holds, a much more pessimistic upper bound is obtained, as stated in the following theorem:

Theorem 2 *Assume the regularity conditions hold for prior Π and G^* is in the support of Π. If the Dirichlet parameters for topic proportions $\alpha_k \in (0, 1]$ and $K^* < K \le |V|$, then as $D \to \infty$ and $N \to \infty$ such as in (3), Eq. (4) holds with*

$$\delta_{D,N} = \left(\frac{\log D}{D} + \frac{\log N}{N} + \frac{\log N}{D} \right)^{\frac{1}{2(K-1)}}. \qquad (6)$$

This theorem gives an upper bound for the contraction rate in general situations (i.e., either the selected number of topics exceeds the truth, or the topics are not known to be well-separated). Note how the upper bound deteriorates with K – it behaves like a nonparametric rate. Note that the size of the vocabulary $|V|$ does not appear in the rate because of the assumption that $K \le |V|$, so that the intrinsic dimension of the topic polytope is equal to $K - 1$. In fact, when $K > |V|$ one can obtain a contraction rate of similar form, except that the exponent rate $\frac{1}{2(K-1)}$ is replaced by $\frac{1}{2|V|}$. Although we give here only an upper bound, a minimax lower bound approximately of the form $\Omega(D^{-2/K})$ is also obtained, confirming the statistical *inefficiency* of the inference in general situations. The proofs of these results, as well as theorems for the broader class of *finite admixture* models are given in the technical report by one of the authors (Nguyen, 2012).

3. Empirical Study

The contraction theorems provide a rigorous characterization of the inference behavior of the LDA with the growth of data. We next validate these theoretical findings using extensive experiments on both synthetic and real-world data sets that present in different scenarios and with various configurations. We find that most empirical findings are consistent with the conclusions of the theorems.

The metric: What we use for the empirical evaluation is the minimum-matching Euclidean distance defined in (1) between the true and learned topic polytopes. It is well-known that when the number of vertices of a polytope in general positions is smaller than the number of dimensions, all such vertices are also the extreme points of their convex hull. Since this is precisely the common setting of topic modeling, the two quantities in (1) can be equivalently expressed as the following:

- Topic precision error (TPE):

$$d(G, G^*) = \max_{1 \le k \le K} \min_{1 \le j \le K^*} \|\phi_k - \phi_j^*\|_2 = \mathrm{TPE}(G, G^*),$$

- Topic recall error (TRE):

$$d(G^*, G) = \max_{1 \le j \le K^*} \min_{1 \le k \le K} \|\phi_j^* - \phi_k\|_2 = \mathrm{TRE}(G^*, G).$$

The TPE measures the largest deviation of the learned topics with respect to the true topics, while the TRE measures the largest deviation of the true topics from the collection of the estimated ones. They are both common information retrieval metrics applied to our topic learning problem.

3.1. Experiments on Synthetic Data

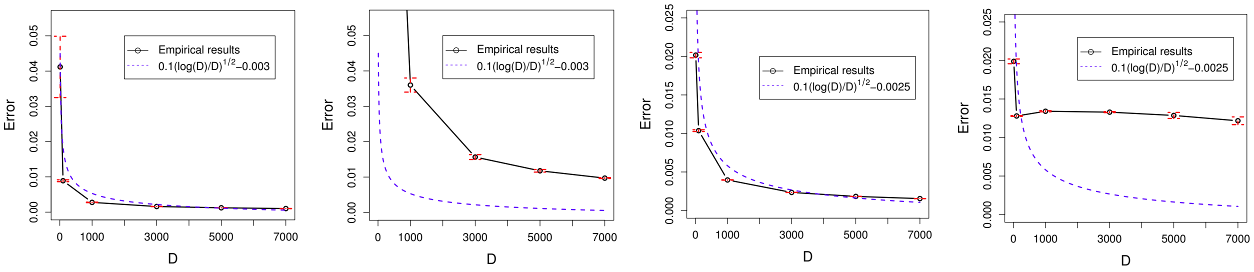
We first investigate the behavior of the LDA using synthetic data sets that are generated by the LDA generative process with different parameter configurations. We set the ground-truth number of topics K^* to be 3, the vocabulary size $|V|$ to be $5,000$, and the parameters of the symmetric Dirichlet priors for topic proportions and word distributions to be 1 and 0.01, respectively, unless otherwise specified. The same values of the hyperparameters are used to generate the data and to learn the topics.

