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

Paul Glasserman[1], and 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. We illustrate the method in simulated data and in a dataset of news articles.

1 INTRODUCTION

Latent Dirichlet Allocation (LDA), introduced in Blei et al. (2003), is a powerful unsupervised learning technique for extracting topics from a collection of documents, with applications in natural language processing, recommendation systems, and other domains. It can be viewed as a method for dimension reduction for categorical data. An important task is to evaluate the quality of the topics extracted by LDA, which is challenging due to the unsupervised nature of the method. A standard approach considers the expected likelihood of held-out documents, averaged over topic distributions drawn from a Dirichlet prior. However, this typically involves high variance due to variability of the predictions of the LDA model under different topic distributions. Consequently, tasks such as model selection—e.g., choosing the number of topics—can require a large simulation budget to obtain reliable estimates.

Early work on LDA focused on two primary inference strategies: variational methods (Blei et al. 2003) and Markov Chain Monte Carlo (MCMC) techniques, notably the collapsed Gibbs sampling approach proposed in Griffiths and Steyvers (2004), which analytically marginalizes the topic distributions. For evaluating model fit and performing model selection, estimating the marginal likelihood of held-out data is critical. Wallach et al. (2009) consider the application of an importance sampling method to this problem, but they find that it can be unstable and have high variance in the context of topic models.

In this paper, we revisit importance sampling for model evaluation in LDA and propose a new estimator based on properties of Dirichlet distributions. We theoretically characterize the asymptotic statistical efficiency of the estimator in large documents. Our analysis highlights how the degree of efficiency depends on the number of topics, the topic vectors, and the optimal document-topic distribution. Empirically, we demonstrate that our importance sampling estimator substantially improves upon the mean squared error of the standard Monte Carlo estimator in a variety of examples. We also confirm that our theoretical analysis is predictive of the method's empirical effectiveness.

2 LATENT DIRICHLET ALLOCATION (LDA)

This section provides a brief description of LDA. We begin with some background on Dirichlet distributions.

2.1 Dirichlet Distributions

A Dirichlet distribution is supported on the set of probability vectors of length K, for some $K \geq 2$. Define Δ_{K-1} as the $(K-1)$-dimensional simplex in $\mathbb{R}^K$, i.e.,

$$\Delta_{K-1} = \left\{ x \in \mathbb{R}^K \,\middle|\, x_i \geq 0 \text{ for all } i = 1, \ldots, K, \ \sum_{i=1}^{K} x_i = 1 \right\}.$$

A Dirichlet distribution Dir_α has the probability density function

$$\mathrm{Dir}_\alpha(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad x \in \Delta_{K-1},$$

with parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$, with $\alpha_i > 0$ for all i. The normalizing constant $B(\alpha)$ is the multivariate Beta function given by

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}. \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function, defined for $x > 0$ as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt.$$

Dirichlet distributions are commonly used to model distributions over probability vectors. The Dirichlet density's shape changes with the parameters α_i: as the α_i increase, the distribution shifts from concentrating near the corners of the simplex (favoring sparse, peaked probability vectors) to concentrating near the center (favoring more uniform probability vectors). We will assume throughout that all Dirichlet vectors have strictly positive components, as this property holds with probability one.

2.2 Latent Dirichlet Allocation (LDA)

LDA posits a generative probabilistic model for a collection of documents based on a latent structure of topics. Bayesian inference is then used to infer the topics from observed text. The generative model assumes that each document is a mixture of latent topics, where each topic is characterized by a probability distribution over a fixed vocabulary of words; each document is then characterized by a probability distribution over topics.

For example, one topic may assign high probability to words like "economy," "inflation," and "market." Another topic may assign high probability to "patient," "treatment," and "symptom." An individual document about the financial performance of drug companies might assign 15% weight to the first topic, 10% weight to the second topic, with the remaining 75% weight distributed among other topics.

The generative process assumes that topic vectors are drawn from a Dirichlet distribution over a fixed vocabulary; topic proportions for each document are drawn from a Dirichlet distribution over the set of topics. Then, for each word in the document, a topic is sampled from the document's distribution over topics, and a word is drawn from the corresponding topic's distribution over the vocabulary. The details are as follows:

Suppose the number of topics K and the vocabulary size V are fixed.

1. **Topic generation:** K topic vectors are drawn from a Dirichlet distribution,

$$\phi_k \overset{\text{iid}}{\sim} \mathrm{Dir}_\beta(\cdot), \quad \text{for } k = 1, \ldots, K,$$

where $\mathrm{Dir}_\beta(\cdot)$ is a Dirichlet distribution with parameters $\beta \in \mathbb{R}_+^V$. Each ϕ_k is a probability distribution over the vocabulary.

2. **Document-level topic proportions:** For each document $j = 1,\ldots,J$, a topic proportion vector is drawn:
$$\theta_j \overset{\text{iid}}{\sim} \text{Dir}_\alpha(\cdot),$$
where $\theta_j \in \Delta_{K-1}$ and $\alpha \in \mathbb{R}_+^K$. Each θ_j is a probability distribution over topics.
3. **Word generation:** A document j is a sequence of N_j words (with $N_1,\ldots,N_J$ fixed), where each word takes a value from the vocabulary. Given the document-specific topic distribution θ_j, the ith word, $i = 1,\ldots,N_j$, is generated as follows:
   (a) Draw a topic assignment:
   $$z_{i,j} \sim \theta_j,$$
   (b) Draw a word from the selected topic:
   $$w_{i,j} \sim \phi_{z_{i,j}}.$$

The generative model determines the probability of any collection of documents, given the latent topic structure. The inverse problem of inferring the latent topic structure from a collection of documents is a problem of Bayesian inference. It is commonly solved using either variational inference or Gibbs sampling.

3 EVALUATION METRIC

We now discuss the evaluation metric used to assess the quality of the learned set of topic vectors ϕ. 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. To do so, we consider the expected likelihood of the document averaged over topic proportions θ drawn from the Dirichlet prior with parameter α,

$$p(w|\phi) = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}\left[p(w|\theta,\phi)\right] = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}\left[\prod_{v=1}^{V}\left(\sum_{k=1}^{K}\theta(k)\phi_k(v)\right)^{n_v}\right]. \tag{2}$$

In this expression, only θ is stochastic; the word counts n_v and topic vectors ϕ_k are fixed.

