Evaluation Methods for Topic Models

Hanna M. Wallach WALLACH@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA

Iain Murray MURRAY@CS.TORONTO.EDU
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4 CANADA

David Mimno MIMNO@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA

Abstract

A natural evaluation metric for statistical topic models is the probability of held-out documents given a trained model. While exact computation of this probability is intractable, several estimators for this probability have been used in the topic modeling literature, including the harmonic mean method and empirical likelihood method. In this paper, we demonstrate experimentally that commonly-used methods are unlikely to accurately estimate the probability of held-out documents, and propose two alternative methods that are both accurate and efficient.

1. Introduction

Statistical topic modeling is an increasingly useful tool for analyzing large unstructured text collections. There is a significant body of work introducing and developing sophisticated topic models and their applications. To date, however, there have not been any papers specifically addressing the issue of evaluating topic models. Evaluation is an important issue: the unsupervised nature of topic models makes model selection difficult. For some applications there may be extrinsic tasks, such as information retrieval or document classification, for which performance can be evaluated. However, there is a need for a universal method that measures the generalization capability of a topic model in a way that is accurate, computationally efficient, and independent of any specific application.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

In this paper we consider only the simplest topic model, latent Dirichlet allocation (LDA), and compare a number of methods for estimating the probability of held-out documents given a trained model. Most of the methods presented, however, are applicable to more complicated topic models. In addition to comparing evaluation methods that are currently used in the topic modeling literature, we propose several alternative methods. We present empirical results on synthetic and real-world data sets showing that the currently-used estimators are less accurate and have higher variance than the proposed new estimators.

2. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA), originally introduced by Blei et al. (2003), is a generative model for text. In this model, a "topic" t is a discrete distribution over words with probability vector $\boldsymbol{\phi}_t$. Dirichlet priors, with concentration parameter β and base measure $\boldsymbol{n}$, are placed over the topics $\Phi = \{\boldsymbol{\phi}_1, \dots \boldsymbol{\phi}_T\}$:

$$P(\Phi) = \textstyle\prod_t \operatorname{Dir}(\boldsymbol{\phi}_t; \beta \boldsymbol{n}). \tag{1}$$

Each document, indexed by d, is assumed to have its own distribution over topics given by probabilities $\boldsymbol{\theta}_d$. The priors over $\Theta = \{\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_D\}$ are also Dirichlet, with concentration parameter α and base measure $\boldsymbol{m}$:

$$P(\Theta) = \textstyle\prod_d \operatorname{Dir}(\boldsymbol{\theta}_d; \alpha \boldsymbol{m}). \tag{2}$$

The tokens in a document $\boldsymbol{w}^{(d)} = \{w_n^{(d)}\}_{n=1}^{N_d}$ are associated with topic assignments $\boldsymbol{z}^{(d)} = \{z_n^{(d)}\}_{n=1}^{N_d}$, drawn i.i.d. from the document-specific topic distribution:

$$P(\boldsymbol{z}^{(d)} \mid \boldsymbol{\theta}_d) = \textstyle\prod_n \theta_{z_n^{(d)}|d}\,. \tag{3}$$

The tokens are drawn from the topics' distributions:

$$P(\boldsymbol{w}^{(d)} \mid \boldsymbol{z}^{(d)}, \Phi) = \textstyle\prod_n \phi_{w_n^{(d)}|z_n^{(d)}}\,. \tag{4}$$

A data set of documents $\mathcal{W} = \{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(D)}\}$ is observed, while the underlying corresponding topic assignments $\mathcal{Z} = \{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, ..., \boldsymbol{z}^{(D)}\}$ are unobserved.

Conjugacy of Dirichlets with multinomials allows the parameters to be marginalized out. For example,

$$P(\boldsymbol{z}^{(d)} \,|\, \alpha\boldsymbol{m}) = \int \mathrm{d}\boldsymbol{\theta}_d \, P(\boldsymbol{z}^{(d)} \,|\, \boldsymbol{\theta}_d) \, P(\boldsymbol{\theta}_d \,|\, \alpha\boldsymbol{m}) = \frac{\Gamma(\alpha)}{\Gamma(N_d + \alpha)} \prod_t \frac{\Gamma(N_{t|d} + \alpha m_t)}{\Gamma(\alpha m_t)}, \quad (5)$$

where topic t occurs $N_{t|d}$ times in $\boldsymbol{z}^{(d)}$ of length N_d.

3. Evaluating LDA

LDA is typically evaluated by either measuring performance on some secondary task, such as document classification or information retrieval, or by estimating the probability of unseen held-out documents given some training documents. A better model will give rise to a higher probability of held-out documents, on average.

The probability of a set of held-out documents $\mathcal{W}$ given a set of training documents $\mathcal{W}'$, can be written as

$$P(\mathcal{W} \,|\, \mathcal{W}') = \int \mathrm{d}\Phi \, \mathrm{d}\alpha \, \mathrm{d}\boldsymbol{m} \, P(\mathcal{W} \,|\, \Phi, \alpha\boldsymbol{m}) \, P(\Phi, \alpha\boldsymbol{m} \,|\, \mathcal{W}').$$

This integral can be approximated by averaging $P(\mathcal{W} \,|\, \Phi, \alpha\boldsymbol{m})$ under samples from $P(\Phi, \alpha\boldsymbol{m} \,|\, \mathcal{W}')$, or evaluating at a point estimate. We take the latter approach. Variational methods (Blei et al., 2003) and MCMC methods (Griffiths & Steyvers, 2004) are effective at marginalizing out the topic assignments $\mathcal{Z}$ associated with the training data to infer Φ and $\alpha\boldsymbol{m}$.