We focus on the variation of the following parameters: the number of documents (D), the length of documents (N), the hyperparameter for the topic-word Dirichlet distribution (β), and the number of topics specified for inference (K). Collapsed Gibbs sampling (Griffiths & Steyvers, 2004) is used for model inference. The minimum-matching Euclidean distance metric, i.e. the maximum one between TPE and TRE metric, is used to evaluate the topic estimates. In particular we report the *learning error*, which is defined as the posterior mean of the aforementioned metric. All the results reported are averaged over 30 simulations in which the true topics and samples are all randomly generated according to the LDA model. The 95% confidence intervals are also plotted.

3.1.1. SCENARIO I: FIXED N AND INCREASING D

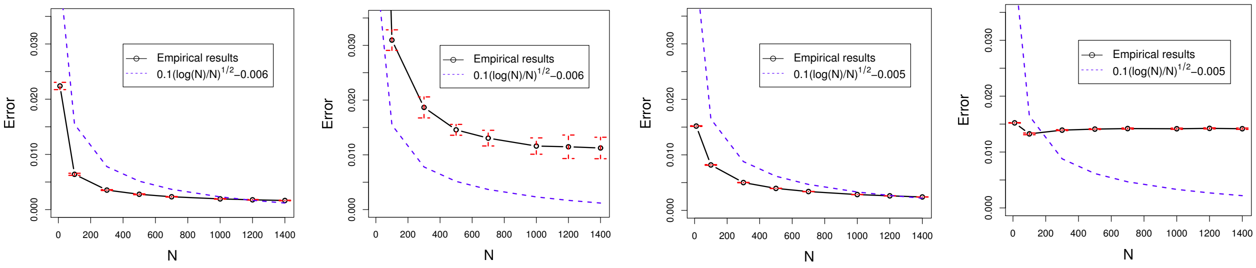
In the first scenario, we fix the length of documents $N = 500$, and let the number of documents increase from 10 to 7,000. We consider two different values of K (corresponding to the exact-fitted ($K = K^* = 3$) and the overfitted ($K = 10, K^* = 3$) scenarios, respectively), two different values of β (corresponding to well-separated topics ($\beta = 0.01$) and more word-diffuse, less distinguishable topics ($\beta = 1$)). The errors of the topics learned are reported in Fig. 1. In this scenario, the main varying term of the theoretical upper bound (5) is $\left(\frac{\log D}{D}\right)^{\frac{1}{2}}$, which we compare the empirical errors against. A few observations can be made:

1. Compare plot (a) with (b), and (c) with (d) (same β but different K), it can be seen that when the LDA is overfitted, the performance degenerates significantly. As predicted by Thm. 2, the error appears to either decrease very



(a) $K = K^*, \beta = 0.01$ (b) $K > K^*, \beta = 0.01$ (c) $K = K^*, \beta = 1$ (d) $K > K^*, \beta = 1$

Figure 1. Scenario I - Fixed N and increasing D. In the exact-fitted case, the error convergence rate well matches the result of Theorem 1. The over-fitted case leads to a much worse rate.



(a) $K = K^*, \beta = 0.01$ (b) $K > K^*, \beta = 0.01$ (c) $K = K^*, \beta = 1$ (d) $K > K^*, \beta = 1$

Figure 2. Scenario II - Fixed D and increasing N. In the over-fitted case, the error fails to vanish.

slowly due to the $\frac{1}{2(K-1)}$ exponent, or gradually become flattened as the constant $\frac{\log N}{N}$ term in the bound becomes the bottleneck.

2. Compare plot (a) with (c), and (b) with (d) (same K but different β): when β is larger, the error curves decay faster when less data is available. A possible explanation is that the topics are clustered thus easier to identify on a coarser level. As more data becomes available, the error decreases more slowly due to the confusion between topics on a finer level of inference. Interestingly, the error rate flats out, even as the amount of data increases. By contrast, when β is small, topics are more word-sparse and distinguishable, resulting in a more efficient learning rate.

