
Annealing Paths for the Evaluation of Topic Models

James Foulds Padhraic Smyth

Department of Computer Science
University of California, Irvine
Irvine, CA 92697, USA
{jfoulds, smyth}@ics.uci.edu

Abstract

Statistical topic models such as latent Dirichlet allocation have become enormously popular in the past decade, with dozens of learning algorithms and extensions being proposed each year. As these models and algorithms continue to be developed, it becomes increasingly important to evaluate them relative to previous techniques. However, evaluating the predictive performance of a topic model is a computationally difficult task. Annealed importance sampling (AIS), a Monte Carlo technique which operates by annealing between two distributions, has previously been successfully used for topic model evaluation (Wallach et al., 2009b). This technique estimates the likelihood of a held-out document by simulating an annealing process from the prior to the posterior for the latent topic assignments, and using this simulation as an importance sampling proposal distribution.

In this paper we introduce new AIS annealing paths which instead anneal *from one topic model to another*, thereby estimating the *relative* performance of the models. This strategy can exhibit much lower empirical variance than previous approaches, facilitating reliable per-document comparisons of topic models. We then show how to use these paths to evaluate the predictive performance of topic model learning algorithms by efficiently estimating the likelihood at each iteration of the training procedure. The proposed method achieves better held-out likelihood estimates for this task than previous algorithms with, in some cases, an order of magnitude less computation.

1 INTRODUCTION

Topic models such as latent Dirichlet allocation (Blei et al., 2003) have become standard tools for analyzing text cor-

pora, with broad applications in areas such as political science (Grimmer, 2010), sociology (McFarland et al., 2013), conversational dialog (Nguyen et al., 2013), and more. A multitude of extensions to the LDA model have been developed for finding meaningful latent structure in text, along with a variety of strategies for fitting these models to increasingly large corpora.

As these new ideas continue to be proposed in the literature it becomes increasingly important to obtain accurate quantitative evaluations of the different approaches. Among the techniques available for evaluating topic models, the prediction of words in held-out documents (via test log-likelihood or perplexity) is perhaps the single most widely-used method for benchmarking the performance of new topic models and inference algorithms. An important point is that speedups for training these models do not necessarily translate to speedups in evaluating them. For example, there now exist very fast learning algorithms for training topic models based on approximate inference techniques, such as stochastic variational inference (Hoffman et al., 2010, 2013; Foulds et al., 2013), making it possible to learn topic models on corpora with millions of documents. Ironically, however, the time taken to compute test-set metrics for these algorithms can be orders of magnitude greater than the time it takes to train them. The evaluation of the predictive performance of topic models on held-out documents is still painfully slow, and relatively unreliable for individual documents as we will see later in the paper.

More specifically, consider a held-out document d , with word vector $w^{(d)}$, in the context of evaluating the quality of an LDA topic model (or one its many extensions). Given point estimates of topics Φ and a potentially document-specific Dirichlet prior α (if learned), we wish to compute the likelihood of the words in this held-out document, $Pr(w^{(d)}|\Phi, \alpha)$.¹ The direct computation of this quantity involves either an intractable sum over the latent topic assignments $z^{(d)}$, or an intractable integral over the distribution over topics $\theta^{(d)}$. Moreover, this already difficult computation must be performed for every document in

¹Or perplexity, a function of this and document length.

the held-out test set, which frequently contains hundreds to thousands of documents. To address this challenge, a wide variety of approximation strategies for estimating $Pr(w^{(d)}|\Phi, \alpha)$ have been proposed in papers such as those from Wallach et al. (2009b), Buntine (2009) and Scott & Baldridge (2013). Although these methods can lead to significantly more accurate results than naive approaches, the reliable and efficient evaluation of topic models remains a relatively open problem of practical significance.

In this paper we investigate new methods for evaluating topic models based on annealed importance sampling (AIS) (Neal, 2001), a Monte Carlo integration technique which was previously applied to topic model evaluation by Wallach et al. (2009b). Given two probability distributions, AIS produces an estimate of the ratio of their partition functions by annealing between them. Wallach et al. leverage this idea by annealing from the prior over the latent topic assignments $z^{(d)}$ to the posterior, resulting in an estimate of held-out document likelihood. AIS can be very accurate given enough computation time, although the amount of time needed may vary greatly between different choices of annealing paths (Grosse et al., 2013).

The first contribution of this paper is to propose and evaluate an alternative annealing strategy, using two AIS paths which anneal from one topic model to another. This strategy (referred to as ratio-AIS) computes the *ratio* of the likelihoods of two models instead of computing the likelihoods of each model separately. The result is an estimate of the relative performance of the models, with significantly lower empirical variance across runs than previous approaches.² This in turn brings computational benefits, as fewer samples or annealing temperatures may be required to achieve reliable results. The reduced variance comes at the cost of potentially increased bias when insufficient iterations are performed to achieve convergence. However, we also show how to detect such bias by annealing between the topic models in both directions and comparing the results. The consequence of this bias-variance trade-off is that the proposed method is useful in cases where we would like to perform in-depth analysis at the per-document level and when the two topic models are similar to each other. The previous high-variance low-bias methods may still be preferred for general full-corpus comparisons of topic models.

Finally, we show how to use the proposed AIS paths for evaluating topic model learning algorithms by computing held-out likelihood curves over the iterations of the learning procedure. This is achieved by annealing between the topic models at each iteration of the learning algorithm in turn, which allows all previous computation to be reused in each of the likelihood estimates. The proposed method outperforms previous algorithms, in some cases even when

it is given an order of magnitude less computation time. Note that although we focus on topic models, the ideas presented here could potentially also be useful for other latent variable models with intractable likelihoods.

2 BACKGROUND

When proposing a new topic model or learning algorithm, it is important to evaluate its performance. When the model is to be used for a certain task it may be possible to evaluate it with respect to an extrinsic, task-specific metric. For example one could evaluate the quality of topics being used as features for a classification algorithm by measuring classification accuracy. More generally, however, given that topic models are generally trained in an unsupervised manner (with a few notable exceptions), a ground-truth evaluation metric is typically not available.