In this paper, we focus on evaluating

$$P(\mathcal{W} \,|\, \Phi, \alpha\boldsymbol{m}) = \textstyle\prod_d P(\boldsymbol{w}^{(d)} \,|\, \Phi, \alpha\boldsymbol{m}). \quad (6)$$

Since the topic assignments for one document are independent of the topic assignments for all other documents, each held-out document can be evaluated separately. For the rest of this paper, we refer to the current document as $\boldsymbol{w}$, its latent topic assignments as $\boldsymbol{z}$, and its document-specific topic distribution as $\boldsymbol{\theta}$.

Many of the evaluation methods in this paper require the ability to obtain a set of topic assignments $\boldsymbol{z}$ for document $\boldsymbol{w}$ using Gibbs sampling. Gibbs sampling involves sequentially resampling each z_n from its conditional posterior given $\boldsymbol{w}$, Φ, $\alpha\boldsymbol{m}$ and $\boldsymbol{z}_{\backslash n}$ (the current latent topic assignments for all other tokens):

$$\begin{aligned} P(z_n = t \,|\, \boldsymbol{w}, \boldsymbol{z}_{\backslash n}, \Phi, \alpha\boldsymbol{m}) &\propto P(w_n \,|\, z_n = t, \Phi) \, P(z_n = t \,|\, \boldsymbol{z}_{\backslash n}, \alpha\boldsymbol{m}) \\ &\propto \phi_{w_n|t} \, \frac{\{N_t\}_{\backslash n} + \alpha m_t}{N - 1 + \alpha}, \end{aligned} \quad (7)$$

where $\{N_t\}_{\backslash n}$ is the number of times topic t occurs in the document in question, excluding position n, and N is the total number of tokens in the document.

4. Estimating $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$

The evaluation probability $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$ for held-out document $\boldsymbol{w}$ can be thought of as the normalizing constant that relates the posterior distribution over $\boldsymbol{z}$ to the joint distribution over $\boldsymbol{w}$ and $\boldsymbol{z}$ in Bayes' rule:

$$P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{z}, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}. \quad (8)$$

There are many existing methods for estimating normalizing constants. In this section, we review some of these methods, as previously applied to topic models, and also outline two alternative methods: a Chib-style estimator and a "left-to-right" evaluation algorithm.

4.1. Importance sampling methods

In general, given a model with observed variables $\boldsymbol{w}$ and unknown variables $\boldsymbol{h}$, importance sampling can be used to approximate the probability of the observed variables, either $P(\boldsymbol{w}) = \sum_{\boldsymbol{h}} P(\boldsymbol{w}, \boldsymbol{h})$ or $\int \mathrm{d}\boldsymbol{h} \, P(\boldsymbol{w}, \boldsymbol{h})$. If $Q(\boldsymbol{h})$ is some simple, tractable distribution over $\boldsymbol{h}$—the "proposal distribution"—then

$$P(\boldsymbol{w}) \simeq \frac{1}{S} \sum_s \frac{P(\boldsymbol{w}, \boldsymbol{h}^{(s)})}{Q(\boldsymbol{h}^{(s)})}, \quad \boldsymbol{h}^{(s)} \sim Q(\boldsymbol{h}), \quad (9)$$

is an unbiased estimator. To ensure low variance, $Q(\boldsymbol{h})$ must be similar to the "target distribution" $P(\boldsymbol{h} \,|\, \boldsymbol{w})$ and must be non-zero wherever $P(\boldsymbol{w}, \boldsymbol{h})$ is non-zero.

In this section, we explain how $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$ can be estimated using importance sampling by either (a) integrating out $\boldsymbol{\theta}$ and using the prior over $\boldsymbol{h} = \boldsymbol{z}$ as the proposal distribution, or (b) using the prior over $\boldsymbol{h} = \boldsymbol{\theta}$ as the proposal distribution, thereby allowing the topic assignments $\boldsymbol{z}$ to be marginalized out directly.

If the proposal distribution is the prior over $\boldsymbol{z}$,

$$\begin{aligned} P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) &= \sum_{\boldsymbol{z}} P(\boldsymbol{w} \,|\, \boldsymbol{z}, \Phi) \, P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m}) \\ &\simeq \frac{1}{S} \sum_s P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)}, \Phi), \end{aligned} \quad (10)$$

where $\boldsymbol{z}^{(s)} \sim P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m})$. Unfortunately, topic assignments drawn from the prior, without consideration of the corresponding tokens, are unlikely to provide a good explanation of $\boldsymbol{w}$. The prior is not usually close to the target distribution unless $\boldsymbol{w}$ is very short.

Better proposal distributions for $\boldsymbol{z}^{(s)}$ can be constructed by taking $\boldsymbol{w}$ into account. The simplest way

is to form a distribution over topics for each token w_n, ignoring dependencies between tokens: $Q(z_n) \propto \alpha m_{z_n} \phi_{w_n|z_n}$. A more sophisticated method, which we call "iterated pseudo-counts," involves iteratively updating $Q(z_n)$ every sampling iteration. After initializing $Q(z_n)^{(0)} \propto \alpha m_{z_n} \phi_{w_n|z_n}$, the update rule is

$$Q(z_n)^{(s)} \propto (\alpha m_{z_n} + \sum_{n' \neq n} Q(z_{n'})^{(s-1)})\, \phi_{w_n|z_n}. \quad (11)$$