3. When the number of topics is exactly fitted, the error rate seems to match the function $\left(\frac{\log D}{D}\right)^{\frac{1}{2}}$ quite well. In the over-fitted case however, the empirical rate is slower. Later we will analyze the exact exponent of this rate.

3.1.2. Scenario II: Fixed D and increasing N

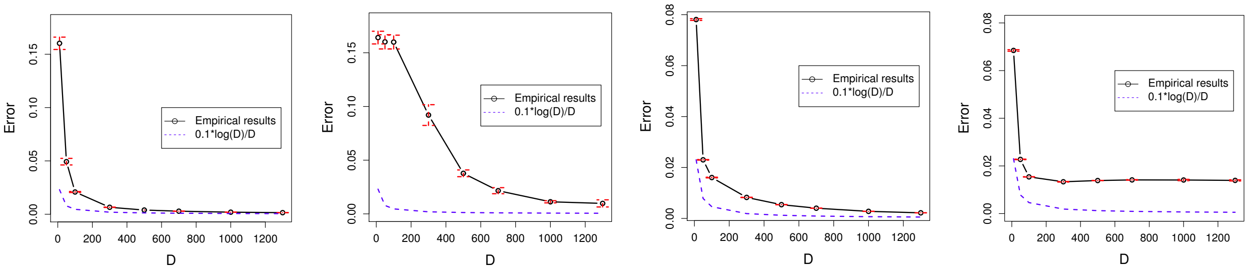
We next fix the number of documents D to $1,000$ and let the document length N range from 10 to $1,400$. We consider the exact-fitted ($K = K^* = 3$) and the over-fitted case ($K = 5$), using the setting $\beta = 0.01$ (word-sparse topics) and $\beta = 1$ (word-diffuse topics). The errors of the topics learned by the LDA are reported in Fig. 2, in which we compare against the varying term $\left(\frac{\log N}{N}\right)^{\frac{1}{2}}$. The behavior of the LDA in this scenario is very similar to Scenario I, as predicted by both theorems. In particular, in the over-fitted case, the error fails to vanish even as N becomes large, possibly due to the presence of the constant term $\frac{\log D}{D}$ in the upper bound.

3.1.3. Scenario III: $N = D$, both increasing

We next apply the LDA to synthetic data generated with $D = N$ that is allowed to increase simultaneously from 10 to $1,300$. As before, both the exact-fitted ($K = K^* = 3$) and an over-fitted case ($K = 5$) with $\beta = \{0.01, 1\}$ are considered. The results are reported in Fig. 3, where the error is plotted against $\frac{\log N}{N} = \frac{\log D}{D}$. Several observations can be made:

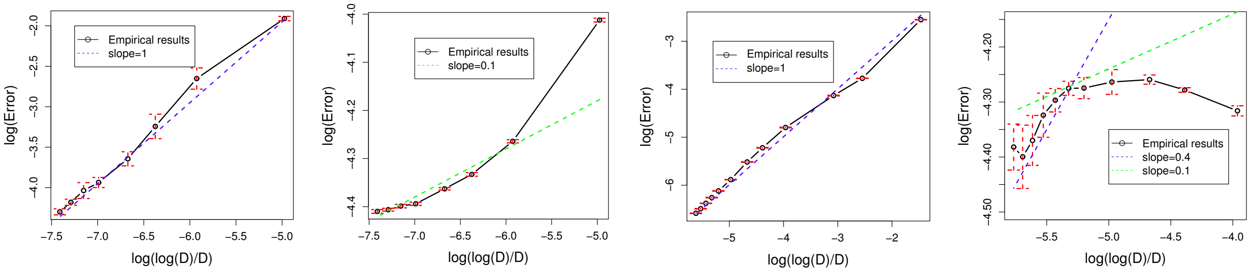
1. As in previous scenarios, the LDA is effective in the exact-fitted setting and when the word-distributions of the topics are sparse (i.e., β is small). When both of these conditions fail, the error rate fails to converge to zero, even as the data sizes $D = N$ increases (see Fig. 3).