Consequently, a number of intrinsic (i.e. task independent) validation strategies for topic models have been developed in the literature. For example, Chang et al. (2009) proposed the use of elicitation of judgments from humans to evaluate the quality of topic models. Given that obtaining these judgments can be expensive and difficult, Newman et al. (2010) and Mimno et al. (2011) proposed automatic surrogate measures of topic coherence, and showed that these measures, which are typically based on word co-occurrence statistics, are correlated with human judgments.

However, as topic models are statistical models, we also would like to be able to evaluate them as such. In the context of unsupervised machine learning, the standard approach for evaluating a statistical model is to compute the probability of held-out data. Regardless of the utility of the aforementioned methods, it is generally useful to demonstrate good predictive performance in addition to any other extrinsic or intrinsic validation results. Intuitively, as our goal is to fit a statistical model to data, we would like to know both how well we are able to fit the model, and how well the model is able to explain unobserved data.

As in Wallach et al. (2009b), we therefore focus on the computation of $Pr(w^{(d)}|\Phi, \alpha)$, the likelihood of the words $w^{(d)}$ in a held-out document d (or equivalently, perplexity), conditioned on point estimates of the topic-word distributions Φ and (possibly document-specific) priors α , where Φ is a $W \times K$ matrix consisting of K discrete distributions $\Phi^{(k)}$ over the W words in the dictionary, and α is a K -dimensional Dirichlet parameter vector.³ This quantity can be used to evaluate a point estimate of the topics, or in an inner loop to evaluate Bayesian evaluation metrics such as the posterior predictive probability of held-out documents.

²“Variance” here refers to variance across Monte Carlo estimates of the difference in log-likelihood between models, per document.

³It is standard practice to learn an asymmetric Dirichlet prior α in LDA models, following Wallach et al. (2009a), so we include it as a parameter to evaluate. The prior may also be learned in a document dependent way for models such as DMR (Mimno & McCallum, 2008).

It is in general infeasible to compute $Pr(w^{(d)}|\Phi, \alpha)$ directly, as it involves an intractable sum $\sum_z^{(d)} Pr(w^{(d)}, z^{(d)}|\Phi, \alpha)$ or an intractable integral $\int_{\theta} Pr(w^{(d)}, \theta^{(d)}|\Phi, \alpha)$. The computational difficulty arises because the topic assignments $z^{(d)}$ and distributions over topics $\theta^{(d)}$ for the held-out document are unknown, and so all possible values must be considered. A variety of approximation strategies were considered by Wallach et al. (2009b), the number of which alone is a testament to the difficulty of the problem. The most widely used of these approaches is the “left-to-right” particle filtering algorithm. In the algorithm, a number of particles are maintained, representing topic assignments up to the current word t in the document. In each iteration, these particles are used to draw samples of the topic assignment for the next word $t + 1$, conditioned on the previous words and topic assignments. A resampling step is also performed, making the algorithm’s run time quadratic in the length of the document. The algorithm was analyzed more closely by Buntine (2009), and a faster, but less accurate, variant of the technique was proposed by Scott & Baldridge (2013).

Alternatively, a strategy for side-stepping some of the computational difficulty is to instead estimate (or sample) $z^{(d)}$ or $\theta^{(d)}$ on a subset $w^{(d,1)}$ of the document, and predict only the remaining portion of the document $w^{(d,2)}$, thus estimating $Pr(w^{(d,2)}|w^{(d,1)}, \Phi, \alpha)$. This method is frequently used in practice (e.g. Rosen-Zvi et al. (2004); Wallach et al. (2009a)). However, this “document completion” scenario changes the task somewhat, and is not the gold standard prediction task we would like it to be. It measures the ability of the model to “orient” itself quickly when given partial documents, rather than how likely the overall document is under the model. The widespread use of the document completion strategy may be largely due to its convenient computational properties (leading in turn to its use as a surrogate for fully held-out prediction), rather than being due to any intrinsic benefit of the metric itself.

It is also unclear how the use of document completion as a surrogate for full-document prediction might affect our conclusions, particularly when using topic models which learn the Dirichlet hyper-parameter α as in Wallach et al. (2009a) and Mimno & McCallum (2008). Learning α may help the model to recover $\theta^{(d)}$ better on the training portion of the document, thus increasing the performance of the model for document completion more than in the fully held-out case.

On the other hand, observing more of the document decreases the relative impact of the prior on the posterior distribution, which could reduce the observed improvement due to learning α . Thus, we suspect that document completion may not always be a good surrogate for the full prediction task. It should be noted that many methods for fully held-out prediction can also be adapted for document completion (including those proposed here).

2.1 ANNEALED IMPORTANCE SAMPLING

One of the more accurate strategies investigated by Wallach et al. (2009b) to estimate held-out likelihood was annealed importance sampling (AIS) (Neal, 2001). AIS is a general technique for estimating an expectation of a function of a random variable x with respect to an intractable distribution of interest p_0 . Consider a distribution p_n (which is typically easy to sample from) and a sequence of “intermediate” distributions p_{n-1}, \dots, p_1 leading from p_n to p_0 . AIS works by annealing from p_n towards p_0 by way of the intermediate distributions, and using importance weights to correct for the fact that an annealing process was used instead of sampling directly from p_0 .

Assume that for each intermediate distribution p_j we have a Markov chain with transition operator $T_j(x, x')$ which is invariant to that distribution. We need to be able to sample from these Markov chains, and for each p_j be able to evaluate some function f_j which is proportional to it. In a manner similar to that of traditional importance sampling, AIS produces a collection of samples $x^{(1)}, \dots, x^{(S)}$ with associated importance weights $w^{(1)}, \dots, w^{(S)}$. As with importance sampling, the expectation of interest is estimated using the samples, weighted by the importance weights.