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 (2) as

$$p(w|\phi) = \mathbb{E}_{\theta\sim\text{Dir}_\alpha}[e^{nH(\theta)}], \tag{3}$$

where

$$H(\theta) = \sum_{v=1}^{V} p_v \log\left(\sum_{k=1}^{K}\theta(k)\phi_k(v)\right). \tag{4}$$

Here, H captures the log-likelihood under the topic mixture. Intuitively, H measures how well the topic-weighted word probabilities explain the empirical word frequencies p_v in the document. However, computing the expectation (3) exactly is intractable in general. 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 $\theta_i \overset{iid}{\sim} \text{Dir}_\alpha$,

$$\hat{p}_{MC} = \frac{1}{N}\sum_{i=1}^{N} e^{nH(\theta_i)}. \tag{5}$$

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

4 DIRICHLET IMPORTANCE SAMPLING ESTIMATOR

Importance sampling seeks to reduce variance by drawing more samples from "important" regions. We consider importance sampling strategies that replace the original Dirichlet distribution Dir_α in (3) with another Dirichlet distribution Dir_γ, for some parameter vector γ. This approach leads to the following representation of our quantity of interest:

$$\mathbb{E}_{\text{Dir}_\alpha}[e^{nH(\theta)}] = \mathbb{E}_{\text{Dir}_\gamma}\left[e^{nH(\theta)} \cdot \frac{\text{Dir}_\alpha}{\text{Dir}_\gamma}(\theta)\right] = \mathbb{E}_{\text{Dir}_\gamma}\left[e^{nH(\theta)} \cdot \frac{B(\gamma)}{B(\alpha)} \prod_{j=1}^{K} \theta_j^{\alpha_j - \gamma_j}\right]. \tag{6}$$

For any choice of γ, $\gamma_i > 0$, $i = 1, \ldots, K$, the expression inside the rightmost expectation in (6) provides an unbiased estimator when sampled under Dir_γ. The likelihood ratio $\text{Dir}_\alpha(\theta)/\text{Dir}_\gamma(\theta)$ corrects for the change of sampling distribution. (For general background on importance sampling, see, e.g., Asmussen and Glynn 2007 and Kroese et al. 2013.)

Ideally, we would like Dir_γ to place more weight on regions that contribute the most to the original integrand. For large n, we expect most of the contribution to (3) will come from values of θ near the maximizer θ^* of $H(\theta)$. This idea is supported by the asymptotic approximation we will use in the analysis of our estimator; see Appendix A.1. We therefore seek to concentrate sampling of θ near θ^*. At the same time, we need to be mindful that the likelihood ratio $\text{Dir}_\alpha(\theta)/\text{Dir}_\gamma(\theta)$ in (6) may diverge at the boundary of the simplex. In fact, $\mathbb{E}_{\text{Dir}_\gamma}\left[(\prod_{j=1}^{K} \theta_j^{\alpha_j - \gamma_j})^2\right]$ will be infinite if $2\alpha_j - \gamma_j < 0$.

To address both considerations, we propose to sample from $\theta_i \sim \text{Dir}_{\alpha + \sqrt{n}\theta^*}$ to produce the following importance sampling (IS) estimator:

$$\hat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} e^{nH(\theta_i)} \frac{\text{Dir}_\alpha(\theta_i)}{\text{Dir}_{\alpha+\sqrt{n}\theta^*}(\theta_i)} \mathbf{1}_{\Delta_{K-1}^\varepsilon}, \tag{7}$$

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

$$\Delta_{K-1}^\varepsilon := \{\theta \in \Delta_{K-1} : \theta_i \geq \varepsilon\}. \tag{8}$$

The restriction to Δ_{K-1}^ε keeps the likelihood ratio bounded. It introduces a small bias,

$$\text{Bias}(\hat{p}_{IS}) = \mathbb{E}_{\theta \sim \text{Dir}_\alpha}[e^{nH(\theta)}] - \mathbb{E}_{\theta \sim \text{Dir}_\alpha}[e^{nH(\theta)} \mathbf{1}_{\Delta_{K-1}^\varepsilon}] = \mathbb{E}_{\theta \sim \text{Dir}_\alpha}[e^{nH(\theta)} \mathbf{1}_{(\Delta_{K-1}^\varepsilon)^c}], \tag{9}$$

which is the expectation over the omitted part of the simplex and which we will show is negligible. The parameter $\varepsilon > 0$ can be arbitrarily small. We only require that the maximizer θ^* be in the strict interior of Δ_{K-1}^ε, which is assumption (A1) below. Our choice of importance sampling distribution concentrates sampling near the maximizer θ^* because

$$\text{Mode}(\text{Dir}_{\alpha+\sqrt{n}\theta^*}) = \frac{\alpha + \sqrt{n}\theta^* - \mathbf{1}}{\sum_{j=1}^{K}(\alpha_j + \sqrt{n}\theta_j^*) - K} \to \theta^*.$$

The $\sqrt{n}$ scaling allows us to apply an asymptotic approximation to evaluate the variance of $\hat{p}_{IS}$; see Appendix A.1.

5 ASYMPTOTIC MEAN SQUARED ERROR (MSE) REDUCTION

Our main result proves an asymptotic expression for the efficiency of our importance sampling estimator as the length of the held-out document increases. We first characterize its asymptotic bias relative to the variance of the naive Monte Carlo estimator. All proofs are in the appendix. We will use the following conditions:

(A1) The maximizing vector θ^* is in the strict interior of Δ_{K-1}^ε, i.e. $\theta_j^* > \varepsilon, j = 1, \ldots, K$.

(A2) The matrix $[\phi_k(v)]_{k=1,\ldots,K, v=1,\ldots,V}$ has rank $K < V$.