2. Interestingly, when both D and N increase simultaneously, the empirical error decays at a faster rate than indicated by the upper bound $\left(\frac{\log D}{D}\right)^{\frac{1}{2}}$ from Thm. 1. As in the subsequent plots, a rough estimate could be $\Omega(1/D)$, which actually matches the theoretical *lower* bound of the error contraction rate (cf. Thm. 3 in Nguyen (2012)). This suggests that the upper bound given in Thm. 1 could be



(a) $K = K^*, \beta = 0.01$ (b) $K > K^*, \beta = 0.01$ (c) $K = K^*, \beta = 1$ (d) $K > K^*, \beta = 1$

Figure 3. Scenario III - $D = N$. In the exactly-fitted case the error concentrates faster than the upper bound in Theorem 1. The over-fitted case leads extremely slow error rates.



(a) Fixed N, $K = K^*$ (b) Fixed N, $K > K^*$ (c) $D = N$, $K = K^*$ (d) $D = N$, $K > K^*$

Figure 4. Log-plot of the topic error rate against $\log(\log(D)/D)$ when D increases. Note that larger D corresponds to the *left* side of the plot. We also plot straight lines with slopes indicating either the upper or the lower bound predicted by the theorems. ($K^* = 3, K = 5$)

quite conservative in certain configurations and scenarios.

3.1.4. Exponential exponents of the error rate

To precisely verify the theoretical bounds provided by the theorems, we study the exact exponent of the empirical error rate of the LDA and present the results in Figure 4.

We consider two scenarios: (1) fixed $N = 5$ and increasing D; and (2) $D = N$ and both increasing. When the LDA is exact-fitted (Fig. 4(a) and 4(c)), the slopes of the logarithm of the error appear to be very close to 1, which matches that of the lower bound $\Omega(1/D)$ mentioned above. On the other hand, in the over-fitted setting (Fig. 4(b) and 4(d)), the slopes of the logarithm of the error rate appear to tend toward the range bounded by $\frac{1}{2 \cdot K} = 0.1$ and $\frac{2}{K} = 0.4$. These are approximations of the exponents of lower/upper bounds predicted by the theory, respectively.

3.2. Experiments on Real Data Sets

The results on synthetic data sets verified the soundness of the theory. In this section, we present some empirical results when applying the LDA model to the following widely studied real-world data sets: (1) Wikipedia articles, (2) news articles from the New York Times (NYT) [2], and (3) short messages from Twitter.com. In all three collections, stop words and infrequent words (i.e., words appearing less than 100 times in the Twitter data set or the Wikipedia data set, and 50 times in the New York Times data set) are excluded. For the Twitter data, we randomly sample 10,000 users who have at least 100 tweets collected by the Twitter API (roughly 10% of all tweets) between Jan. 14th and Jan. 20th of 2013. The statistics of the three data sets are reported in Table 1.

Our goal is to test the effects of the four limiting factors on the performance of the LDA, namely the document length (N), the number of documents (D), and the hyperparameters (α and β). In order to test the effect of document lengths on the LDA's performance, we concatenate all tweets posted by the same author as "pseudo-documents," which allows us to obtain longer documents. For the same purpose, Wikipedia articles that contain less than 50 words after preprocessing are excluded from the analysis. We then construct data sets with various values of N and D by randomly sampling words from each document or randomly sampling documents from the corpus. All results are averaged over 10 randomized simulations. Hyper-parameters are set as $\alpha = 1$ and $\beta = 0.01$ unless otherwise specified.

As the ground-truth topics in real-world data are not available, it is necessary to introduce an alternative metric for evaluation purposes. We adopt the widely used average

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

Table 1. Statistics of the Real Data Sets

Data Set	# documents	# training documents	# average document length	vocabulary size (training)
Wikipedia	2,395,616	10,000	300.7	109,611
NYT	299,752	10,000	313.8	52,847
Twitter	81,553	10,000	417.3	122,035

For each data set, the entire set of documents are used for calculating PMI.

point-wise mutual information (PMI) to measure the quality of the learned topics (Newman et al., 2011). We have also done the evaluation using perplexity on held-out data. The findings are consistent with those observed with PMI. Due to the space limitation, we omit these results.

