Complexity of Inference in Topic Models

David Sontag, Daniel M. Roy
Massachusetts Institute of Technology
{dsontag,droy}@csail.mit.edu

Abstract

We consider the computational complexity of finding the MAP assignment of topics to words in Latent Dirichlet Allocation. We show that, when the effective number of topics per document is small, exact inference takes polynomial time. In contrast, we show that, when a document has a large number of topics, finding the MAP assignment in LDA is NP-hard. Our results motivate further study of the structure in real-world topic models, and raise a number of questions about the requirements for accurate inference during both learning and test-time use of topic models.

1 Introduction

Probabilistic models of text and topics, known as topic models, are powerful tools for exploring large data sets and for making inferences about the content of documents. Topic models are frequently used for deriving low-dimensional representations of documents that are then used for information retrieval, document summarization, and classification [Blei & McAuliffe, 2008; Lacoste-Julien *et al.*, 2009]. Almost all uses of topic models require inference. For example, unsupervised learning of topic models using Expectation Maximization requires the repeated computation of marginal probabilities of what topics are present in the documents. For applications in information retrieval and classification, each new document necessitates inference to determine what topics are present.

Although there is a wealth of literature on approximate inference algorithms for topic models, such Gibbs sampling and variational inference [Blei *et al.*, 2003; Griffiths & Steyvers, 2004; Mukherjee & Blei, 2009; Porteous *et al.*, 2008; Teh *et al.*, 2007], little is known about the complexity of exact inference. In this paper, we consider the computational complexity of inference in topic models, beginning with one of the simplest and most popular models, Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003]. We chose to study LDA because we believe that it captures the essence of what makes inference easy or hard in topic models. Our hope is that our results will motivate discussion of the following questions, guiding research of both new topic models and approximate inference for topic models:

1. **What is the structure of real-world LDA inference problems?**
   Might there be structure in "natural" problem instances that makes them different from hard instances (e.g., those used in our reductions)?
2. **How much does having accurate inference affect the results of learning?**
   With a large training set or sufficiently long documents, might there be enough "averaging" for learning to succeed even with somewhat inaccurate inference?
3. **What are the requirements of applications that use test-time inference?**
   How accurate does test-time inference need to be? What quantities are needed (e.g., marginals, likelihood, most likely assignment)?

2 MAP inference

We will consider the inference problem for a single document. The LDA model states that the document, represented as a collection of words $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, is generated as follows: a distribution over the T topics is sampled from a Dirichlet distribution, $\theta \sim \mathrm{Dir}(\alpha)$; then, for $i = 1, \ldots, N$, we sample a topic $z_i \sim \mathrm{Multinomial}(\theta)$ and word $w_i \sim \Pr(w|z_i)$. Assume that these word distributions have been previously estimated, and denote $l_{it} = \log \Pr(w_i|z_i = t)$ as the log probability of the ith word being generated from topic t. After integrating out the topic distribution vector, the joint distribution of the topic assignments is given by

$$\Pr(z_1, \ldots, z_N) = \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(n_t + \alpha_t)}{\Gamma(\sum_t \alpha_t + N)} \prod_{i=1}^{N} \Pr(w_i|z_i), \tag{1}$$

where n_t is the total number of words assigned to topic t.

In this paper, we will focus on the inference problem of finding the most likely assignment of topics to words, i.e. the maximum a posteriori (MAP) assignment. Taking the logarithm of Eq. 1 and ignoring constants, finding the MAP assignment is seen to be equivalent to the following combinatorial optimization problem:

$$\begin{aligned} \Phi = \max_{x_{it} \in \{0,1\}, n_t} \quad & \sum_t \lg \Gamma(n_t + \alpha_t) + \sum_{i,t} x_{it} l_{it} \\ \text{subject to} \quad & \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t, \end{aligned} \tag{2}$$

where the indicator variable $x_{it} = \mathbb{I}[z_i = t]$ denotes the assignment of word i to topic t.

2.1 Exact maximization for small number of topics

Suppose a document only uses $\tau \ll T$ topics. That is, T could be large, but we are guaranteed that the MAP assignment for a document uses at most τ different topics. In this section, we show how we can use this knowledge to efficiently find a maximizing assignment of words to topics.

We first observe that, if we knew the *number* of words assigned to each topic, finding the MAP assignment is easy. For $i \in \{1, \ldots, T\}$, let n_i^* be the number of words assigned to topic i in the MAP assignment. Then, the MAP assignment $\vec{x}$ is found by solving the following optimization problem:

$$\begin{aligned} \max_{x_{it} \in \{0,1\}} \quad & \sum_{i,t} x_{it} l_{it} \\ \text{subject to} \quad & \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t^*, \end{aligned} \tag{3}$$

which is equivalent to weighted b-matching in a bipartite graph (the words are on one side, the topics on the other) and can be optimally solved in time $O(bm^3)$, where $b = \max_t n_t^* = O(N)$ and $m = N + T$ [Schrijver, 2003].