The strategy for drawing each sample $x^{(i)}$ is to begin by drawing a sample x_{n-1} from p_n , then drawing a sequence of points x_{n-2}, \dots, x_0 which “anneal” towards p_0 . Each of the remaining x_j ’s in the sequence are generated from x_{j+1} via T_j . Importance weights $w^{(i)}$ are computed by viewing (x_0, \dots, x_{n-1}) as an augmented state space, and performing importance sampling on this new state space. The above procedure is used as a proposal distribution Q for importance sampling from another distribution P :

$$Q(x_0, \dots, x_{n-1}) \propto f_n(x_{n-1}) \prod_{j=n-1}^1 T_j(x_j, x_{j-1})$$

$$P(x_0, \dots, x_{n-1}) \propto f_0(x_0) \prod_{j=1}^{n-1} \tilde{T}_j(x_{j-1}, x_j),$$

where $\tilde{T}_j(x, x') = T_j(x', x) \frac{f_j(x')}{f_j(x)}$ is the reversal of the transition defined by T_j . This leads to importance weights for each of the samples,

$$w^{(i)} = \frac{P(x_0, \dots, x_{n-1})}{Q(x_0, \dots, x_{n-1})} = \prod_{j=0}^{n-1} \frac{f_j(x_j)}{f_{j+1}(x_j)}. \quad (1)$$

Note that the marginal probability of x_0 under P is $p_0(x_0)$, so after letting $x^{(i)} = x_0$ the procedure correctly carries out importance sampling from p_0 . AIS also provides an estimate for the ratio of normalizing constants for f_0 and f_n . The normalizing constant for P is the same as the normalizing constant for f_0 , and the normalizing constant for Q is the same as the normalizing constant for f_n , and so

the average of the importance weights, $\frac{\sum w^{(i)}}{N}$, converges to $\frac{\int f_0(x)dx}{\int f_n(x)dx}$.

2.2 AIS FOR TOPIC MODELS

Wallach et al. (2009b) showed how to apply the AIS procedure to the problem of calculating LDA likelihoods. The likelihood of a test document for a topic model can be estimated by using AIS to estimate a normalization constant, operating on the latent topic assignments $z^{(d)}$ for the document.⁴ We can set $f_0 = Pr(w^{(d)}, z^{(d)}|\Phi, \alpha)$, $f_n = Pr(z^{(d)}|\alpha)$, with intermediate distributions $f_j = Pr(w^{(d)}|z^{(d)}, \Phi, \alpha)^{\beta_j} f_n$ and the transition operators T_j being the Gibbs sampler for f_j . The ratio of normalizing constants is

$$\begin{aligned} \frac{\sum w^{(i)}}{S} &\approx \frac{\sum_{z^{(d)}} Pr(w^{(d)}, z^{(d)}|\Phi, \alpha)}{\sum_{z^{(d)}} Pr(z^{(d)}|\alpha)} \\ &= \frac{Pr(w^{(d)}|\Phi, \alpha)}{1} = Pr(w^{(d)}|\Phi, \alpha). \end{aligned} \quad (2)$$

The procedure for producing each importance sample, then, is to draw an initial $z^{(d)}$ from the prior, and anneal it towards f_0 by performing $r_j \geq 1$ Gibbs iterations at each intermediate distribution. After repeating this procedure for each sample, the likelihood is estimated as the average of the importance weights. Note that in what follows we define a *run* as the full procedure averaging over importance samples, while a *sample* refers to a single importance sample.

3 ALTERNATIVE ANNEALING PATHS FOR THE EVALUATION OF TOPIC MODELS

The AIS method described above can be very accurate if given enough computation time (Wallach et al., 2009b). However, it is subject to several potentially avoidable sources of variability. Firstly, the method estimates the ratio of the desired quantity $Pr(w^{(d)}|\Phi, \alpha)$ and the denominator $\sum_{z^{(d)}} Pr(z^{(d)}|\alpha)$ in Equation 2, which equals one, introducing stochastic noise on behalf of the denominator even though this is a constant. We would also expect that the prior may typically be very different from the posterior, thereby requiring many annealing iterations to prevent the importance weights $w^{(i)}$ from having a large variance. This has consequences for the efficiency of the sampler, which is reduced by a factor of approximately

$1 + \text{Var}_q[w^{(i)} / E_q[w^{(i)}]]$ relative to direct sampling from the target density (Neal, 2001).⁵

Making matters worse, we typically must perform the AIS procedure many times across all held-out documents, and therefore have a relatively limited computational budget per document, preventing us from compensating for the high variance by collecting many importance samples with a large number of temperatures. In this section, we introduce new AIS annealing paths for the evaluation of topic models which can have lower variance than the standard approach. We first introduce AIS paths which compare two topic models by annealing between them. We then show how to use these paths for evaluating topic model learning algorithms by computing per-iteration predictive performance efficiently, reusing all previous computation.

3.1 COMPARING TOPIC MODELS BY ANNEALING BETWEEN THEM

The most typical evaluation scenario is model comparison—we want to determine whether a particular model (model 1) performs better at predicting held-out documents than a baseline method (model 2) such as vanilla LDA or a model trained using a previous learning algorithm. Thus, in such situations, the quantity of interest is the *relative* log-likelihood score of the model and the baseline:

$$\begin{aligned} &\log Pr(w^{(d)}|\Phi^{(1)}, \alpha^{(1)}) - \log Pr(w^{(d)}|\Phi^{(2)}, \alpha^{(2)}) \\ &= \log \frac{Pr(w^{(d)}|\Phi^{(1)}, \alpha^{(1)})}{Pr(w^{(d)}|\Phi^{(2)}, \alpha^{(2)})}. \end{aligned} \quad (3)$$

To compute this in the framework proposed by Wallach et al., we must perform the AIS procedure once for each model, incurring the stochastic error twice. To avoid this and the aforementioned sources of variability with that approach, and given that the procedure is already designed to compute a ratio, we propose to instead use AIS to compute Equation 3 directly. Let $f_0(z^{(d)}) = Pr(w^{(d)}, z^{(d)}|\Phi^{(1)}, \alpha^{(1)})$ and $f_n(z^{(d)}) = Pr(w^{(d)}, z^{(d)}|\Phi^{(2)}, \alpha^{(2)})$. Then the desired quantity can be estimated via

$$\begin{aligned} \sum \frac{w^{(i)}}{N} &\approx \frac{\sum_{z^{(d)}} Pr(w^{(d)}, z^{(d)}|\Phi^{(1)}, \alpha^{(1)})}{\sum_{z^{(d)}} Pr(w^{(d)}, z^{(d)}|\Phi^{(2)}, \alpha^{(2)})} \\ &= \frac{Pr(w^{(d)}|\Phi^{(1)}, \alpha^{(1)})}{Pr(w^{(d)}|\Phi^{(2)}, \alpha^{(2)})}. \end{aligned} \quad (4)$$