Alternatively, $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$ can be written as an integral over the document-specific topic distribution $\boldsymbol{\theta}$:

$$\begin{aligned} P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) &= \int \mathrm{d}\boldsymbol{\theta}\, P(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \Phi)\, P(\boldsymbol{\theta} \,|\, \alpha\boldsymbol{m}) \\ &\simeq \frac{1}{S} \sum_s P(\boldsymbol{w} \,|\, \boldsymbol{\theta}^{(s)}, \Phi), \end{aligned} \quad (12)$$

where $\boldsymbol{\theta}^{(s)}$ is drawn from $P(\boldsymbol{\theta} \,|\, \alpha\boldsymbol{m}) = \mathrm{Dir}\,(\boldsymbol{\theta}; \alpha\boldsymbol{m})$. The estimator in (12) is easily computed because the topic assignments are independent given $\boldsymbol{\theta}$:

$$\begin{aligned} P(\boldsymbol{w} \,|\, \boldsymbol{\theta}^{(s)}, \Phi) &= \prod_n P(w_n \,|\, \boldsymbol{\theta}^{(s)}, \Phi) \\ &= \prod_n \sum_{z_n} P(w_n, z_n \,|\, \boldsymbol{\theta}^{(s)}, \Phi). \end{aligned} \quad (13)$$

If the probabilities $P(w \,|\, \boldsymbol{\theta}^{(s)}, \Phi)$ are estimated from a synthetic document, randomly-generated using $\boldsymbol{\theta}^{(s)}$, the resultant estimator corresponds to the empirical likelihood method described by Li and McCallum (2006). Used directly, however, (13) will give the same result as using infinitely long synthetic documents and is how the empirical likelihood method is implemented in MALLET (McCallum, 2002).

Importance sampling does not work well when sampling from high-dimensional distributions. Unless the proposal distribution is a near-perfect approximation to the target distribution, the variance of the estimator will be very large. When sampling continuous values, such as $\boldsymbol{\theta}$, the estimator may have infinite variance.

4.2. Harmonic mean method

The harmonic mean method (Newton & Raftery, 1994) is based on the following unbiased estimator:

$$\frac{1}{P(\boldsymbol{w})} = \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{z} \,|\, \boldsymbol{w})}{P(\boldsymbol{w} \,|\, \boldsymbol{z})} \simeq \frac{1}{S} \sum_s \frac{1}{P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)})}, \quad (14)$$

where $\boldsymbol{z}^{(s)}$ is drawn from $P(\boldsymbol{z} \,|\, \boldsymbol{w})$. Conditioning on Φ and $\alpha\boldsymbol{m}$ gives an estimator for $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$:

$$\begin{aligned} P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) &\simeq \frac{1}{\frac{1}{S}\sum_s \frac{1}{P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)}, \Phi)}} \\ &= \mathrm{HM}\,(\{P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)}, \Phi)\}_{s=1}^S), \end{aligned} \quad (15)$$

where $\boldsymbol{z}^{(s)} \sim P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})$ and $\mathrm{HM}\,(\cdot)$ denotes the harmonic mean. In practice, $\{\boldsymbol{z}^{(s)}\}_{s=1}^S$ are S samples taken from a Gibbs sampler after a burn-in period of B iterations. Since the samples are used to approximate an expectation, they need not be independent and thinning is unnecessary. Consequently, the cost of the estimator is that of $S + B$ Gibbs iterations.

Newton and Raftery (1994) expressed reservations about the harmonic mean method when introducing it, and Neal added further criticism in the discussion. Despite these criticisms, it has been used in several topic modeling papers (Griffiths & Steyvers, 2004; Griffiths et al., 2005; Wallach, 2006), due to its ease of implementation and relative computational efficiency.

4.3. Annealed importance sampling

Annealed importance sampling (AIS) can be viewed as a variant of simple importance sampling defined on a higher-dimensional state space (Neal, 2001). Many auxiliary variables are introduced in order to make the proposal distribution closer to the target distribution. When used to approximate $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$, AIS uses the following sequence of probability distributions:

$$P_s(\boldsymbol{z}) \propto P(\boldsymbol{w} \,|\, \boldsymbol{z}, \Phi)^{\tau_s}\, P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m}),$$

defined by a set of "inverse temperatures," $0 = \tau_0 < \tau_1 < \ldots < \tau_S = 1$. When $s = 0$, $\tau_s = 0$ and so $P_0(\boldsymbol{z})$ is the prior distribution $P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m})$. Similarly, when $s = S$, $P_S(\boldsymbol{z})$ is the posterior distribution $P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})$. Intermediate values of s interpolate between the prior and posterior distributions. For each $s = 1, \ldots, S-1$, a Markov chain transition operator $T_s(\boldsymbol{z}' \leftarrow \boldsymbol{z})$ that leaves $P_s(\boldsymbol{z})$ invariant must also be defined. When approximating $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$, $T_s(\boldsymbol{z}' \leftarrow \boldsymbol{z})$ is the Gibbs sampling operator that samples sequentially from

$$P_s(z_n \,|\, \boldsymbol{z}_{\setminus n}) \propto P(w_n \,|\, z_n, \Phi)^{\tau_s}\, P(z_n \,|\, \boldsymbol{z}_{\setminus n}, \alpha\boldsymbol{m}). \quad (16)$$

Sampling from (16) is as easy as sampling from (7).