Lemma 1 Suppose conditions (A1) and (A2) hold. Then for some positive constant $\delta > 0$, we have

$$\frac{\text{Bias}(\hat{p}_{\text{IS}})^2}{\text{Var}(\hat{p}_{\text{MC}})} = O(e^{-2n\delta}). \tag{10}$$

Next, we analyze the variance reduction achieved by the IS estimator (7). Under the same conditions, we can show that the the variance of the IS estimator decays at a faster polynomial rate in comparison to the MC estimator (5).

Theorem 1 Suppose (A1) and (A2) hold. Then, as $n \to \infty$,

$$\frac{\text{Var}(\hat{p}_{\text{IS}})}{\text{Var}(\hat{p}_{\text{MC}})} \sim \; n^{-\frac{K-1}{4}} \cdot (2\pi)^{\frac{K-1}{2}} \cdot \text{Dir}_{\alpha}(\theta^*) \cdot \left(\prod_k \theta_k^* \right)^{1/2}. \tag{11}$$

Combining the exponential bias decay from (10) with the polynomial variance reduction from (11), we can quantify the asymptotic mean squared error improvement of our IS estimator over the MC estimator.

Corollary 1 Under conditions (A1) and (A2),

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

Remark. Condition (A2) is easily achieved as long as the vocabulary size V is much larger than K, which is always the case in practice. We show in the appendix that (A2) implies that the Hessian $\nabla^2 H$ is negative definite on Δ_{K-1}, a property we will use in our asymptotic analysis. We will also note that since the topic-word distributions $\{\phi_k\}_{k=1,\ldots,K}$ have strictly positive components, $\nabla^2 H$ is continuous.

In Corollary 1, the MSE reduction using IS is theoretically greater with a larger number of topics K. We provide some insight into this feature in the remark at the end of Appendix A.5.

6 NUMERICAL RESULTS

We run experiments to provide numerical evidence for the result in Corollary 1 and to demonstrate the MSE reduction of our proposed importance sampling estimator on a standard empirical dataset.

6.1 Synthetic Data Experiments

We run synthetic data experiments using an LDA model with a vocabulary of $V = 1000$ words and $K = 5$ topics. We fix $\alpha = 0.1 \cdot \mathbf{1}_K \in \mathbb{R}^K$ and $\beta = 0.1 \cdot \mathbf{1}_V \in \mathbb{R}^V$; small values of these parameters are commonly used in LDA because they produce "spiky" vectors.

For the evaluation of (2), a document is characterized by its length n and its word frequencies p_v. We generate 100 instances of p_v and vary n. To generate an instance of p_v, we draw $K = 5$ independent latent topic vectors from $\phi_k \sim \text{Dir}_{\beta}$ on Δ_V and a random topic distribution $\theta_{\text{true}} \sim \text{Dir}_{\tilde{\alpha}}$ on Δ_{K-1}. We then set $p_v = \sum_{k=1}^{K} \theta_{\text{true},k} \phi_k(v)$. This construction makes $\theta^* = \theta_{\text{true}}$; choosing $\tilde{\alpha} = \mathbf{1}_K$ helps ensure that θ^* will be in the interior of $\Delta_{K-1}^{\varepsilon}$. For each of the 100 instances generated this way, we vary n from 50 to 1000.

For the standard Monte Carlo estimator, we draw $N = 10^6$ samples of $\theta \sim \text{Dir}_{\alpha}$ and evaluate (5) for each instance of p_v and each n. For our proposed importance sampling estimator $\hat{p}_{\text{IS}}$, we draw $N = 10^6$ samples of $\theta \sim \text{Dir}_{\alpha+\sqrt{n}\theta^*}$ and evaluate (7) for each instance of p_v and each n. We set $\varepsilon = 0.01$ for the simplex truncation parameter. We observe that while the importance sampling estimator introduces the additional step of computing the maximizer θ^*, the wall-clock time of doing so with standard gradient-based methods was 100x faster (2.2 ms on a Intel(R) Xeon(R) CPU 2.20GHz) than sampling from the Dirichlet distribution with $N = 10^6$ samples (340 ms on the same machine), contributing only negligibly to the overall computational cost of the estimator.

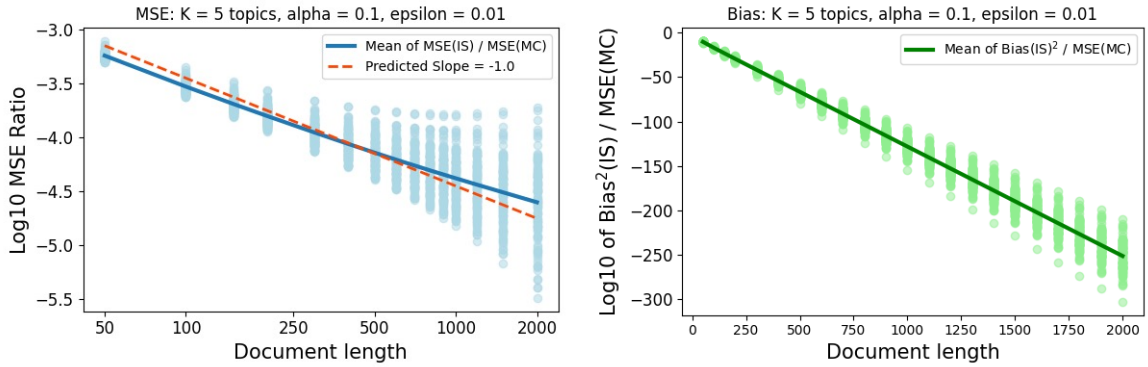

Figure 1: Results for 100 instances of synthetic data with $K = 5$, $V = 1000$, and $\varepsilon = 0.01$. Left: MSE ratio decays as predicted. Right: Bias ratio decays exponentially.