We will refer to this strategy as “ratio-AIS.” To implement this method, it remains to choose the annealing path, i.e. the sequence of intermediate distributions. We first consider a geometric average $f_j(z^{(d)}) =$

⁴The derivation here differs slightly from that of Wallach et al. (2009b). The present derivation suggests that the procedure described in Wallach et al. produces just one importance sample. This may be repeated, finally producing as output the average of the resulting importance weights. In practice however, we found that a single sample with a longer annealing run, as in Wallach et al., may still be the best strategy on a computational budget.

⁵Note that $E_q[w^{(i)}]$ is equal to the ratio of normalizing constants of the target and proposal densities, which in our case is the quantity of interest, e.g. the likelihood.

$f_0(z^{(d)})^{\beta_j} f_n(z^{(d)})^{1-\beta_j}$ of the initial and final distributions, a strategy suggested by Neal (2001) with analogy to simulated annealing, where β_j can be viewed as an “inverse temperature.” To choose a transition operator T_j invariant to f_j , we straightforwardly select the Gibbs sampler. We have importance weights

$$\begin{aligned} w^{(i)} &= \prod_{j=0}^{n-1} \frac{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(1)}, \alpha^{(1)})^{\beta_j}}{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(1)}, \alpha^{(1)})^{\beta_{j+1}}} \\ &\times \prod_{j=0}^{n-1} \frac{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(2)}, \alpha^{(2)})^{1-\beta_j}}{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(2)}, \alpha^{(2)})^{1-\beta_{j+1}}} \\ &= \prod_{j=0}^{n-1} \frac{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(1)}, \alpha^{(1)})^\tau}{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(2)}, \alpha^{(2)})^\tau} \\ \log w^{(i)} &= \frac{1}{n} \sum_{j=0}^{n-1} \log \frac{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(1)}, \alpha^{(1)})}{Pr(w^{(d)}, z_j^{(d)} | \Phi^{(2)}, \alpha^{(2)})}, \quad (5) \end{aligned}$$

assuming $\beta_j - \beta_{j+1} = \tau = n^{-1} \forall j, 0 \leq j < n-1$. Elegantly, the log importance weights are the average of the log ratios of the probabilities of $w^{(d)}$ and $z^{(d)}$ according to each model. Observe that the same z assignments are used for the numerator and denominator in each of the ratios in Equation 5, further reducing the variance of the estimate relative to the standard AIS strategy.

Although geometric averages are the standard choice for an annealing path, in many cases there exist annealing paths which perform much better. Grosse et al. (2013) introduced an alternative annealing path for exponential families which converges much more quickly, constructed by annealing averages of the moments of the sufficient statistics. The Dirichlet-multinomial distribution $Pr(z^{(d)} | \alpha)$ is not an exponential family so their method does not directly apply to LDA. Nevertheless, we consider an annealing path inspired by their work, where intermediate distributions are constructed by taking convex combinations of the parameters:

$$\begin{aligned} f_j(z^{(d)}) &= Pr(w^{(d)}, z^{(d)} | \Phi_j = \beta_j \Phi^{(1)} + (1 - \beta_j) \Phi^{(2)}, \\ \alpha_j &= \beta_j \alpha^{(1)} + (1 - \beta_j) \alpha^{(2)}). \quad (6) \end{aligned}$$

The intermediate distributions are topic models, so we set T_j to be the corresponding Gibbs sampler. This T_j does not require power operations, providing substantial execution time savings over the geometric path and Equation 2. The importance weights are

$$\begin{aligned} \log w^{(i)} &= \sum_{j=0}^{n-1} \left(\log Pr(w^{(d)}, z_j^{(d)} | \Phi_j, \alpha_j) \right. \\ &\quad \left. - \log Pr(w^{(d)}, z_j^{(d)} | \Phi_{j+1}, \alpha_{j+1}) \right). \quad (7) \end{aligned}$$

To implement this method we need to draw initially from $f_n(z^{(d)})$, which we accomplish via Gibbs sampling. These

initial samples from $f_n(z^{(d)})$ need not be independent for the procedure to work, although we may choose to run independent chains if the cost of burn-in is deemed to be less than the time wasted due to running the annealing on correlated samples. Finally, AIS will be more likely to converge if the initial and target distributions are similar to each other. We therefore align the topics before running the algorithm, using the Hungarian algorithm to minimize the L1 distances between topics. This operation, which is $O(K^3)$, where K is the number of topics, is not a computational bottleneck (relative to performing AIS) and needs only to be performed once per corpus. Pseudo-code for ratio-AIS using the path from Equation 6 is given in Algorithm 1.

Algorithm 1 Ratio-AIS, using the convex path

```

for  $i = 1 : S$  //importance samples
   $\log \omega[i] := 0$ 
   $\Phi^{(next)} := \Phi^{(2)}$ 
   $\alpha^{(next)} := \alpha^{(2)}$ 
  draw  $z^{(i)} \sim Pr(z | \alpha^{(2)})$ 
  for  $j = n-1, n-2, \dots, 0$  //temperatures
     $\Phi^{(curr)} := \Phi^{(next)}$ 
     $\alpha^{(curr)} := \alpha^{(next)}$ 
     $\Phi^{(next)} := \beta_j \Phi^{(1)} + (1 - \beta_j) \Phi^{(2)}$ 
     $\alpha^{(next)} := \beta_j \alpha^{(1)} + (1 - \beta_j) \alpha^{(2)}$ 
    for  $a = 1 : r_j$  //  $r_{n-1}$  is large, for burn in
      for  $l = 1 : \text{length}(w^{(d)})$  //words
        //draw  $z_l^{(i)}$ 
         $Pr(z_l^{(i)} = k | \cdot) \propto (n_k^{(i)} + \alpha_k^{(curr)}) \Phi_{w_l^{(d)}, k}^{(curr)}$ 
       $\log \omega[i] := \log \omega[i]$ 
       $+ \log Pr(w^{(d)}, z^{(i)} | \Phi^{(next)}, \alpha^{(next)})$ 
       $- \log Pr(w^{(d)}, z^{(i)} | \Phi^{(curr)}, \alpha^{(curr)})$ 
  return  $\log \text{SumExp}(\log \omega) - \log(S)$ 