AIS builds a proposal distribution $Q(Z)$ over the extended state space $Z = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(S)}\}$ by first sampling from the tractable prior $P_0(\boldsymbol{z})$ and then applying a series of transition operators $T_1, T_2, \ldots, T_{S-1}$ that "move" the sample through the intermediate distributions $P_s(\boldsymbol{z})$ towards the posterior $P_S(\boldsymbol{z})$. The probability of the resultant state sequence Z is given by

$$Q(Z) = P_0(\boldsymbol{z}^{(1)}) \prod_{s=1}^{S-1} T_s(\boldsymbol{z}^{(s+1)} \leftarrow \boldsymbol{z}^{(s)}). \quad (17)$$

The target distribution for the proposal $Q(Z)$ is

$$P(Z) = P_S(\boldsymbol{z}^{(S)}) \prod_{s=1}^{S-1} \widetilde{T}_s(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s+1)}), \quad (18)$$

1: initialize $0 = \tau_0 < \tau_1 < \ldots < \tau_S = 1$
2: sample $\boldsymbol{z}^{(1)}$ from the prior $P_0(\boldsymbol{z}) = P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m})$.
3: **for** $s = 2 : S$ **do**
4: sample $\boldsymbol{z}^{(s)} \sim T_{s-1}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s-1)})$
5: **end for**
6: $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) \simeq \prod_{s=1}^{S} P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)}, \Phi)^{\tau_s - \tau_{s-1}}$

Algorithm 1: Annealed importance sampling.

where $\widetilde{T}_s$ is the reverse transition operator, given by

$$\widetilde{T}_s(\boldsymbol{z}' \leftarrow \boldsymbol{z}) = T_s(\boldsymbol{z} \leftarrow \boldsymbol{z}') \frac{P_s(\boldsymbol{z}')}{P_s(\boldsymbol{z})}. \tag{19}$$

Having sampled a sequence of topic assignments from $Q(Z)$, a scalar importance weight is constructed:

$$\begin{aligned} w_{\text{AIS}} &= \frac{P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})\, P(Z)}{Q(Z)} \\ &= \frac{P(\boldsymbol{w}, \boldsymbol{z}^{(S)} \,|\, \Phi, \alpha\boldsymbol{m}) \prod_{s=1}^{S-1} \widetilde{T}_s(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s+1)})}{P_0(\boldsymbol{z}^{(1)}) \prod_{s=1}^{S-1} T_s(\boldsymbol{z}^{(s+1)} \leftarrow \boldsymbol{z}^{(s)})} \\ &= \prod_{s=1}^{S} P(\boldsymbol{w} \,|\, \boldsymbol{z}^{(s)}, \Phi)^{\tau_s - \tau_{s-1}}. \end{aligned}$$

Given a set of samples from $Q(Z)$, the corresponding importance weights can be used to approximate $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$ because of the following equality:

$$\begin{aligned} P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) &= P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) \sum_{Z} P(Z) \\ &= \mathbb{E}_{Q(Z)} \left[w_{\text{AIS}} \right]. \end{aligned} \tag{20}$$

The transition operators do not necessarily need to be ergodic. The simple importance sampling approximation in (10), in which the proposal distribution is $P(\boldsymbol{z} \,|\, \alpha\boldsymbol{m})$, is recovered by using transition operators that do nothing: $T_s(\boldsymbol{z}' \leftarrow \boldsymbol{z}) = \delta(\boldsymbol{z}' - \boldsymbol{z})$ for all s.

The AIS algorithm is summarized in algorithm 1.

4.4. Chib-style estimation

For any "special" set of latent topic assignments $\boldsymbol{z}^\star$, Bayes' rule gives rise to the following identity:

$$P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{z}^\star \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})}. \tag{21}$$

Chib (1995) introduced a family of estimators that first pick a $\boldsymbol{z}^\star$ and then estimate the denominator, $P(\boldsymbol{z}^\star \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})$. The numerator $P(\boldsymbol{z}^\star, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) = P(\boldsymbol{w} \,|\, \boldsymbol{z}^\star, \Phi)\, P(\boldsymbol{z}^\star \,|\, \alpha\boldsymbol{m})$ is known from (4) and (5).

Any Markov chain operator T for sampling from the posterior, including the Gibbs sampler, satisfies

$$\begin{aligned} &P(\boldsymbol{z}^\star \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}) \\ &\qquad = \sum_{\boldsymbol{z}} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z})\, P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}). \end{aligned} \tag{22}$$

1: initialize $\boldsymbol{z}^*$ to a high posterior probability state
2: sample s uniformly from $\{1, \ldots, S\}$
3: sample $\boldsymbol{z}^{(s)} \sim \widetilde{T}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^\star)$
4: **for** $s' = (s+1) : S$ **do**
5: sample $\boldsymbol{z}^{(s')} \sim T(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'-1)})$
6: **end for**
7: **for** $s' = (s-1) : -1 : 1$ **do**
8: sample $\boldsymbol{z}^{(s')} \sim \widetilde{T}(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'+1)})$
9: **end for**
10: $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) \simeq$
$P(\boldsymbol{w}, \boldsymbol{z}^\star \,|\, \Phi, \alpha\boldsymbol{m}) \,\Big/\, \frac{1}{S} \sum_{s'} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}^{(s')})$

Algorithm 2: A Chib-style estimator.

(22) can be substituted into (21) to give

$$\begin{aligned} P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) &= \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}{\sum_{\boldsymbol{z}} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z})\, P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})} \\ &\simeq \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}{\frac{1}{S} \sum_{s=1}^{S} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}^{(s)})}, \end{aligned}$$

where $Z = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(S)}\}$ can be obtained by Gibbs sampling from $P(\boldsymbol{z} \,|\, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})$. Murray and Salakhutdinov (2009) showed that this estimator can overestimate the desired probability in expectation. Instead, they constructed the following proposal distribution:

$$\begin{aligned} Q(Z) = &\frac{1}{S} \sum_{s=1}^{S} \widetilde{T}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^\star) \prod_{s'=s+1}^{S} T(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'-1)}) \\ &\cdot \prod_{s'=1}^{s-1} \widetilde{T}(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'+1)}). \end{aligned}$$

Since the forward operator transition T consists of sequentially applying (7) for positions 1 to N (in that order), the reverse transition operator $\widetilde{T}$ can be constructed by simply applying (7) in the reverse order.