Fig. 5 reports the empirical performance of the LDA on the above real-world data sets via 10-fold cross-validations. Rows in Fig. 5 correspond to the three data sets, and columns correspond to different parameter configurations.

The results are consistent with the theory and the empirical analysis on synthetic data. When the data presents extreme properties (e.g., very short or very few documents) or when the hyperparameters are not appropriately set, the LDA's performance suffers. A turning point and a diminished return of the performance can be observed when increasing N or D, suggesting favorable ranges of values of the parameters. A benign range of the hyperparameters also can be observed, e.g., a small β and either a small α (Wikipedia) or a large α (NYT articles, Twitter) on different data sets.

4. Discussion

Related work. Our main focus is to understand the limiting factors of the LDA through studying the posterior distribution of the latent topic structure, as the amount of data increases. Our results are based on mild geometric assumptions and are independent of inference algorithms (however, see a discussion on the limitations below). We note some recent papers on the recoverability of the parameters of the LDA (Arora et al., 2012; Anandkumar et al., 2012). Both works rely on arguably stronger separability/sparsity conditions on the true topics and are specific to certain computationally efficient matrix factorization algorithms.

The performance of the LDA has also been studied empirically for various tasks and data sets using different evaluation measures (Newman et al., 2010; 2011). The conclusions are generally consistent with our findings. Meanwhile, Wallach et al. (2009) studied the role of asymmetric Dirichlet priors. Mukherjee & Blei (2009) presented an analysis of the difference between two inference algorithms, collapsed variational inference and mean field variational inference, and provided a relative performance guarantee for the approximate inference. These studies serve as a nice complement to our findings.

There are also interesting explorations of how to configure the data to improve the performance of LDA in reality. For example, Hong & Davison (2010) showed that LDA can obtain better topics when trained on "pseudo-documents" of aggregated tweets than on individual tweet. Mimno & McCallum (2007) proposed to break digital books into pages to obtain documents that are shorter and more concentrated on a few topics. These practices reconfirmed our findings about the limiting factors. Our work provides a theoretical justification and a thorough simulation based validation of these heuristic procedures.

Implications and guidelines for lay users of the LDA. We have presented the theory and supporting empirical study of the LDA model-based posterior inference. These findings are translated into the following guidance on the use of the LDA in practice:

(1) The number of documents plays perhaps the most important role; it is theoretically *impossible* to guarantee identification of topics from a small number of documents, no matter how long. Once there are sufficiently many documents, further increasing the number may not significantly improve the performance, unless the document length is also suitably increased. In practice, the LDA achieves comparable results even if thousands of documents are sampled from a much larger collection.
(2) The length of documents also plays a crucial role: poor performance of the LDA is expected when documents are too short, even if there is a very large number of them. Ideally, the documents need to be sufficiently long, but need not be too long: in practice, for very long documents, one can sample a fraction of each document and the LDA still yields comparable topics.
(3) When a very large number of topics than needed are used to fit the LDA, the statistical inference may become inescapably inefficient. In theory, the convergence rate deteriorates quickly to a nonparametric rate, depending on the number of topics used to fit the LDA. This implies, in practice, the user needs to exercise extra caution to avoid selecting overly large number of topics for the model.
(4) The LDA performs well when the underlying topics are well-separated in the sense of Euclidean metric; e.g., this is the case if the topics are concentrated at a small number of words. Another favorable scenario is concerned with the distribution of documents within the topic polytope: when individual documents are associated mostly with small subsets of topics, so that they are geometrically concentrated