```

Detecting Convergence Failures

AIS can produce poor estimates if the annealing fails to converge to a high-probability state in the target distribution within the given set of iterations. In general, this may be very difficult to detect. However, in our case we can interchange f_0 and f_n in our AIS strategy to compute the reciprocal of the desired ratio, and compare the reciprocal of this to our estimate. If these two values are wildly different, then we will know that the annealing has failed to converge. This means that we are able to detect convergence failures in many practical cases. In our experiments, we were easily able to catch convergence failures by observing a systematic bias across documents in the results of the different annealing directions (see Section 4).

3.2 EFFICIENTLY EVALUATING TOPIC MODEL LEARNING ALGORITHMS WITH ITERATION-AIS

When evaluating algorithms for learning topic models (or monitoring their convergence), we would ideally like to compute and plot held-out log-likelihood scores per learning iteration (or unit of computation time) for each algorithm under consideration. This is extremely expensive, requiring $|H| \times I \times M$ Monte Carlo approximations of already intractable high-dimensional integrals, where H is the held-out test set, I is the number of iterations of the learning algorithms to evaluate at, and M is the number of competing learning methods.

Fortunately, for many learning algorithms such as the collapsed Gibbs sampler, the topics at successive iterations are similar to each other, and the topics typically vary smoothly from “high temperature” high entropy distributions at early iterations to more complicated later distributions. This suggests using a single AIS path to perform the entire evaluation across all of the iterations, with the topic models at each iteration (or a subset of them) as intermediate distributions. We can accomplish this by annealing from the prior $Pr(z^{(d)}|\alpha^{(1)})$ to the first topic model $Pr(w^{(d)}, z^{(d)}|\Phi^{(1)}, \alpha^{(1)})$ as in Wallach et al. (2009b), and then using ratio-AIS to anneal between successive topic models $Pr(w^{(d)}, z^{(d)}|\Phi^{(t)}, \alpha^{(t)})$. For the topic model at iteration t , the average $S^{-1} \sum_i w^{(i,t)}$ of the importance weights computed up to that point $w^{(i,t)}$ converges to the ratio of normalizing constants,

$$\frac{\sum_{z^{(d)}} Pr(w^{(d)}, z^{(d)}|\Phi^{(t)}, \alpha^{(t)})}{\sum_{z^{(d)}} Pr(z^{(d)}|\alpha^{(1)})} = Pr(w^{(d)}|\Phi^{(t)}, \alpha^{(t)}). \quad (8)$$

With n temperatures per learning iteration k , importance weights can be written recursively as

$$\begin{aligned} \log w^{(i,t)} &= \sum_{t'=1}^t \sum_{j=0}^{n-1} \log \frac{f_{t',j}(z_{t',j})}{f_{t',j+1}(z_{t',j})} \\ &= \log w^{(i,t-1)} + \sum_{j=0}^{n-1} \log \frac{f_{t,j}(z_{t,j})}{f_{t,j+1}(z_{t,j})}. \end{aligned} \quad (9)$$

This method, which we refer to as *iteration-AIS*, exploits all of the computation for selecting z assignments and importance weights from the likelihood estimates at previous learning iterations, leading to successively longer annealing runs, and therefore potentially better likelihood estimates, as k increases.

4 EXPERIMENTS

We explored the performance of the proposed techniques using a corpora of scientific articles from the Association

of Computational Linguistics (ACL) conference⁶ (Radev et al., 2013), and another from the Neural Information Processing Systems (NIPS) conference.⁷ The ACL dataset consists of the 3286 articles from the years 1987 to 2011, while the NIPS corpus contains the 1740 articles published between 1987 and 1999. In each experiment, topic models with 50 topics were fit to each corpus by performing 1000 iterations of collapsed Gibbs sampling using the MALLET toolkit (McCallum, 2002). Roughly 10% of the documents in each corpus were withheld for testing (130 NIPS articles, and 300 ACL articles). Although cross-validation would have been a preferable option to using a single hold-out set, the computational expense of the experiments prevented this. For example, across all algorithms and learning iterations, Figure 2 required a total of 6.6 million Gibbs iterations for each one of the 300 test articles.

When using AIS we must select the number of temperatures n , the number of importance samples S , and the temperature schedule $\beta_0, \beta_1, \dots, \beta_n$. The variability of an AIS estimator can be reduced by increasing S (due to the law of large numbers) or by increasing n (which reduces the variance of the $w^{(i)}$). In the experiments, we focused on the case where $S = 1$, as in Wallach et al. (2009b). We found in preliminary experiments that $S = 1$ gave essentially exactly the same answer as $S = 100$ importance samples for Ratio-AIS with 10,000 temperatures. For simplicity, we used a uniform spacing of the temperatures β_j .⁸

We also compared to the left-to-right (LR) particle filtering algorithm of Wallach et al. (2009b), using the implementation provided in MALLET. The left-to-right method requires $N_d(N_d + 1)/2$ word-level Gibbs updates per particle for a document of length N_d . The execution of $p = 2 \times n/(N_d + 1)$ particles corresponds to the same number of Gibbs updates as AIS with n temperatures and $S = 1$. We select the number of LR particles by rounding p to the nearest integer greater than zero.