Using the definition of $\widetilde{T}$ in (19) it can be shown that

$$P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) \simeq \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})}{\frac{1}{S} \sum_{s=1}^{S} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}^{(s)})}, \tag{23}$$

under samples from $Q(Z)$. In this application, with forwards and reverse Gibbs samplers, the estimator is formally unbiased, even for finite runs of the chain.

The probability of moving to $\boldsymbol{z}^\star$ is given by

$$T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}) = \prod_{n} P(z_n^\star \,|\, \boldsymbol{z}_{<n}^\star, \boldsymbol{z}_{>n}, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}). \tag{24}$$

This Chib-style estimator is valid for any choice of "special state" $\boldsymbol{z}^\star$. We set $\boldsymbol{z}^\star$ by iteratively maximizing (7) for positions $1, \ldots, N$, after a few iterations of regular Gibbs sampling. In all our experiments, less than 1% of computer time was spent setting $\boldsymbol{z}^\star$.

The Chib-style method is summarized in algorithm 2.

initialize $l := 0$
for each position n in $\boldsymbol{w}$ **do**
  initialize $p_n := 0$
  for each particle $r = 1$ to R **do**
    for $n' < n$ **do**
      sample $z_{n'}^{(r)} \sim P(z_{n'}^{(r)} \mid w_{n'}, \{\boldsymbol{z}_{<n}^{(r)}\}_{\setminus n'}, \Phi, \alpha\boldsymbol{m})$
    end for
    $p_n := p_n + \sum_t P(w_n, z_n^{(r)} = t \mid \boldsymbol{z}_{<n}^{(r)}, \Phi, \alpha\boldsymbol{m})$
    sample $z_n^{(r)} \sim P(z_n^{(r)} \mid w_n, \boldsymbol{z}_{<n}^{(r)}, \Phi, \alpha\boldsymbol{m})$
  end for
  $p_n := p_n \,/\, R$
  $l := l + \log p_n$
end for
$\log P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \simeq l$

Algorithm 3: A "left-to-right" evaluation algorithm.

4.5. "Left-to-right" evaluation algorithm

Another approach for approximating $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$ was recently proposed by Wallach (2008). This method, which operates in an incremental, "left-to-right" fashion, decomposes $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$ as

$$\begin{aligned} &P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \\ &\quad = \textstyle\prod_n P(w_n \mid \boldsymbol{w}_{<n}, \Phi, \alpha\boldsymbol{m}) \\ &\quad = \textstyle\prod_n \sum_{\boldsymbol{z}_{\leq n}} P(w_n, \boldsymbol{z}_{\leq n} \mid \boldsymbol{w}_{<n}, \Phi, \alpha\boldsymbol{m}). \end{aligned} \quad (25)$$

Each sum over $\boldsymbol{z}_{\leq n}$ can then be approximated using an approach inspired by sequential Monte Carlo methods (Del Moral et al., 2006), as in algorithm 3. This method is appropriate for a wider range of applications—including predictive text entry and speech recognition systems—than the other methods in this section, because of its "left-to-right" operation.

4.6. Relative costs of the methods

The majority of the methods described above are based on Gibbs sampling, which dominates their costs: computing $P(z_n \mid w_n, \boldsymbol{z}_{\setminus n}, \Phi, \alpha\boldsymbol{m})$ is significantly more costly than computing $P(w_n \mid z_n, \Phi)$—the quantity used to construct the estimators given the samples. The Chib-style method is an exception: constructing the estimator itself has a cost roughly equal to that of Gibbs sampling. None-the-less, the approximate cost of each method can be reported in terms of the number of Gibbs sampling site updates required (i.e., the number of z_n variables updated) as shown in table 1.

Importance sampling using the prior over $\boldsymbol{\theta}$ as the sampling distribution does not involve Gibbs sampling. However, $\sum_{z_n} P(z_n, w_n \mid \boldsymbol{\theta}^{(s)}, \Phi)$ must be computed for each held-out token w_n, which has a similar cost to a Gibbs sampling site update. The cost of simple importance sampling using a distribution over $\boldsymbol{z}$ is harder to express, and will be implementation dependent. Slightly unfairly to these methods, we assume

initialize $0 = \tau_0 < \tau_1 < \ldots < \tau_S = 1$
sample $\boldsymbol{z}^{(1)}$ from $P_0(\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m})$.
for $s = 2 : S$ **do**
  sample $\boldsymbol{z}^{(s)} \sim T_{s-1}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s-1)})$
end for
$P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) \simeq \prod_{s=1}^{S} P(\boldsymbol{w}^{(2)} \mid \boldsymbol{z}^{(s)}, \boldsymbol{w}^{(1)}, \Phi)^{\tau_s - \tau_{s-1}}$

Algorithm 4: AIS for document completion.

that the cost of generating samples is directly comparable to Gibbs sampling. The cost could be examined more closely were such a method to yield good results.

5. Document completion

Another way of evaluating topic models is to compare predictive performance by estimating the probability of the second half of a document, given the first (Rosen-Zvi et al., 2004). This is typically accomplished by adding the first half of each held-out document to the training data, while retaining the second half for evaluation. Letting $\boldsymbol{w}^{(1)}$ be the first half of $\boldsymbol{w}$ and $\boldsymbol{w}^{(2)}$ be the second half, the goal is to compute

$$P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{w}^{(2)}, \boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})}, \quad (26)$$

which is a ratio of normalizing constants. Any of the methods for estimating $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \equiv P(\boldsymbol{w}^{(2)}, \boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})$ described in the previous section can be re-run on only $\boldsymbol{w}^{(1)}$ to estimate $P(\boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})$, thereby allowing evaluation of (26). However, specialized techniques may be more efficient.