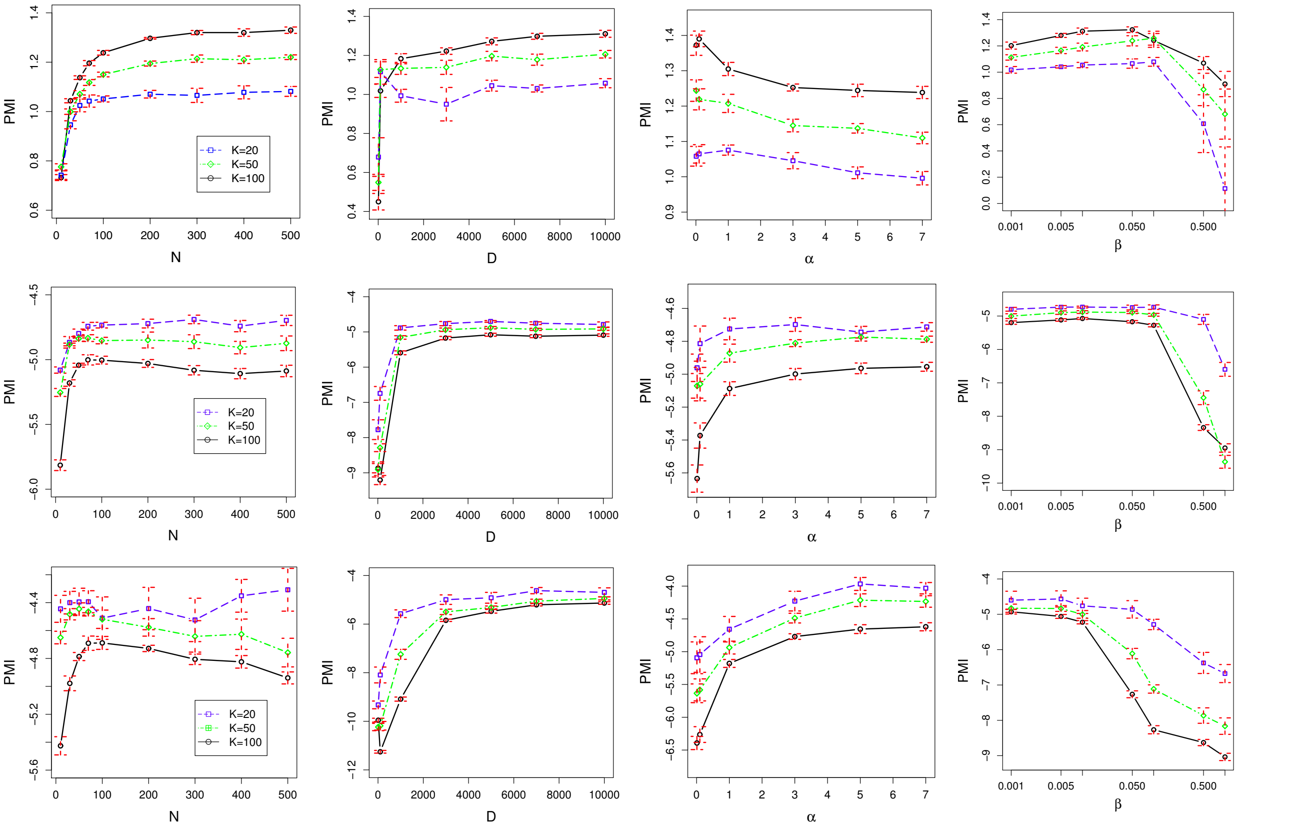


Figure 5. Empirical performance of the LDA on real-world data sets evaluated by averaged PMI. Rows correspond to different data sets, from top to bottom: (1) Wikipedia, (2) New York Times, and (3) Twitter. Columns correspond to different parameter configurations, from left to right: (a) fixing D and increasing N, (b) fixing N and increasing D, (c) fixing N and D and varying α, and (d) varying β.

mostly near the boundary of the topic polytope.

(5) If it is believed that each document is associated with few topics, the Dirichlet parameter of the document-topic distributions should be set small (e.g. $\alpha \approx 0.1$). If the topics are known to be word-sparse, the Dirichlet parameter of the word distributions β is set small (e.g., 0.01), in which case learning is efficient. Large β means more word-diffuse and similar topics, which might be inefficient to learn.

Limitations of existing results. There are several limitations of our results, particularly as we try to establish an asymptotic theory and link them to the empirical behaviors of the LDA model in practice:

(1) As the first attempt to address the posterior contraction behavior of the LDA model, we required several geometrically intuitive assumptions, some of which may not be commonly enforced in practice. For instance, in reality we do not know how separated the topics are, and whether their convex hull is geometrically degenerate or not. These unfavorable conditions, if true, would probably lead to worse learning rates than what is established by the current theory. This suggests that it may be beneficial to impose additional geometric constraints on the prior distribution in a computationally efficient way to prevent degenerate situations.