Ratio-AIS was designed for reliable per-document comparisons. To explore this, we ran each algorithm twice on each document, and reported results comparing the two runs across documents. To remove the effect of document length in the results, instead of reporting the differences in log-likelihood scores for each model we consider instead perplexity scores $\exp(-\frac{\log Pr(w^{(d)}|\Phi, \alpha)}{N_d})$. The ratio of the perplexity of model 1 over the perplexity of model 2 for a document is easily computed from the output of Ratio-AIS as $\exp(\frac{L_2 - L_1}{N_d})$, where L_j is the log-likelihood for model j . We considered two evaluation scenarios: comparing learned topics to perturbed versions of the same top-

⁶Available at <http://clair.eecs.umich.edu/aan/index.php>.

⁷The NIPS dataset, due to Gregor Heinrich, is available at <http://www.arbylon.net/resources.html>.

⁸Neal (2001) suggests that a geometric spacing of the β_j 's may be beneficial, at least for the geometric annealing path.

ics (Section 4.1), and comparing topic models learned with symmetric and asymmetric Dirichlet priors (Section 4.2). Finally, we evaluated iteration-AIS for estimation of per-iteration likelihood (Section 4.3).

4.1 LEARNED TOPICS VERSUS PERTURBED TOPICS

As the likelihoods we are trying to estimate are intractable, we do not in general have access to ground truth. However, after learning topics Φ on a dataset and then creating a noisy copy of them Φ' , we have good reason to believe that the original topics Φ are better than the copy. This style of experiment was previously performed by Wallach et al. (2009b). We took the word-topic assignments learned by MALLET, and created Φ' by re-assigning 5% of them to new word-topic assignments uniformly at random.⁹

Ratios of the perplexities for the two models were computed with both cheap (100 temperatures) and expensive runs (10,000 temperatures). Overall results are given in Table 1, and per-document results for the ACL dataset are plotted in Figure 1.¹⁰ The two ratio-AIS paths were both the most accurate and the most consistent methods, in both temperature regimes.

In the cheap regime, the ratio-AIS points are slightly off-diagonal in Figure 1, with one annealing direction giving systematically lower results, representing a detectable bias due to convergence failure in at least one annealing direction. Nevertheless, these results have much lower variance, and the bias disappears in the expensive regime. Surprisingly, the standard AIS method performed extremely poorly, with most data points falling outside of the boundaries of the figure, which are tight around the results of the other methods. Using many importance samples would very likely mitigate this, at greater computational cost. It should be noted that the task of comparing two very similar topic models is difficult for standard methods, but is relatively easy for ratio-AIS due to the distance to anneal between the distributions being smaller.

4.2 SYMMETRIC VERSUS ASYMMETRIC DIRICHLET PRIORS

Learning asymmetric α hyperparameters can improve the predictive performance of topic models (e.g., Wallach et al. (2009a)). To explore this, on each corpus we learned a topic model with asymmetric α , and a model where α was fixed to be flat but its concentration parameter was learned. The AIS and LR algorithms were used to compare the resulting models, using runs with 1000 temperatures and 10,000 temperatures.

⁹MALLET’s left-to-right takes as input a count matrix, so the perturbed topics must be representable as counts.

¹⁰Results for the NIPS corpus are similar, and are provided in Foulds (2014).

It was found that in the “cheap” 1000 temperature regime, the ratio-AIS estimates were the most closely correlated with left-to-right estimates in the expensive regime, the best available proxy for ground truth (Table 2, top).¹¹ In all cases the ratio-AIS paths had one to two orders of magnitude lower empirical variances in the estimates of per-document perplexity ratios than the previous methods, with the convex path having the least variance (Table 2, middle). Ratio-AIS therefore achieves the original goal of greatly reducing the variance of per-document comparisons of topic models. This is particularly important if we want to perform detailed analysis at a per-document level, such as exploring the effect of covariates on topic model performance. In such a scenario, the previous methods have unacceptably high variance for a reasonable level of computation (see also Figure 1), while the ratio-AIS estimates of relative performance have very small empirical variance with just one importance sample.

Unfortunately, this reduction comes at a price of potentially increased bias in the estimated perplexity ratio when given insufficient computation. Topic models which learn an asymmetric α tend to perform better than those with a symmetric α (Wallach et al., 2009a), and the previous methods detected a larger advantage for the asymmetric approach (Table 2, bottom). The direction of the ratio-AIS annealing path also made a difference to the outcome. In particular, the forward direction of annealing did not detect an overall advantage to the asymmetric hyper-parameter model. On the other hand, the difference per direction allowed us to detect a convergence failure, which is difficult to do in general. Also note that for the task in Section 4.1, the overall perplexity ratios were very consistent between annealing directions, and showed a clearer difference between models than the baseline algorithms did – see Foulds (2014).

4.3 EVALUATING TOPIC MODELS PER ITERATION

The iteration-AIS annealing path evaluates the performance of topic model learning algorithms on a per-iteration basis. We explored its performance using the convex path with 1000 and 10,000 temperatures per learned model, annealing between the models at every 10th learning iteration. At the first learning iteration $\Phi^{(1)}$, the algorithms were given an extra 1000 temperatures to compensate for the cold-start from the prior.

Results on ACL are shown in Figure 2. It was found that iteration-AIS estimated higher log-likelihoods than left-to-right and standard AIS in both temperature regimes (Figure 2, left). The main failure mode of these algorithms is to underestimate the likelihood by failing to find high probability regions, so higher values are likely to be better

¹¹The standard AIS estimate of the perplexity ratios had too high a variance to be used (see Table 2).

% Correct	Left to Right	Standard AIS	Ratio-AIS Geometric	Ratio-AIS Geom. (reverse)	Ratio-AIS Convex	Ratio-AIS Convex (reverse)
NIPS (cheap)	63.8	48.8	83.8	89.2	84.6	87.7
NIPS (expensive)	84.6	62.3	86.9	87.7	87.7	87.7
ACL (cheap)	80.2	50.8	88.3	92.0	88.3	92.3
ACL (expensive)	90.7	75.2	90.3	90.3	90.3	90.3

Table 1: Percentage of documents where the learned topics Φ were estimated to have higher likelihood than the perturbed topics Φ' .

