Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

IMPORTANCE SAMPLING FOR LATENT DIRICHLET ALLOCATION

Ayeong Lee[1]

[1]Graduate School of Business, Columbia University, New York, NY, USA

ABSTRACT

Latent Dirichlet Allocation (LDA) is a method for finding topics in text data. Evaluating an LDA model entails estimating the expected likelihood of held-out documents. This is commonly done through Monte Carlo simulation, which is prone to high relative variance. We propose an importance sampling estimator for this problem and characterize the theoretical asymptotic statistical efficiency it achieves in large documents.

1 INTRODUCTION

Latent Dirichlet Allocation (LDA), introduced in Blei et al. (2003), is a popular unsupervised learning technique for extracting topics from a collection of documents, with applications in natural language processing, recommendation systems, and other domains.

An important task is to evaluate the quality of the topics extracted by LDA which is often done by considering the expected likelihood of on a set of held-out document. The expected likelihood is used not only to evaluate the fit of the LDA model, but also to perform model selection, e.g. determining the optimal number of topics.

Suppose the number of topics K and the set of vocabulary V are fixed. Let w be the set of words in a held-out document, and let n_v be the counts of each vocabulary element v. Given the word counts, we would like to evaluate the likelihood of observing these words w in the document under the topic model ϕ. To do so, we consider the expected likelihood of the document averaged over topic proportions θ drawn from the Dirichlet prior with parameter α,

$$p(w|\phi) = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}\left[p(w|\theta,\phi)\right] = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}\left[\prod_{v=1}^{V}\left(\sum_{k=1}^{K}\theta(k)\phi_k(v)\right)^{n_v}\right]. \tag{1}$$

Letting n denote the total number of words in the document and $p_v = n_v/n$ denote the frequency of words in the document, we can write (1) as

$$p(w|\phi) = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}[e^{nH(\theta)}], \tag{2}$$

where

$$H(\theta) = \sum_{v=1}^{V} p_v \log\left(\sum_{k=1}^{K}\theta(k)\phi_k(v)\right) \tag{3}$$

is a concave function. Given that there is no closed form for $p(w|\phi)$, a standard procedure is to estimate the quantity through Monte Carlo sampling: with $i = 1,...,N$ samples of $\theta_i \overset{iid}{\sim} \text{Dir}_\alpha$,

$$\hat{p}_{MC} = \frac{1}{N}\sum_{i=1}^{N} e^{nH(\theta_i)}. \tag{4}$$

However, due to the exponential estimand, when n is large, the variance of $\hat{p}_{MC}$ is high relative to the mean.

In this paper, we propose an alternative estimator based on importance sampling and the properties of Dirichlet distributions. We theoretically characterize the asymptotic statistical efficiency of the estimator in large documents, and show numerically that it substantially reduces variance over the standard estimator.

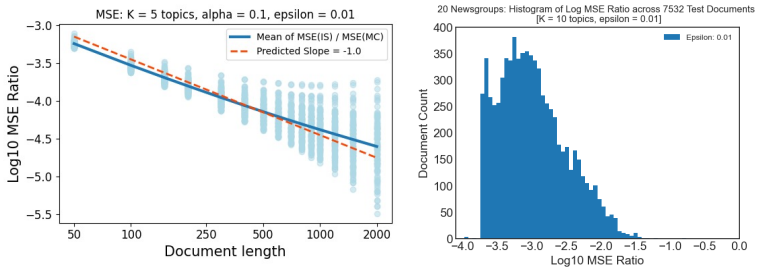

Figure 1: Left: MSE Ratio results for 100 instances of synthetic data with $K = 5$, $V = 1000$, and $\varepsilon = 0.01$. MSE ratio decays as predicted. Right: Histogram of $\log(\text{MSE}(\hat{p}_{\text{IS}})/\text{MSE}(\hat{p}_{\text{MC}}))$ across 7,532 test documents in the 20newsgroups dataset using $K = 10$ topics.

2 DIRICHLET IMPORTANCE SAMPLING ESTIMATOR

Suppose the maximizer θ^* of H is in the interior of the simplex Δ_{K-1}. Inspired from classical Laplace approximation, we propose to sample from $\theta_i \sim \text{Dir}_{\alpha+\sqrt{n}\theta^*}$ to produce the following importance sampling (IS) estimator:

$$\hat{p}_{IS} = \frac{1}{N}\sum_{i=1}^{N} e^{nH(\theta_i)} \frac{\text{Dir}_{\alpha}(\theta_i)}{\text{Dir}_{\alpha+\sqrt{n}\theta^*}(\theta_i)} \mathbf{1}_{\Delta^{\varepsilon}_{K-1}}, \tag{5}$$

where $\mathbf{1}_{\Delta^{\varepsilon}_{K-1}}$ is an indicator on the ε-truncated $(K-1)$ simplex defined as

$$\Delta^{\varepsilon}_{K-1} := \{\theta \in \Delta_{K-1} : \theta_i \geq \varepsilon\}. \tag{6}$$

The design of the estimator is motivated as follows. For large n, we expect most of the contribution to the expected likelihood (2) will come from values of θ near the maximizer θ^*. By sampling from a Dirichlet distribution with parameter $\alpha + \sqrt{n}\theta^*$, the estimator places more weight on these values as n grows large. However, if the proposal distribution concentrates too heavily on θ^* the likelihood ratio may diverge near the boundary; in fact, $\mathbb{E}[(\text{Dir}_{\alpha}(\theta)/\text{Dir}_{\gamma}(\theta))^2] = \infty$ for any γ where $\gamma > 2\alpha$. The $\sqrt{n}$ scaling of the Dirichlet parameter yields the benefits of targeted sampling while carefully controlling this behavior, and ε-truncation prevents divergence near the boundary.

3 CONCLUSION

Our main result proves an asymptotic expression for the efficiency of our importance sampling estimator as the document length n increases. As $n \to \infty$, ratio of the mean-squared error (MSE) of importance sampling compared to the standard Monte Carlo estimator decays at a polynomial rate:

$$\frac{\text{MSE}(\hat{p}_{\text{IS}})}{\text{MSE}(\hat{p}_{\text{MC}})} = \Theta(n^{-\frac{K-1}{4}}). \tag{7}$$

We provide numerical evidence of our theoretical result on a synthetic dataset and on a text corpus (20 newsgroup dataset). Figure 1 shows that the MSE ratio decays as predicted on simulated documents. For the real dataset, 96% of test documents show a log-MSE ratio less than -2, and for more than 53% of documents the ratio is less than -3.

ACKNOWLEDGMENTS

The author thanks her advisor Paul Glasserman for his comments and feedback.

REFERENCES

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3:993–1022.