(2) Our theoretical analysis focuses on the behavior of the *true* posterior distribution of the latent topic variables. In practice, however, the posterior is obtained via approximation techniques (e.g., MCMC sampling). Therefore our empirical results inevitably contain such approximation error, which is dependent on the choice of the inference algorithm and its settings. Although there is currently no mature theory or practically simple way of jointly assessing both the computational complexity and statistical efficiency in a latent variable model such as the LDA, this is an interesting and important question worth exploring further.

Acknowledgements

This work is partially supported by the NSF Grants IIS-0968489, IIS-1054199, CCF-1048168 (QM), CCF-1115769 and OCI-1047871 (XN). It is also supported in part by the National Natural Science Foundation of China (NSFC Grant No. 61272343) and Doctoral Fund of Ministry of Education of China (MOEC RFDP Grant No. 20130001110032).

References

Anandkumar, Animashree, Foster, Dean P, Hsu, Daniel, Kakade, Sham M, and Liu, Yi-Kai. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *arXiv preprint arXiv:1204.6703*, 2012.

Arora, Saneev, Ge, Rong, and Moitra, Ankur. Learning topic models—going beyond SVD. *arXiv preprint arXiv:1204.1956*, 2012.

Bing, Lidong, Lam, Wai, and Wong, Tak-Lam. Using query log and social tagging to refine queries based on latent topics. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 583–592. ACM, 2011.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

Carman, Mark J, Crestani, Fabio, Harvey, Morgan, and Baillie, Mark. Towards query log based personalization using topic models. In *CIKM*, pp. 1849–1852. ACM, 2010.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

Grimmer, Justin. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Hong, Liangjie and Davison, Brian D. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 80–88, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0217-3.

Ling, Xu, Jiang, Jing, He, Xin, Mei, Qiaozhu, Zhai, Chengxiang, and Schatz, Bruce. Generating gene summaries from biomedical literature: A study of semi-structured summarization. *Information Processing & Management*, 43(6):1777–1791, 2007.

Liu, Yan, Niculescu-Mizil, Alexandru, and Gryc, Wojciech. Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 665–672. ACM, 2009.

McCallum, A, Mann, GS, and Mimno, D. Bibliometric impact measures leveraging topic analysis. In *JCDL*, pp. 65–74. IEEE, 2006.

Mimno, David. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(1):3, 2012.

Mimno, David and McCallum, Andrew. Organizing the oca: learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 376–385. ACM, 2007.

Mukherjee, Indraneel and Blei, David M. Relative performance guarantees for approximate inference in latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 21:1129–1136, 2009.

Newman, David, Chemudugunta, Chaitanya, Smyth, Padhraic, and Steyvers, Mark. Analyzing entities and topics in news articles using statistical topic models. *Intelligence and Security Informatics*, pp. 93–104, 2006.

Newman, David, Noh, Youn, Talley, Edmund, Karimi, Sarvnaz, and Baldwin, Timothy. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pp. 215–224. ACM, 2010.

Newman, David, Bonilla, Edwin V., and Buntine, Wray L. Improving topic coherence with regularized topic models. In *NIPS*, pp. 496–504, 2011.

Nguyen, XuanLong. Posterior contraction of the population polytope in finite admixture models. *arXiv preprint arXiv:1206.0068*, 2012.

Pritchard, Jonathan K, Stephens, Matthew, and Donnelly, Peter. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Ramage, Daniel, Rosen, Evan, Chuang, Jason, Manning, Christopher D, and McFarland, Daniel A. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, 2009.

Ramage, Daniel, Dumais, Susan, and Liebling, Dan. Characterizing microblogs with topic models. In *ICWSM*, 2010.

Wallach, Hanna M, Mimno, David M, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *NIPS*, volume 22, pp. 1973–1981, 2009.

Wang, Chong and Blei, David M. Collaborative topic modeling for recommending scientific articles. In *KDD*, pp. 448–456, 2011.

Zhao, Wayne, Jiang, Jing, Weng, Jianshu, He, Jing, Lim, Ee-Peng, Yan, Hongfei, and Li, Xiaoming. Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval*, pp. 338–349, 2011.