We call $(n_1, \ldots, n_T)$ a *valid partition* when $n_i \geq 0$ and $\sum_t n_t = N$. Using weighted b-matching, we can find a MAP assignment of words to topics by trying all $\binom{T}{\tau} = \Theta(T^\tau)$ choices of τ topics and all possible valid partitions with at most τ non-zeros.

for all subsets $A \subseteq \{1, 2, \ldots, T\}$ such that $|A| = \tau$ **do**
  for all valid partitions $\vec{n} = (n_1, n_2, \ldots, n_T)$ such that $n_t = 0$ for $t \notin A$ **do**
    $\Phi_{A,\vec{n}} \leftarrow$ WEIGHTED-B-MATCHING$(A, \vec{n}, l) + \sum_t \lg \Gamma(n_t + \alpha_t)$
  end for
end for
return $\arg\max_{A,\vec{n}} \Phi_{A,\vec{n}}$

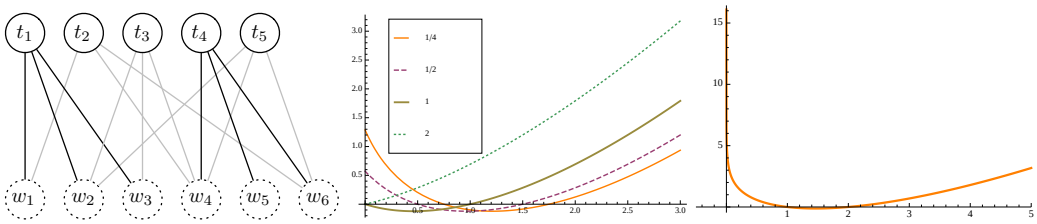

Figure 1: **(Left)** A LDA instance derived from a k-set packing instance. **(Center)** Plot of $F(n_t) = \lg \Gamma(n_t + \alpha)$ for various values of α. The x-axis varies n_t, the number of words assigned to topic t, and the y-axis shows $F(n_t)$. **(Right)** Behavior of $\lg \Gamma(n_t + \alpha)$ as $\alpha \to 0$. The function is stable everywhere but at zero, where the reward for sparsity increases without bound.

There are at most N^τ valid partitions with τ non-zero counts. For each of these, we solve the b-matching problem to find the most likely assignment of words to topics that satisfies the cardinality constraints. Thus, the total running time is $O((NT)^\tau N(N+\tau)^3)$. This is tractable when the number of topics τ appearing in a document is a constant.

2.2 Inference is NP-hard for large numbers of topics

In this section, we show that probabilistic inference is NP-hard in the general setting where a document may have a large number of topics in its MAP assignment. Let MAX-LDA(α) denote the decision problem of whether $\Phi > V$ (see Eq. 2) for some $V \in \mathbb{R}$, where the hyperparameters $\alpha_t = \alpha$ for all topics. We consider both $\alpha < 1$ and $\alpha \geq 1$ because, as shown in Figure 1, the optimization problem is qualitatively different in these two cases.

Theorem 1. MAX-LDA(α) *is NP-hard for all* $\alpha > 0$.

Proof. Our proof is a straightforward generalization of the approach used by Halperin & Karp [2005] to show that the minimum entropy set cover problem is hard to approximate.

The proof is done by reduction from k-set packing (k-SP), for $k \geq 3$. In k-SP, we are given a collection of k-element sets over some universe of elements Σ with $|\Sigma| = n$. The goal is to find the largest collection of *disjoint* sets. There exists a constant $c > 1$ such that it is NP-hard to decide whether a k-SP instance has (i) a solution with n/k disjoint sets covering all elements (called a *perfect matching*), or (ii) at most cn/k disjoint sets (called a (cn/k)-matching).

We now describe how to construct a LDA inference problem from a k-SP instance. This requires specifying the words in the document, the number of topics, and the word log probabilities l_{it}. Let each element $i \in \Sigma$ correspond to a word w_i, and let each set correspond to one topic. The document consists of all of the words (i.e., Σ). We assign uniform probability to the words in each topic, so that $\Pr(w_i | z_i = t) = \frac{1}{k}$ for $i \in t$, and 0 otherwise. Figure 1 illustrates the resulting LDA model. The topics are on the top, and the words from the document are on the bottom. An edge is drawn between a topic (set) and a word (element) if the corresponding set contains that element.

What remains is to show that we can solve some k-SP problem by using this reduction and solving a MAX-LDA(α) problem. For technical reasons involving $\alpha > 1$, we require that k is sufficiently large. We will use the following result, proved in the Appendix.

Lemma 2. *Let P be a k-SP instance for $k > (1+\alpha)^2$, and let P' be the derived* MAX-LDA*(α) instance. There exists constants C_U and $C_L < C_U$ such that, if there is a perfect matching in P, then $\Phi \geq C_U$. If, on the other hand, there is at most a (cn/k)-matching in P, then $\Phi < C_L$.*

Let P be a k-SP instance for $k > (3+\alpha)^2$, P' be the derived MAX-LDA(α) instance, and C_U and $C_L < C_U$ be as in Lemma 2. Then, by testing $\Phi < C_L$ and $\Phi > C_U$ we can decide whether P is a perfect matching or at best a (cn/k)-matching. Hence k-SP reduces to MAX-LDA(α). □

The bold lines in Figure 1 indicate the MAP assignment, which for this example corresponds to a perfect matching for the original k-set packing instance. More realistic documents would have significantly more words than topics used. Although this is not possible while keeping $k = 3$, since the MAP assignment always has $\tau \geq N/k$, we can instead reduce from a k-set packing problem with $k \gg 3$. Lemma 2 shows that this is hard as well.