5.1. Estimated $\boldsymbol{\theta}$

The estimated $\boldsymbol{\theta}$ method involves drawing samples $\boldsymbol{z}^{(1,s)} \sim P(\boldsymbol{z}^{(1)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m})$ and then forming

$$\hat{\theta}_t^{(s)} = P(t \mid \boldsymbol{z}^{(1,s)}, \alpha\boldsymbol{m}) = \frac{N_t^{(1,s)} + \alpha m_t}{N^{(1)} + \alpha}, \quad (27)$$

where $N^{(1)}$ is the number of tokens in $\boldsymbol{w}^{(1)}$. If the predictive probability of t is clamped to $\hat{\theta}_t^{(s)}$ for the remainder of the document, i.e., for $\boldsymbol{w}^{(2)}$, then

$$P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) \simeq \frac{1}{S} \sum_s \prod_n \sum_t \phi_{w_n^{(2)}|t}\, \hat{\theta}_t^{(s)}.$$

5.2. Importance sampling and AIS

The importance sampling algorithms described in sections 4.1 and 4.3 can all be adapted to estimate (26) directly by using samples conditioned on $\boldsymbol{w}^{(1)}$. For AIS, we use the following sequence of distributions:

$$P_s(\boldsymbol{z}) \propto P(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)} \mid \boldsymbol{z}, \Phi)^{\tau_s}\, P(\boldsymbol{z} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}).$$

(26) can then be approximated as in algorithm 4.

Table 1. Summary of methods for estimating $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$ with approximate costs for a document of length N. CS ($S = 1000$) and LR ($R = 20$) on a 200 word synthetic document each run in 3.2 seconds on a 1000MHz Opteron 175.

Method	Parameters	Description	Cost (# Gibbs site updates)
AIS	# temperatures S	Annealed importance sampling	SN
HM	burn-in B, # samples S	Harmonic mean method	$N(B+S)$
LR	# particles R	"Left-to-right" evaluation algorithm	$RN(N-1)\,/\,2$
CS	chain length S	Chib-style estimator	$2SN$
IS-PT	# samples S	Importance sampling from $P(\boldsymbol{\theta} \mid \alpha\boldsymbol{m})$	SN
IS-IP	# iterations I, # samples S	Importance sampling, $Q(\boldsymbol{z})$ from (11)	$(I+S)N$
IS-PZW	# samples S	Importance sampling, $Q(z_n) \propto \alpha m_{z_n} \phi_{w_n \mid z_n}$	SN

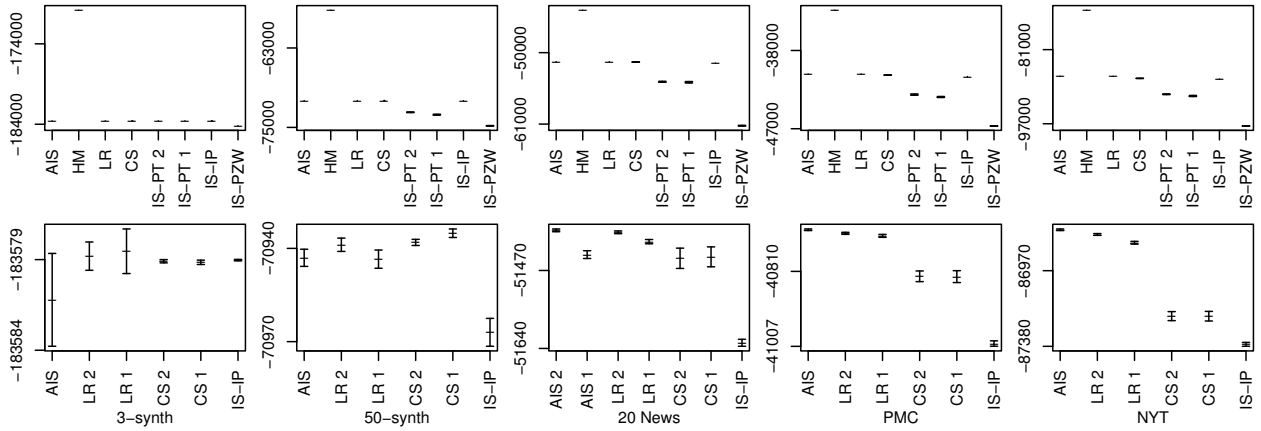

Figure 1. $\sum_d \log P(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$ for all five data sets. The top figures show results for all methods, while the bottom figures focus on the most competitive methods. To demonstrate convergence, in some cases, we report results for the same method twice, the second time with double the computation: e.g., "CS 1" uses $S = 1000$ while "CS 2" uses $S = 2000$.

5.3. "Left-to-right" evaluation algorithm

The "left-to-right" algorithm described in section 4.5 can estimate $P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m})$ directly. If the words in $\boldsymbol{w}$ are ordered such that $\boldsymbol{w}^{(1)}$ is fully observed before any words from $\boldsymbol{w}^{(2)}$ are observed, then a second estimator can be accumulated as in line 12 of algorithm 3, for the positions involving tokens in $\boldsymbol{w}^{(2)}$.

6. Results

In this section, we present experimental results comparing the evaluation methods described in the previous two sections on both real and synthetic data. Our MATLAB and Java implementations are available from `http://www.cs.umass.edu/~wallach/code/etm/`.