Because $\hat{p}_{\mathrm{IS}}$ is biased, we calculate both the MSE ratio and the relative bias normalized by the MSE of the Monte Carlo estimator (which is the same as its variance). These ratios are

$$\frac{\mathrm{MSE}(\hat{p}_{\mathrm{IS}})}{\mathrm{MSE}(\hat{p}_{\mathrm{MC}})} \quad \text{and} \quad \frac{\mathrm{Bias}^2(\hat{p}_{\mathrm{IS}})}{\mathrm{MSE}(\hat{p}_{\mathrm{MC}})}. \tag{12}$$

As the document length n increases, our asymptotic results in Corollary 1 predict that the MSE ratio should decay at rate $n^{-(K-1)/4} = n^{-1}$, and the relative bias should decay exponentially.

We calculate the ratios in (12) for each of the 100 instances, for each value of n. The results are shown in the left panel of Figure 1, where the blue line shows the average MSE ratio. The right panel shows corresponding results for the relative bias. Consistent with our theoretical predictions, the MSE ratio decays like n^{-1}, and the relative bias decays exponentially. The results indicate that our IS estimator indeed provides significant variance reduction and negligible bias compared with the standard MC estimator. Most of the variability in the figures comes from estimating the variance of the MC estimator.

In separate experiments we have confirmed that we do not get similar performance with $\varepsilon = 0$: the restriction to some $\Delta^{\varepsilon}_{K-1}$, $\varepsilon > 0$, in our IS estimator is important. We need ε to be small enough to satisfy (A1), and we generally observe better results with smaller ε, particularly with larger values of K.

6.2 Newsgroups Data

We next evaluate our importance sampling estimator on the "20 newsgroups" dataset, a widely used text corpus, often attributed to Lang (1995). The dataset contains approximately 18000 newsgroup posts, split into a training set and a test set. The vocabulary size is roughly 10^5 for the entire corpus. For our model, we choose $K = 10$ topics and use variational inference to extract the topic-word distributions ϕ_k from approximately $11,000$ training documents. Then, for each of approximately $7,500$ test documents we calculate the empirical frequencies p_v and evaluate the MSE ratio between the standard Monte Carlo estimator and our proposed importance sampling estimator. With $\alpha = 0.1 \cdot \mathbf{1}_K$, we draw $N = 10^5$ samples from Dir_{α} and $\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}$ for the MC and IS estimators, respectively. For IS we set the truncation parameter to $\varepsilon = 0.01$. The maximizer θ^* for each document can be calculated quickly through a simple recursion.

This setting does not strictly fall within the scope of our theoretical analysis because here we cannot vary n while holding p_v fixed; we simply have documents of different lengths. Also, we cannot guarantee (A1). We can nevertheless check for error reduction through the MSE ratio.

Figure 2 plots the distribution of log-MSE ratios across the test documents. The results show consistent error reduction across all documents. Furthermore, 96% of test documents show a log-MSE ratio less than -2, and for more than 53% the ratio is less then -3.

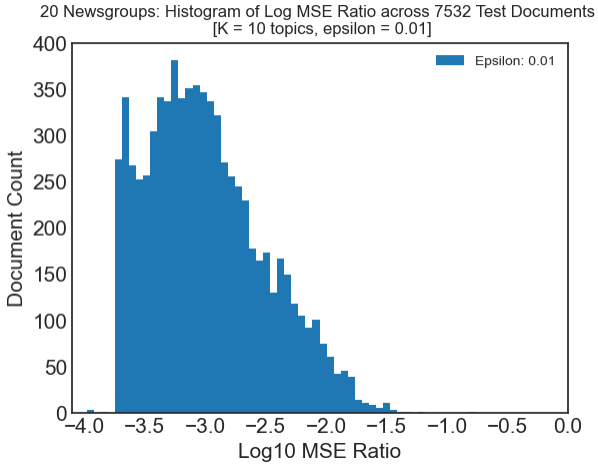

Figure 2: Histogram of $\log(\mathrm{MSE}(\hat{p}_{\mathrm{IS}})/\mathrm{MSE}(\hat{p}_{\mathrm{MC}}))$ across 7,532 test documents using $K = 10$ topics.

7 CONCLUSION

We proposed and analyzed an importance sampling estimator for the evaluation of LDA topic models on test documents. Compared to standard Monte Carlo, the estimator's relative efficiency grows with document length. Future work includes exploring cases in which θ^* lies on the boundary of the simplex, testing larger models, and analyzing the bias–variance trade-off for different restriction thresholds ε.

A APPENDIX

A.1 Laplace Approximations

The classical Laplace approximation (see, e.g., Breitung 1994, Theorem 41, p.56) takes the form

$$\int_F h(\mathbf{x})\exp(nf(\mathbf{x}))d\mathbf{x} \sim (2\pi)^{k/2}\frac{h(\mathbf{x}^*)}{\sqrt{|\det(H_f(\mathbf{x}^*)|}}\cdot\exp(nf(\mathbf{x}^*))\cdot n^{-k/2}, \quad \text{as } n\to\infty, \tag{13}$$

where the integration domain $F \subset \mathbb{R}^k$ is a compact set and the maximizer $\mathbf{x}^*$ of f is an interior point of F. Here, $h(\cdot)$ is a continuous function assumed to be neither vanishing nor singular at $\mathbf{x}^*$, and the Hessian of $f(\cdot)$, $H_f(\mathbf{x}^*)$ must be negative definite. The key point is that the integrand has an exponential dependence on a large parameter n, as in our setting. The approximation is based on the idea that the integral is dominated by the behavior of the integrand near the maximum of f. We will apply (13) to expectations like $\mathbb{E}_{\theta\sim\mathrm{Dir}_\alpha}[e^{nH(\theta)}]$. In fact, under mild conditions, it suffices to restrict to smaller neighborhoods V of $\mathbf{x}^*$, instead of the entire F (Breitung 1994, Lemma 38, p.53):

$$\int_F h(\mathbf{x})\exp(nf(\mathbf{x}))d\mathbf{x} \sim \int_{F\cap V} h(\mathbf{x})\exp(nf(\mathbf{x}))d\mathbf{x}, \quad n\to\infty. \tag{14}$$

We will also use the more general approximation (Breitung 1994, Theorem 44, p.62),

$$\int_F h(\mathbf{x})\exp\left(nf(\mathbf{x})+\sqrt{n}g(\mathbf{x})\right)d\mathbf{x} \sim (2\pi)^{k/2}h(\mathbf{x}^*)\cdot\frac{\exp\left(-\frac{1}{2}\nabla g(\mathbf{x}^*)^\top H_f^{-1}(\mathbf{x}^*)\nabla g(\mathbf{x}^*)\right)}{\sqrt{|\det(H_f(\mathbf{x}^*))|}} \cdot\exp\left(nf(\mathbf{x}^*)+\sqrt{n}g(\mathbf{x}^*)\right)\cdot n^{-k/2} \tag{15}$$

where h, f, F satisfy the same assumptions as in (13), and g is assumed to be in $C^1(F)$. This result allows us to include a component g that is non-concave near $\mathbf{x}^*$ by making its contribution to the exponent of smaller order in n.