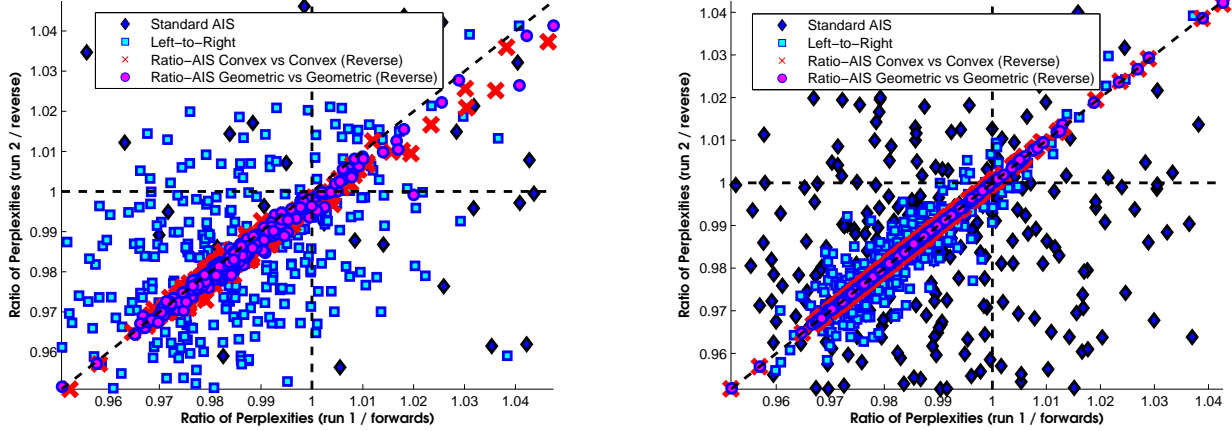


Figure 1: Comparing learned topics with perturbed versions of them, on the ACL dataset. In the figures, every point corresponds to a document. Each axis corresponds to estimated $\frac{\text{perp}(\Phi)}{\text{perp}(\Phi')}$ for a repeat of the experiment, with the ratio-AIS repeats being performed in different annealing directions. Points in the lower left quadrant are those which (likely correctly) predict the unperturbed topics as the winner in both trials. Points near the diagonal have consistent results across the two trials. **Left**: 100 temperatures. **Right**: 10,000 temperatures. Missing Standard AIS results are outside of the bounds of the plots. Figure best viewed in color.

Correlation with Long LR Run	Left to Right	Standard AIS	Ratio-AIS Geometric	Ratio-AIS Geom. (reverse)	Ratio-AIS Convex	Ratio-AIS Convex (reverse)
NIPS (cheap)	0.947	0.619	0.973	0.975	0.976	0.981
NIPS (expensive)	0.993	0.852	0.981	0.982	0.981	0.982
ACL (cheap)	0.965	0.578	0.984	0.983	0.987	0.986
ACL (expensive)	0.995	0.892	0.989	0.989	0.990	0.989

Variance of Perplexity Ratio	Left to Right	Standard AIS	Ratio-AIS Geometric	Ratio-AIS Geom. (reverse)	Ratio-AIS Convex	Ratio-AIS Convex (reverse)
NIPS (cheap)	2.6×10^{-4}	2.6×10^{-3}	2.0×10^{-5}	1.5×10^{-5}	8.2×10^{-6}	9.8×10^{-6}
NIPS (expensive)	1.7×10^{-5}	6.0×10^{-4}	1.4×10^{-6}	1.2×10^{-6}	6.9×10^{-7}	5.8×10^{-7}
ACL (cheap)	1.7×10^{-4}	3.6×10^{-3}	1.6×10^{-5}	1.3×10^{-5}	7.7×10^{-6}	6.6×10^{-6}
ACL (expensive)	1.4×10^{-5}	5.6×10^{-4}	1.1×10^{-6}	9.4×10^{-7}	7.4×10^{-7}	5.1×10^{-7}

Corpus-Level Perplexity Ratio	Left to Right	Standard AIS	Ratio-AIS Geometric	Ratio-AIS Geom. (reverse)	Ratio-AIS Convex	Ratio-AIS Convex (reverse)
NIPS (cheap)	0.984	0.975	1.01	0.992	1.01	0.994
NIPS (expensive)	0.989	0.990	1.00	0.999	1.00	0.998
ACL (cheap)	0.984	0.980	1.00	0.985	1.00	0.988
ACL (expensive)	0.987	0.989	0.994	0.992	0.996	0.992

Table 2: Comparing asymmetric α and symmetric α topic models. Correlation coefficient with the perplexity ratio estimates from a run of left-to-right in the expensive regime (**top**), average empirical variance (evaluated across two runs per document) of the per-document perplexity ratio (**middle**), and the overall perplexity ratio for the entire corpus (**bottom**).

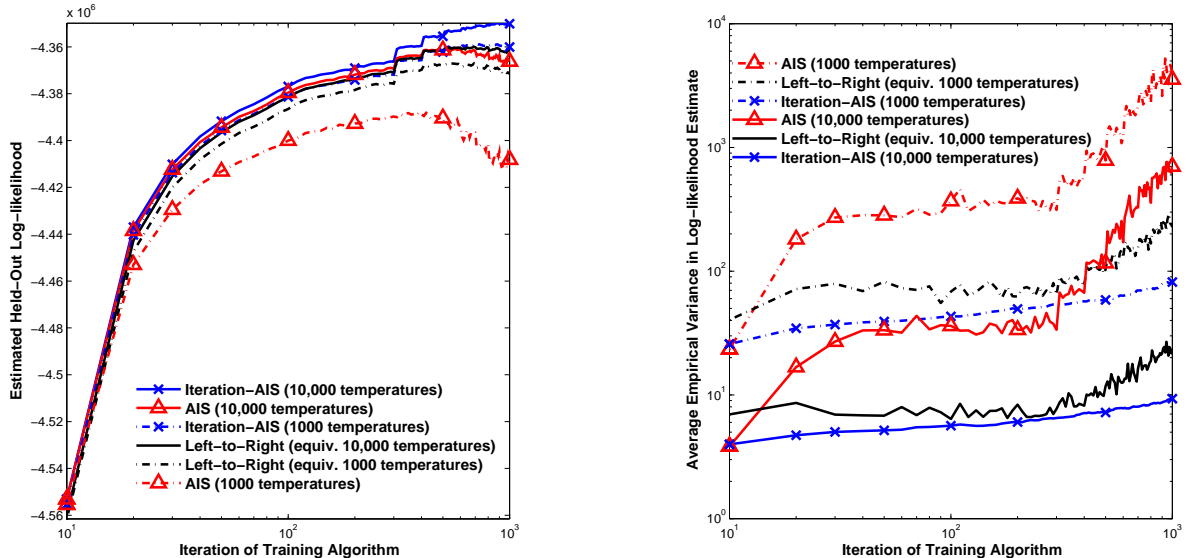


Figure 2: Evaluating iteration-AIS on ACL. Jumps in log-likelihood are due to hyper-parameter optimization. Figure best viewed in color.