6.1. Description of data

Two synthetic data sets and three real-world corpora were used to compare the methods. The synthetic data sets were generated using two LDA models. In order to make the statistics of the synthetic documents as close as possible to real documents, the values of Φ, α and $\boldsymbol{m}$ were inferred from a collection of computer science abstracts using an MCMC implementation in the MALLET software package (McCallum, 2002).

Table 2. Data sets used in the experiments. V is the vocabulary size, $\bar{N}$ is the mean document length, "St. Dev." is the estimated standard deviation in document length.

Data set	V	$\bar{N}$	St. Dev.
Synthetic, 3 topics	9242	500	0
Synthetic, 50 topics	9242	200	0
20 Newsgroups	22695	120.4	296.2
PubMed Central abstracts	30262	101.8	49.2
New York Times articles	50412	230.6	250.5

Each of the three real-world data sets was divided into training and held-out documents. For each data set, Φ, α and $\boldsymbol{m}$ values were inferred using the training documents. Given these values, the evaluation methods were compared using the held-out documents. For all three data sets the number of topics was set to 200. Descriptions of all five data sets are given in table 2.

6.2. Estimating $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$

For each of the five data sets, $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$ was estimated for fifty held-out documents. The evaluation methods are summarized in table 1. Like Murray and Salakhutdinov (2009), AIS with 10,000 temperatures was intended as a gold standard. This method is computationally expensive, but is often accurate. For the harmonic mean method, $B = 50,000$ burn-in iterations were used, followed by $S = 50,000$ Gibbs sam-

pling iterations (without any thinning). The computation time for this parameterization roughly matches the computation time for AIS with 10,000 temperatures. For each of the data sets, the Chib-style estimator was run with two different parameterizations: $S = 1,000$ and $S = 2,000$. The remaining methods were also run with two parameterizations, chosen to result in computation times equivalent to those of the Chib-style estimator (which are much smaller than those of either AIS or the harmonic mean method).

For each method and data set, the estimates of $P(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$ for each held-out document $\boldsymbol{w}^{(d)}$ were averaged over 10 runs, to give $\bar{P}(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$. The log probability of the held-out documents was then estimated as $\sum_d \log \bar{P}(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$. Standard deviations were obtained using a bootstrap method, in which 10,000 log probabilities for each data set were obtained by sampling with replacement from the 10 runs for each held-out document. Figure 1 shows $\sum_d \log \bar{P}(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$ for each method and data set.

The harmonic mean method wildly overestimated $\sum_d \log P(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m})$ for all of the data sets. Estimates were effectively from one or very few samples, making the harmonic mean method very unstable.

The main failure mode for the simple importance sampling methods, AIS and the Chib-style estimator is inadequately sampling the upper tail of the distribution over importance weights. This causes underestimation of both $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})$ and the variance of the estimator. Consequently, the largest estimate is likely to be the best. Formal statements could be made using the bounds discussed by Gogate et al. (2007). AIS exhibited this failure mode on the 20 Newsgroups data set, yielding a lower probability than the "left-to-right" algorithm due to poor performance on some long documents. Increasing the accuracy by using 20,000 temperatures gave larger probabilities in agreement with the "left-to-right" algorithm. Long documents in the synthetic data set with only 3 topics also caused AIS to exhibit higher variance than other methods.

The Chib-style estimator and the "left-to-right" algorithm both performed well, with the latter consistently performing better on the real-world data sets. Results on the synthetic data sets show that the "left-to-right" algorithm does not universally dominate the Chib-style method, however. While more accurate than harmonic mean, none of the simpler importance sampling methods were competitive and generally performed exceedingly poorly, usually giving large underestimates. The "iterated pseudo-counts" method was the best of the simple importance sampling methods, but was still significantly worse than the Chib-style estimator.

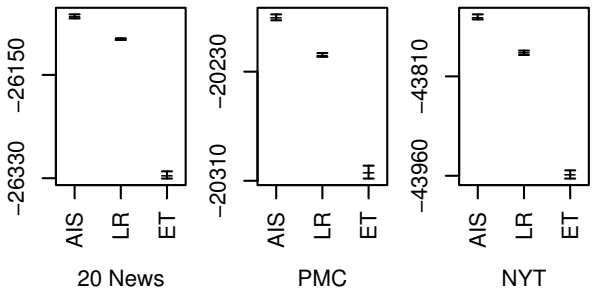

Figure 2. $\sum_d \log P(\boldsymbol{w}^{(2,d)} \mid \boldsymbol{w}^{(1,d)}, \Phi, \alpha\boldsymbol{m})$ for document completion methods. "ET" is the estimated $\boldsymbol{\theta}$ method.

6.3. Document completion

Another way of evaluating topic models is to estimate $P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m})$. As explained in section 5, this probability can be directly approximated using either the estimated $\boldsymbol{\theta}$ method, the "left-to-right" algorithm or AIS. These methods were compared using $B = 5,000$ burn-in iterations and $S = 20,000$ samples for the estimated $\boldsymbol{\theta}$ method, $B = 500$ burn-in iterations and $1,500$ temperatures for AIS, and $R = 4 \cdot 2000 \,/\, N$ particles for the "left-to-right" algorithm. Despite being allowed significantly more computation time than the other methods, the estimated $\boldsymbol{\theta}$ method exhibited relatively poor performance and did not achieve results close to those of AIS. The "left-to-right" algorithm did approach the estimates obtained using AIS, with substantially less computation time.