The approximation in (15) is the main reason we chose the $\sqrt{n}$ scaling in setting the IS distribution to $\text{Dir}_{\alpha+\sqrt{n}\theta^*}$. The likelihood ratio contributes a non-concave term $-\theta^* \log \theta$ to the exponent of the integrand, and the $\sqrt{n}$ scaling allows us to analyze its performance using (15).

A.2 Negative Definiteness of $\nabla^2 H(\theta)$

Lemma 2 Suppose (A2) holds. Then the Hessian $\nabla^2 H(\theta)$ is negative definite at every $\theta \in \Delta_{K-1}$.

Proof. Write $\phi(v)$ for the vector $(\phi_1(v), \ldots, \phi_K(v))^\top$, which is the vth column of the matrix in (A2). Then $\theta^\top \phi(v) > 0$, for all $\theta \in \Delta_{K-1}$. The Hessian of $H(\theta)$ is given by

$$\nabla^2 H(\theta) = -\sum_{v=1}^{V} p_v \cdot \frac{\phi(v)\phi(v)^\top}{(\theta^\top \phi(v))^2}. \tag{16}$$

To show negative definiteness, consider any nonzero vector $x \in \mathbb{R}^K$ and note that

$$x^\top \nabla^2 H(\theta) x = -\sum_{v=1}^{V} p_v \cdot \frac{(x^\top \phi(v))^2}{(\theta^\top \phi(v))^2}.$$

This expression is strictly negative unless $x^\top \phi(v) = 0$ for all v. But this would imply that x is orthogonal to all $\phi(v)$, contradicting the assumption that the vectors $\{\phi(v)\}_{v=1}^V$ span $\mathbb{R}^K$.

A.3 Bias Decay Bound

Proof of Lemma 1. Since H is C^2 and its Hessian is negative definite on Δ_{K-1} with $\theta^* \in \Delta_{K-1}^{\varepsilon}$, there is a $\delta > 0$ such that

$$\sup_{\theta \in (\Delta_{K-1}^{\varepsilon})^c} H(\theta) \le H(\theta^*) - \delta. \tag{17}$$

Letting $\eta = \inf_{(\Delta_{K-1}^{\varepsilon})^c} ||\theta^* - \theta||$, this δ is bounded away from zero by $\eta^2/2$ times the smallest eigenvalue of $\left(\sum_{v=1}^V p_v \phi(v)\phi(v)^\top\right)$, which is strictly positive. From (17) we get

$$\text{Bias}(\hat{p}_{IS}) = \mathbb{E}_{\theta \sim \text{Dir}_\alpha}[e^{nH(\theta)} \mathbf{1}_{(\Delta_{K-1}^{\varepsilon})^c}] \le e^{n(H(\theta^*)-\delta)}. \tag{18}$$

To bound $\text{Var}(\hat{p}_{MC})$ from below, we first note the following the lower bound of $\mathbb{E}_{\theta \sim \text{Dir}_\alpha}\left[e^{2nH(\theta)}\right]$. The function H is convex in $\log \theta$. Combining this property with the lower bound for the beta function in (36) yields the following lower bound:

$$\begin{aligned}
\mathbb{E}_{\theta \sim \text{Dir}_\alpha}\left[e^{2nH(\theta)}\right] &\ge \mathbb{E}_{\theta \sim \text{Dir}_\alpha}\left[e^{2nH(\theta^*)+2n\theta^* \cdot (\log\theta - \log\theta^*)}\right] \\
&= \frac{e^{2nH(\theta^*)-2n\theta^* \cdot \log\theta^*}}{B(\alpha)} B(\alpha + 2n\theta^*) \\
&\ge C \cdot R \cdot \frac{e^{2nH(\theta^*)}}{B(\alpha)} \left(\frac{\pi}{n}\right)^{\frac{K-1}{2}} \left(\prod_{k=1}^{K} \theta_k^{*(\alpha_k - \frac{1}{2})}\right),
\end{aligned} \tag{19}$$

with $C, R \in (0, \infty)$ as defined in (39)–(40). From (29), we have that $(\mathbb{E}_{\theta \sim \text{Dir}_\alpha}\left[e^{nH(\theta)}\right])^2 = o(n^{-\frac{K-1}{2}} e^{2nH(\theta^*)})$, which, together with (19), implies that there exist a constant $C' > 0$ such that, for large n,

$$\text{Var}_\alpha(e^{nH(\theta)}) \ge C' n^{-\frac{K-1}{2}} e^{nH(\theta^*)}. \tag{20}$$

Combining (18) and (20), we get

$$\begin{aligned}\frac{\mathrm{Bias}(\hat{p}_{\mathrm{IS}})^2}{\mathrm{Var}(\hat{p}_{\mathrm{MC}})} &\leq \frac{e^{2n(H(\theta^*)-\delta)}}{C' n^{-\frac{K-1}{2}} e^{nH(\theta^*)}} \\ &= (C')^{-1} e^{-2n\delta} n^{\frac{K-1}{2}} = O(e^{-2n\delta'}) \quad \forall 0 < \delta' < \delta.\end{aligned} \tag{21}$$

A.4 Second Moment of MC Estimator

For the second moment of the MC estimator, we have the following asymptotic result.