(Wallach et al., 2009b). Consistent with this observation, the iteration-AIS likelihood curve at 1000 temperatures coincided with the likelihood curves of the baselines when they were given ten times more computation. The proposed method also exhibited much lower variance in the likelihood estimates (Figure 2, right). This is expected, as the effective number of annealing temperatures is higher, which is known to reduce the variance of the importance weights (Neal, 2001). Similar results were observed on NIPS (see Foulds (2014) for these and other additional results).

The baselines reported decreasing held-out likelihood in later iterations, while iteration-AIS did not. Such a decrease could be due to over-fitting, but is more likely to be caused by convergence failures due to the topics becoming more complex. As evidence for this, the dip in likelihood was smaller with increased computation, and all methods exhibited higher variance in the likelihood estimates for later learning iterations (Figure 2, right, computed based on two evaluations of the likelihood per document, and averaged across documents). The prior probability of the topic models also decreased from around iteration 300 (the same point where standard AIS began to report a decrease in performance), and this is likely to make inference more difficult (see Foulds (2014)).

5 CONCLUSIONS

We have introduced ratio-AIS, a strategy for comparing topic models, and empirically evaluated its performance relative to previous methods using two datasets. Ratio-AIS was found to have low empirical variance, making it useful for document-level analysis. It should be noted that importance sampling can suffer from bias with a finite number of

samples, e.g. approaches such as those described by Wallach et al. (2009b) will typically underestimate the likelihood. For ratio-AIS this results in the potential for a bias that favors a particular model when an insufficient number of samples or temperatures is used, due to the directional nature of the approach. Such a convergence failure of a Monte Carlo algorithm is in general very difficult to detect, but in the proposed method the bias is frequently easily detectable by comparing the results of two Monte Carlo runs. When applied to the evaluation of the per-iteration performance of topic model training algorithms (iteration-AIS), the method outperforms traditional approaches even when given an order of magnitude less computation. Based on our results, we recommend ratio-AIS for document-level analysis, or in cases where the topics are very similar to each other. The method should be performed using both annealing directions as a convergence sanity check, at least for a subset of the held-out documents. Left-to-right is still generally preferred for corpus-level perplexity comparisons, unless per-iteration curves are desired, in which case we recommend that iteration-AIS be used.

For future work, it is straightforward to adapt ratio-AIS to the document completion task. It may also be possible to find other AIS paths with better mixing properties, and the ideas in this work may be applicable to other latent variable models such as RBMs. See Foulds (2014) for a discussion on these ideas, as well as on the use of ratio-AIS with multiple topic models, and where the models differ in the number of topics or in their parametric forms.

Acknowledgements

This work was supported by the Office of Naval Research under MURI grant N00014-08-1-1015.

References

- Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- Buntine, W. Estimating likelihoods for topic models. In *Advances in Machine Learning. First Asian Conference on Machine Learning, ACML 2009, Nanjing, China, November 2-4, 2009. Proceedings*, volume 5828 of *Lecture Notes in Computer Science*, pp. 51–64. Springer, 2009.
- Chang, Jonathan, Boyd-Graber, Jordan, Gerrish, Sean, Wang, Chong, and Blei, David. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pp. 288–296, 2009.
- Foulds, J. R. *Latent Variable Modeling for Networks and Text: Algorithms, Models and Evaluation Techniques*. PhD thesis, University of California, Irvine, 2014.
- Foulds, J. R., Boyles, L., DuBois, C., Smyth, P., and Welling, M. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 446–454, 2013.
- Grimmer, Justin. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35, 2010.
- Grosse, Roger, Maddison, Chris, and Salakhutdinov, Ruslan. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems 26*, pp. 2769–2777, 2013.
- Hoffman, Matt, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- Hoffman, Matthew, Bach, Francis R, and Blei, David M. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pp. 856–864, 2010.
- McCallum, Andrew Kachites. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
- McFarland, Daniel A, Ramage, Daniel, Chuang, Jason, Heer, Jeffrey, Manning, Christopher D, and Jurafsky, Daniel. Differentiating language usage through topic models. *Poetics*, 41(6):607–625, 2013.
- Mimno, D. and McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth International Conference on Uncertainty in Artificial Intelligence*, pp. 411–418, 2008.
- Mimno, David, Wallach, Hanna M, Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272. Association for Computational Linguistics, 2011.
- Neal, R.M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.
- Newman, David, Lau, Jey Han, Grieser, Karl, and Baldwin, Timothy. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010.
- Nguyen, Viet-An, Boyd-Graber, Jordan, Resnik, Philip, Cai, Deborah A, Midberry, Jennifer E, and Wang, Yuanxin. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, pp. 1–41, 2013.
- Radev, Dragomir R., Muthukrishnan, Pradeep, Qazvinian, Vahed, and Abu-Jbara, Amjad. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944, 2013.
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 487–494. AUAI Press, 2004.
- Scott, J.G. and Baldridge, J. A recursive estimate for the predictive likelihood in a topic model. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 527–535, 2013.
- Wallach, Hanna M, Mimno, David M, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems 22*, pp. 1973–1981, 2009a.
- Wallach, H.M., Murray, I., Salakhutdinov, R., and Mimno, D. Evaluation methods for topic models. In *International Conference on Machine Learning*, pp. 1105–1112. ACM, 2009b.