6.4. Sensitivity to perturbations in Φ and $\boldsymbol{m}$

In this section, we investigate how the evaluation methods described in sections 4 and 5 are affected by perturbations in Φ and $\boldsymbol{m}$. The methods were compared using the synthetic data set with 50 topics, for which the true Φ, α and $\boldsymbol{m}$ values are known.

Sensitivity to perturbations in Φ was investigated by interpolating between Φ and a randomly-generated set of topic-specific distribution over words Φ'. For each method, either $P(\boldsymbol{w} \mid (1-\lambda)\,\Phi + \lambda\Phi', \alpha\boldsymbol{m})$ or $P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, (1-\lambda)\,\Phi + \lambda\Phi', \alpha\boldsymbol{m})$ were calculated for $\lambda \in \{0, .25, .5, .75\}$. Figure 3 shows the log ratios between the values computed using interpolated parameters and the values computed using Φ only ($\lambda = 0$) for the estimated $\boldsymbol{\theta}$ method, importance sampling from $P(\boldsymbol{\theta} \mid \alpha\boldsymbol{m})$ and the "left-to-right" algorithm. Results for the Chib-style estimator closely track those of the "left-to-right" algorithm, so we report only the latter. Compared to the "left-to-right" algorithm, importance sampling and the estimated $\boldsymbol{\theta}$ method understate the difference between the values computed using the true Φ and the values computed using a perturbed Φ.

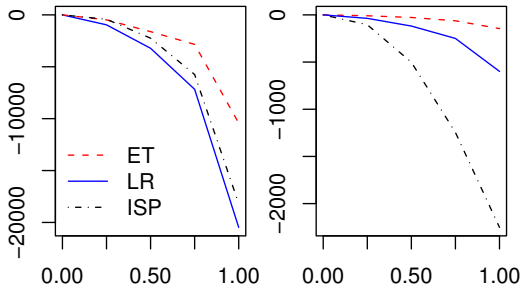

Figure 3. The effects of perturbing Φ (left) and $\boldsymbol{m}$ (right). The x-axis shows the degree of perturbation λ. The y-axis shows the log ratio between the probabilities reported by each estimator with the given λ and with $\lambda = 0$.

Sensitivity to perturbations in the base measure $\boldsymbol{m}$ was investigated similarly. A random $\boldsymbol{m}'$ was generated and either $P(\boldsymbol{w} \mid \Phi, \alpha((1-\lambda)\,\boldsymbol{m} + \lambda \boldsymbol{m}'))$ or $P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha((1-\lambda)\,\boldsymbol{m} + \lambda \boldsymbol{m}'))$ were calculated for $\lambda \in \{0, .25, .5, .75\}$. Log ratios between the values computed using interpolated base measures and the values computed using $\boldsymbol{m}$ are shown in figure 3. Importance sampling is strongly affected by perturbations in $\boldsymbol{m}$; the estimated $\boldsymbol{\theta}$ method is less sensitive.

7. Discussion

The evaluation methods currently used in the topic modeling community, including the harmonic mean method, importance sampling from $P(\boldsymbol{\theta} \mid \alpha \boldsymbol{m})$, and document completion methods, are generally inaccurate. Even if these methods do result in a correct ranking of different models, the relative advantage of one model over another may be incorrectly represented.

Most of the evaluation methods described in this paper extend readily to more complicated topic models—including non-parametric versions based on hierarchical Dirichlet processes (Teh et al., 2006)—since they only require a MCMC algorithm for sampling the latent topic assignments $\boldsymbol{z}$ for each document and a way of evaluating probability $P(\boldsymbol{w} \mid \boldsymbol{z}, \Phi, \alpha \boldsymbol{m})$. Importance sampling from $P(\boldsymbol{\theta} \mid \alpha \boldsymbol{m})$ is not obviously directly applicable to non-parametric topic models, however.

Estimating the probability of held-out documents provides a clear, interpretable metric for evaluating the performance of topic models relative to other topic-based models as well as to other non-topic-based generative models. We provide empirical evidence that several recently-used methods for estimating the probability of held-out documents are inaccurate and can change the results of model comparison. In contrast, the Chib-style estimator and "left-to-right" algorithm presented in this paper provide a clear methodology for accurately assessing and selecting topic models.

Acknowledgments

This work was supported by the Center for Intelligent Information Retrieval and CIA, NSA & NSF under NSF grant #IIS-0326249. Any opinions, findings and conclusions or recommendations are the authors' and do not necessarily reflect those of the sponsor.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *J. Machine Learning Res.*, *3*, 993–1022.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. American Stat. Assoc.*, *90*, 1313–1321.

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. Royal Stat. Soc. B*, *68*, 1–26.

Gogate, V., Bidyuk, B., & Dechter, R. (2007). Studies in lower bounding probability of evidence using the Markov inequality. *Proc. Conf. on Uncertainty in Artificial Intelligence* (pp. 141–148).

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci.*, *101*, 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Proc. Neural Information Processing Systems* (pp. 537–544).

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proc. Int'l. Conf. on Machine Learning* (pp. 577–584).

McCallum, A. (2002). MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Murray, I., & Salakhutdinov, R. (2009). Evaluating probabilities under high-dimensional latent variable models. *Proc. Neural Information Processing Systems* (pp. 1137–1144).

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*, 125–139.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Royal Stat. Soc. B*, *56*, 3–48.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proc. Conf. on Uncertainty in Artificial Intelligence* (pp. 487–494).

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *J. American Stat. Assoc.*, *101*, 1566–1581.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. *Proc. Int'l. Conf. on Machine Learning* (pp. 977–984).

Wallach, H. M. (2008). *Structured topic models for language*. PhD thesis, University of Cambridge.