Lemma 3 Suppose (A2) and θ^* is an interior point of the $(K-1)$ simplex, i.e. $\theta_j^* > 0,\ j = 1,\ldots,K$. Then, as $n \to \infty$,

$$\mathbb{E}_{\theta\sim\mathrm{Dir}_\alpha}[e^{2nH(\theta)}] \sim \frac{e^{2nH(\theta^*)}}{(|\det(\nabla^2 H(\theta^*)|)^{1/2}} (\frac{\pi}{n})^{(K-1)/2} \mathrm{Dir}_\alpha(\theta^*). \tag{22}$$

Proof. Define the projected simplex in $\mathbb{R}^{K-1}$,

$$\tilde{\Delta}_{K-1} = \{\tilde{\theta} \in \mathbb{R}^{K-1} | \sum_{i=1}^{K-1} \tilde{\theta}_i \leq 1, \tilde{\theta}_i \geq 0, i = 1,\ldots,K-1\}.$$

With this we can express the expectation in (22) as

$$\mathbb{E}_{\theta\sim\mathrm{Dir}_\alpha}[e^{2nH(\theta)}] = \int_{\tilde{\Delta}_{K-1}} e^{2nH(T(\tilde{\theta}))} \mathrm{Dir}_\alpha(T(\tilde{\theta})) d\tilde{\theta}, \tag{23}$$

where $T : \mathbb{R}^{K-1} \to \mathbb{R}^K$ is the mapping $T(y) := (y_1,\ldots,y_{K-1}, 1-\sum_{i=1}^{K-1} y_i)$. In other words,

$$T(y) = Ay + b, \quad \text{with} \quad A = \begin{bmatrix} I_{K-1} \\ -\mathbf{1}_{K-1}^\top \end{bmatrix}, \quad b = e_K, \tag{24}$$

where I_{K-1} is $K-1$ dimensional identity matrix, $\mathbf{1}_{K-1}$ a vector of ones, and $b = e_K$ a unit vector. With this reparameterization, the maximizer of the exponent is now in the strict interior of $\tilde{\Delta}_{K-1}$. As in (14), we can further restrict the domain by looking at the projected truncated simplex $\tilde{\Delta}_{K-1}^\varepsilon$,

$$\tilde{\Delta}_{K-1}^\varepsilon = \{\tilde{\theta} \in \mathbb{R}^{K-1} | \sum_{i}^{K-1} \tilde{\theta}_i \leq 1-\varepsilon, \tilde{\theta}_i \geq \varepsilon, i = 1,\ldots,K-1\}. \tag{25}$$

where $\varepsilon > 0$ is small enough that $\theta^* \in \Delta_{K-1}^\varepsilon$. Now $\mathrm{Dir}_\alpha(T(\tilde{\theta}))$ is continuous on $\tilde{\Delta}_{K-1}^\varepsilon$. We can now use (14) and (13) to get

$$\int_{\tilde{\Delta}_{K-1}} e^{2nH(T(\tilde{\theta}))} \mathrm{Dir}_\alpha(T(\tilde{\theta})) d\tilde{\theta} \sim \frac{e^{2nH(\theta^*)}}{(|\det(A^\top \nabla^2 H(\theta^*) A|)^{1/2}} (\frac{\pi}{n})^{(K-1)/2} \mathrm{Dir}_\alpha(\theta^*). \tag{26}$$

By (A2) and the fact that A has full rank, $|\det(A^\top \nabla^2 H(\theta^*) A)| = \det(-A^\top \nabla^2 H(\theta^*) A)$. The right hand side can be further evaluated by defining $C = \begin{bmatrix} A & 1_K \end{bmatrix} \in \mathbb{R}^{K\times K}$ using the Schur complement of $-A^\top \nabla^2 H(\theta^*) A$ in $-C^\top \nabla^2 H(\theta^*) C$ to get

$$\det(-A^\top \nabla^2 H(\theta^*) A) = \det\left(-\nabla^2 H(\theta^*)\right) \cdot \left(\mathbf{1}_K^\top (-\nabla^2 H(\theta^*))^{-1} \mathbf{1}_K\right). \tag{27}$$

To evaluate the rightmost factor, we note that the gradient of H at θ^* is

$$\nabla_\theta H(\theta^*) = \sum_v p_v \nabla_\theta \log(\theta^\top \phi(v))\Big|_{\theta=\theta^*} = \sum_v p_v \frac{\phi(v)}{\theta^{*\top}\phi(v)},$$

which equals $-\nabla^2 H(\theta^*)\theta^*$, in light of the expression in (16). By assumption (A1) and KKT conditions,

$$\nabla_\theta H(\theta^*) = \lambda \mathbf{1}_K.$$

Since $\nabla_\theta H(\theta^*)\cdot\theta^* = 1$, we have that $\lambda = 1$, which implies $(-\nabla^2 H(\theta^*))\theta^* = \mathbf{1}_K$. This further implies $(-\nabla^2 H(\theta^*))^{-1}\mathbf{1}_K = \theta^*$, and therefore that $\mathbf{1}_K^\top(-\nabla_\theta^2 H(\theta^*))^{-1}\mathbf{1}_K = \mathbf{1}_K^\top\theta^* = 1$. Thus, (27) becomes

$$\det(-A^\top \nabla^2 H(\theta^*)A) = \det\left(-\nabla^2 H(\theta^*)\right). \tag{28}$$

Making this substitution in (26) yields the result in (22).

By taking n instead of $2n$ we get the asymptotics for the first moment

$$\mathbb{E}_{\theta\sim\mathrm{Dir}_\alpha}[e^{nH(\theta)}] \sim \frac{e^{nH(\theta^*)}}{(|\det(\nabla^2 H(\theta^*)|)^{1/2}}(\frac{2\pi}{n})^{(K-1)/2}\mathrm{Dir}_\alpha(\tilde{\theta}^*). \tag{29}$$

A.5 Second Moment of Importance Sampling Estimator

We introduce the following asymptotic result for the second moment of the IS estimator in (7).