3 Conclusion

In this paper, we have shown that the complexity of inference in LDA strongly depends on the effective number of topics per document. When we can guarantee that a document is generated from a small number of topics (regardless of the number of topics in the model), MAX-LDA can be solved in polynomial time. On the other hand, if a document can use an arbitrary number of topics, MAX-LDA is NP-hard. The choice of hyperparameters for the Dirichlet does not affect our results.

It would be interesting to show analogous results for computing marginals and the partition function. It is straightforward to extend both our positive and negative results to related models, such as probabilistic latent semantic analysis (PLSA) [Hofmann, 1999] or correlated topic models [Blei & Lafferty, 2006].

Appendix: Proof of Lemma 2

Proof of Lemma 2. Assume there are T sets each having $k \geq 3$ elements, and let Φ be the optimal LDA objective. Define $F(n) = \log \Gamma(n + \alpha)$. Since l_{it} is constant across all topics, the linear term in Eq. 2 will be a constant K. First, note that, if there is a perfect matching,

$$\Phi \geq \frac{n}{k} F(k) + (T - \frac{n}{k}) F(0) + K. \tag{4}$$

The $F(0)$ term is the contribution of unused topics. Otherwise, assume that the best packing has $\gamma \leq cn/k$ sets, each with k elements. Then, by the properties of the log-gamma function,

$$\Phi \leq \gamma F(k) + \frac{n - \gamma k}{k - 1} F(k - 1) + (T - \frac{n}{k}) F(0) + K, \tag{5}$$

where we assume, conservatively, that all of the remaining words are explained by topics assigned $(k-1)$ words. Also, since there was no perfect matching, there were at most $T - \frac{n}{k}$ unused topics. Using our bound on γ, we have

$$\begin{aligned}
\Phi &\leq \frac{cn}{k} F(k) + \frac{n - \frac{cn}{k} k}{k - 1} F(k - 1) && + (T - \frac{n}{k}) F(0) + K && (6)\\
&= \frac{cn}{k} F(k) + \frac{n(1 - c)}{k - 1} F(k - 1) && + (T - \frac{n}{k}) F(0) + K && (7)\\
&= \frac{dn}{k} F(k) && + (T - \frac{n}{k}) F(0) + K, && (8)
\end{aligned}$$

where

$$d := c + (1 - c)\beta, \qquad \text{for } \beta := \frac{k}{F(k)} \frac{F(k - 1)}{k - 1}. \tag{9}$$

Note that $F(k)/k \to \infty$ as $k \to \infty$. Along with the convexity of F, it follows that there exists a k_0 such that $\beta < 1$ for all $k > k_0$. Note that $k > (3 + \alpha)^2$ suffices. This implies that $d < 1$, which shows that there is a non-zero gap between the possible values of Φ. We have that $\frac{k}{k-1} \downarrow 1$ as $k \to \infty$. Therefore, $\beta < 1$ for some k if and only if the slope of F exceeds one at some point. □

Note that the maximum concentration objective, $F(n) = n \log n$, satisfies the conditions on F and, in particular, we have $\beta < 1$ for $k = 3$.

References

Blei, David, & Lafferty, John. 2006. Correlated Topic Models. *Pages 147–154 of:* Weiss, Y., Schölkopf, B., & Platt, J. (eds), *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press.

Blei, David, & McAuliffe, Jon. 2008. Supervised Topic Models. *Pages 121–128 of:* Platt, J.C., Koller, D., Singer, Y., & Roweis, S. (eds), *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Griffiths, Thomas L., & Steyvers, Mark. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(Suppl 1), 5228–5235.

Halperin, Eran, & Karp, Richard M. 2005. The minimum-entropy set cover problem. *Theor. Comput. Sci.*, **348**(2), 240–250.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. *Pages 50–57 of: SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM.

Lacoste-Julien, Simon, Sha, Fei, & Jordan, Michael. 2009. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. *Pages 897–904 of:* Koller, D., Schuurmans, D., Bengio, Y., & Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*.

Mukherjee, Indraneel, & Blei, David M. 2009. Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation. *Pages 1129–1136 of:* Koller, D., Schuurmans, D., Bengio, Y., & Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*.

Porteous, Ian, Newman, David, Ihler, Alexander, Asuncion, Arthur, Smyth, Padhraic, & Welling, Max. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. *Pages 569–577 of: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM.

Schrijver, Alexander. 2003. *Combinatorial optimization. Polyhedra and efficiency. Vol. A*. Algorithms and Combinatorics, vol. 24. Berlin: Springer-Verlag. Paths, flows, matchings, Chapters 1–38.

Teh, Y. W., Newman, D., & Welling, M. 2007. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *In: Advances in Neural Information Processing Systems*, vol. 19.