Lemma 4 Suppose (A1) and (A2) hold. Then as $n\to\infty$,

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}&\left[e^{2nH(\theta)}\left(\frac{\mathrm{Dir}_\alpha(\theta)}{\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}(\theta)}\right)^2 \mathbf{1}_{\Delta^\varepsilon_{K-1}}\right]\\ &\sim \frac{e^{2nH(\theta^*)}}{\sqrt{|\det\nabla^2 H(\theta^*)|}}\left(\frac{2\pi^2}{n^{3/2}}\right)^{\frac{K-1}{2}}\mathrm{Dir}_\alpha(\theta^*)^2\left(\prod_k \theta_k^*\right)^{\frac{1}{2}}.\end{aligned} \tag{30}$$

Proof. Using (25) and $T(\tilde{\theta})$ as in (26), the left side of (30) can be expressed as

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}\left[\mathbf{1}_{\Delta^\varepsilon_{K-1}} e^{2nH(\theta)}\left(\frac{\mathrm{Dir}_\alpha(\theta)}{\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}(\theta)}\right)^2\right] &= \mathbb{E}_{\theta\sim\mathrm{Dir}_\alpha}\left[\mathbf{1}_{\Delta^\varepsilon_{K-1}} e^{2nH(\theta)}\frac{\mathrm{Dir}_\alpha(\theta)}{\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}(\theta)}\right]\\ &= \frac{B(\alpha+\sqrt{n}\theta^*)}{B(\alpha)^2}\int_{\tilde{\Delta}^\varepsilon_{K-1}}\prod_{i=1}^K T(\tilde{\theta})_i^{\alpha_i-1}\\ &\quad\cdot\exp\left(2nH(T(\tilde{\theta})) - \sqrt{n}\,\theta^*\cdot\log T(\tilde{\theta})\right)d\tilde{\theta}.\end{aligned} \tag{31}$$

We can now apply (15) to get

$$\begin{aligned}&\int_{\tilde{\Delta}^\varepsilon_{K-1}}\prod_{i=1}^K T(\tilde{\theta})_i^{\alpha_i-1}\exp\left(2nH(T(\tilde{\theta})) - \sqrt{n}\,\theta^*\cdot\log T(\tilde{\theta})\right)d\tilde{\theta}\\ &\sim n^{-\frac{K-1}{2}}e^{2nH(\theta^*)-\sqrt{n}\theta^*\cdot\log\theta^*}(2\pi)^{\frac{K-1}{2}}\left(\prod_{i=1}^K(\theta_i^*)^{\alpha_i-1}\right)\frac{\exp\left(-\frac{1}{2}(A^\top\mathbf{1}_K)^\top(A^\top\nabla^2 H(\theta^*)A)^{-1}(A^\top\mathbf{1}_K)\right)}{(|\det(A^\top\nabla^2 H(\theta^*)A)|)^{1/2}}\\ &= (2\pi)^{\frac{K-1}{2}}\left(\prod_{i=1}^K(\theta_i^*)^{\alpha_i-1}\right)\cdot\frac{e^{2nH(\theta^*)-\sqrt{n}\theta^*\cdot\log\theta^*}\cdot n^{-\frac{K-1}{2}}}{(|\det(\nabla^2 H(\theta^*))|)^{1/2}}.\end{aligned} \tag{32}$$

Thus, (31) is asymptotic to (32) times $B(\alpha+\sqrt{n}\theta^*)/B(\alpha)^2$. Replacing $B(\alpha+\sqrt{n}\theta^*)$ with the Stirling approximation in (35) yields (30).

Remark. Equation (31) provides some insight into the finding (in Corollary 1) that the relative performance of the IS estimator theoretically improves with larger K. The integral in (31) and the normalizing factor $B(\alpha+\sqrt{n}\theta^*)$ contribute factors of the order of $n^{-(K-1)/2}$ and $n^{-(K-1)/4}$, respectively, to the second moment of the IS estimator through their respective asymptotic approximations. For the second moment of the plain MC estimator (22), we have only a single factor of order $n^{-(K-1)/2}$.

A.6 Variance Ratio Asymptotics

Proof of Theorem 1. To simplify notation, we take $N=1$ replications in the definitions of $\hat{p}_{\mathrm{MC}}$ and $\hat{p}_{\mathrm{IS}}$. Then

$$\frac{\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}\left[\hat{p}_{\mathrm{IS}}^2\right]-\left(\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}\left[\hat{p}_{\mathrm{IS}}\right]\right)^2}{\mathbb{E}_{\theta\sim\mathrm{Dir}(\alpha)}\left[\hat{p}_{\mathrm{MC}}^2\right]}\le\frac{\mathrm{Var}(\hat{p}_{\mathrm{IS}})}{\mathrm{Var}(\hat{p}_{\mathrm{MC}})}\le\frac{\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}\left[\hat{p}_{\mathrm{IS}}^2\right]}{\mathbb{E}_{\theta\sim\mathrm{Dir}(\alpha)}\left[\hat{p}_{\mathrm{MC}}^2\right]-\left(\mathbb{E}_{\theta\sim\mathrm{Dir}(\alpha)}\left[\hat{p}_{\mathrm{MC}}\right]\right)^2}. \quad (33)$$

On both sides of the inequalities, the second-moment terms dominate, and the squared first moments are of higher order in n; see (29). Thus, making the substitutions from (22) and (30) shows that the variance ratio is asymptotic to

$$\begin{aligned}\frac{\mathbb{E}_{\theta\sim\mathrm{Dir}_{\alpha+\sqrt{n}\theta^*}}\left[\hat{p}_{\mathrm{IS}}^2\right]}{\mathbb{E}_{\theta\sim\mathrm{Dir}(\alpha)}\left[\hat{p}_{\mathrm{MC}}^2\right]}&\sim\frac{\frac{e^{2nH(\theta^*)}}{|\det(-\nabla^2H(\theta^*))|^{1/2}}\cdot\left(\frac{2\pi^2}{n^{3/2}}\right)^{(K-1)/2}\cdot\mathrm{Dir}_\alpha(\theta^*)^2\cdot\left(\prod_k\theta_k^*\right)^{1/2}}{\frac{e^{2nH(\theta^*)}}{|\det(-\nabla^2H(\theta^*))|^{1/2}}\cdot\left(\frac{\pi}{n}\right)^{(K-1)/2}\cdot\mathrm{Dir}_\alpha(\theta^*)}\\&=n^{-\frac{(K-1)}{4}}(2\pi)^{(K-1)/2}\cdot\mathrm{Dir}_\alpha(\theta^*)\cdot\left(\prod_k\theta_k^*\right)^{1/2}.\end{aligned}$$

A.7 MSE Reduction

Proof of Corollary 1. Recalling that $\hat{p}_{\mathrm{MC}}$ is unbiased, we know that

$$\frac{\mathrm{Var}(\hat{p}_{\mathrm{IS}})}{\mathrm{Var}(\hat{p}_{\mathrm{MC}})}\le\frac{\mathrm{MSE}(\hat{p}_{\mathrm{IS}})}{\mathrm{MSE}(\hat{p}_{\mathrm{MC}})}\le\frac{\mathrm{Var}(\hat{p}_{\mathrm{IS}})+O(e^{-2n\delta})}{\mathrm{Var}(\hat{p}_{\mathrm{MC}})}. \quad (34)$$

The order of the lower bound is $\Theta(n^{-(K-1)/4})$, in light of (11). The upper bound has the same order because, for $\delta>0$, $O(e^{-2n\delta})$ is $o(n^{-k})$, for all $k>0$, and therefore does not change the order of the ratio.

A.8 Stirling's Formula for Beta Function

Lemma 5 Suppose $\theta^*\in\Delta_{K-1}$, with $\theta_k^*>0$, $k=1,\dots,K$. Then, as $n\to\infty$,

$$B\left(\alpha+\sqrt{n}\,\theta^*\right)\sim(2\pi)^{\frac{K-1}{2}}\,n^{-\frac{K-1}{4}}\exp\left\{\sqrt{n}\sum_{k=1}^K\theta_k^*\ln\theta_k^*+\sum_{k=1}^K\left(\alpha_k-\frac{1}{2}\right)\ln\theta_k^*\right\}. \quad (35)$$

Proof. Stirling's approximation states that, as $x\to\infty$,

$$\Gamma(x)\sim\sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x}.$$

Since $\theta_k^*>0$ and $\sum_{k=1}^K\theta_k^*=1$, applying Stirling's formula to (1) evaluated at $\alpha+\sqrt{n}\theta^*$ and simplifying yields (35).

We also note the following (non-asymptotic) lower bound of the Beta constant

$$B(\alpha+n\theta^*) \geq C_n R_n (2\pi)^{\frac{K-1}{2}} \left(\prod_{k=1}^{K} \theta_k^{*(\alpha_k-\frac{1}{2})}\right) n^{-\frac{(K-1)}{2}} e^{n\theta^*\cdot\log\theta^*}, \tag{36}$$

where

$$C_n = e^{-\frac{1}{12(\alpha_0+n)}} \tag{37}$$

$$R_n = \prod_{k=1}^{K}(1+\frac{\alpha_k}{n\theta_k^*})^{\alpha_k-1/2} \left(\prod_{k=1}^{K}(1+\frac{\alpha_k}{n\theta_k^*})^{(n\theta_k^*)}\right) \left(1-\frac{\alpha_0}{\alpha_0+n}\right)^{\alpha_0+n}. \tag{38}$$

This follows from Theorem 5 in Gordon (1994): $\forall x > 0$

$$\sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x} \leq \Gamma(x) \leq \sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x+\frac{1}{12x}}.$$

Note that for all $n \geq 1$ we have

$$C_n \geq C \equiv e^{-\frac{1}{12\alpha_0}} \tag{39}$$

$$R_n \geq R \equiv \prod_{k=1}^{K} \min\left\{1, \left(1+\frac{\alpha_k}{\theta_k^*}\right)^{\alpha_k-\frac{1}{2}}\right\} \left(1-\frac{\alpha_0}{\alpha_0+1}\right)^{\alpha_0+1}. \tag{40}$$

We may therefore replace C_n with C and R_n with R in (36), as we do in (19).

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer.

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

Breitung, K. W. 1994. *Asymptotic Approximations for Probability Integrals*. Berlin: Springer.

Gordon, L. 1994. "A Stochastic Approach to the Gamma Function". *The American Mathematical Monthly* 101(9):858–865.

Griffiths, T. L., and M. Steyvers. 2004. "Finding Scientific Topics". *Proceedings of the National Academy of Sciences* 101(Supplement 1):5228–5235.

Kroese, D. P., T. Taimre, and Z. I. Botev. 2013. *Handbook of Monte Carlo Methods*. Hoboken, New Jersey: John Wiley & Sons.

Lang, K. 1995. "NewsWeeder: Learning to Filter Netnews". In *Machine Learning: Proceedings of the Twelfth International Conference on Machine Learning*, edited by A. Prieditis and S. Russell, 331–339. San Francisco: Morgan Kaufmann.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. "Evaluation Methods for Topic Models". In *Proceedings of the 26th Annual International Conference on Machine Learning*, edited by A. P. Danyluk, L. Bottou, and M. L. Littman, 1105–1112. New York: ACM.

AUTHOR BIOGRAPHIES

PAUL GLASSERMAN is the Jack R. Anderson Professor of Business at Columbia University. He is author of the book *Monte Carlo Methods in Financial Engineering*, which was awarded the INFORMS Lanchester Prize, and he is a past recipient of the I-Sim Outstanding Simulation Publication Award. His recent research includes applications of text analysis in finance, which motivated his interest in topic modeling. His email address is pg20@columbia.edu.

AYEONG LEE is a PhD student in Decision, Risk, and Operations Program at Columbia University. She holds a Bachelor's degree in Applied Mathematics from Columbia University and a Master's degree in Applied Mathematics from Brown University. Her interests are stochastic simulation, applied probability, and large deviations. Her email address is ayeong.lee@columbia.edu