Robust Evaluation of Topic Models

James Foulds **Padhraic Smyth**
Department of Computer Science
University of California, Irvine
{jfoulds, smyth}@ics.uci.edu

Abstract

Statistical topic models such as latent Dirichlet allocation (LDA) have become enormously popular in the past decade, with dozens of extensions being proposed each year in conferences such as NIPS, ICML, KDD, EMNLP, and others. Test set perplexity is frequently the method of choice used in these papers for comparing new models with older variants, yet relatively little attention has been paid (with a few notable exceptions) to the details of how these perplexity values are computed. In this paper we take a close look at the how to accurately compare the predictive performance of different topic models, introducing an extension of annealed importance sampling (as first proposed by [5]) that yields significantly more reliable estimates of model differences relative to current methods.

1 Introduction

When proposing a new topic model it is important to evaluate its performance. In the context of unsupervised learning the standard approach (in machine learning at least) for evaluating a statistical model is to compute the probability of held-out data.[1] In the case of topic models, for a held-out document d with word vector $w^{(d)}$, and given point estimates of topics Φ and the Dirichlet prior $\alpha^{(d)}$ (if learned), this corresponds to to computing the quantity $Pr(w^{(d)}|\Phi, \alpha^{(d)})$. This is intractable to compute and difficult to reliably approximate. A wide variety of approximation strategies are proposed and discussed in papers such as [5], [1] and [4]. Although the methods provide a significant advance over naive approaches, it is clear that the reliable, efficient evaluation of topic models remains an open problem. A simpler alternative is to use the method of document completion, which is easier to estimate, but is not the gold standard prediction task that we would ideally like. For example, document completion can obscure small differences between models based on the prior, such as differences between different Dirichlet multinomial regression models, causing difficulty in evaluating the relative performance of such models.

In this paper we propose a new method for comparing the predictive performance of any topic model relative to a baseline model. The strategy is based on the annealed importance sampling (AIS) method as applied to topic model evaluation in [5]. The key idea is to focus on the *ratio* of the likelihoods of two models rather than computing the likelihoods of each model separately. This strategy results in a much more reliable estimate of the relative performance of the models, with lower variance across samples than previous approaches. As a bonus, by computing the reciprocal ratio using the same technique, we can in some cases detect convergence failures of the sampler.

[1] Although a number of techniques alternative to this prediction task have been proposed, such as measures of topic coherence, posterior predictive checks and human evaluations of topic models (citations omitted for space), prediction of held-out documents remains an essential evaluation technique.

2 Background

As in [5], we focus on the computation of $P(w^{(d)}|\Phi, \alpha^{(d)})$, the likelihood of a held out document d. This quantity (or perplexity, which is a function of it) can be used to evaluate a point estimate of the topics Φ, or as an inner loop to evaluate Bayesian evaluation metrics such as the posterior predictive probability of held out documents or the marginal likelihood. It is in general intractable to compute $P(w^{(d)}|\Phi, \alpha^{(d)})$ directly, as it involves either an intractable sum or an intractable integral. Note that we allow the Dirichlet prior to be document dependent, to account for topic models such as Dirichlet multinomial regression [2], which allows the prior to depend on each document's features. Notationally, in this work, when we write $P(w^{(d)}|\Phi, \alpha^{(d)})$, we are implicitly conditioning on any features or learned parameters used to compute $\alpha^{(d)}$.

2.1 Annealed Importance Sampling

Annealed Importance Sampling (AIS) [3] is a general technique for estimating an expectation of some function of a variable x with respect to an intractable distribution of interest p_0. Consider another distribution p_n (which is typically easy to sample from) and a sequence of "intermediate" distributions $p_{n-1}, \ldots, p_1$ 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 we have a Markov chain with transition operator $T_i(x, x')$ which is invariant to that distribution. We need to sample from these Markov chains, and for each p_i be to able to evaluate some function f_i which is proportional to it. Similarly to traditional importance sampling, AIS produces a collection of samples $x^{(1)}, \ldots, x^{(S)}$ with associated importance weights $w^{(1)}, \ldots, w^{(S)}$. As for 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 $x_{n-2}, \ldots, x_0$ 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 $(x_0, \ldots, x_{n-1})$ 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:

$$Q(x_0, \ldots, x_{n-1}) \propto f_n(x_{n-1}) \prod_{s=n-1}^{1} T_s(x_s, x_{s-1}), \qquad P(x_0, \ldots, x_{n-1}) \propto f_0(x_0) \prod_{s=1}^{n-1} \tilde{T}_s(x_{s-1}, x_s)\,, \tag{1}$$

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

$$w^{(i)} = \frac{P(x_0, \ldots, x_{n-1})}{Q(x_0, \ldots, x_{n-1})} = \prod_{s=0}^{n-1} \frac{f_s(x_s)}{f_{s+1}(x_s)}\,. \tag{2}$$

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. [5] describe the AIS procedure, as applied to LDA. The likelihood of a test document given a topic model can be estimated using this strategy of exploiting AIS to estimate a normalization constant, operating on the latent topic assignments $z^{(d)}$ for the document.[2] We can

[2]The derivation here differs slightly from that of Wallach et al. [5]. The present derivation suggests that the procedure described in [5] should be repeated many times, returning the average of the resulting values.

set $f_0 = Pr(w^{(d)}, z^{(d)}|\phi, \alpha^{(d)})$, $f_n = Pr(z^{(d)}|\alpha^{(d)})$, with intermediate distributions $f_j = f_0^{\beta_j} f_n$ and the transition operators T_j being the Gibbs sampler for f_j. The ratio of normalizing constants is

$$\frac{\sum w^{(i)}}{N} \approx \frac{\sum_{z^{(d)}} Pr(w^{(d)}, z^{(d)}|\phi, \alpha^{(d)})}{\sum_{z^{(d)}} Pr(z^{(d)}|\alpha^{(d)})} = \frac{Pr(w^{(d)}|\phi, \alpha^{(d)})}{1} = Pr(w^{(d)}|\phi, \alpha^{(d)}) \,. \tag{3}$$

3 Robust Comparisons of Topic Models

The above method computes the ratio of the desired quantity $Pr(w^{(d)}|\phi, \alpha^{(d)})$ and a quantity which equals one, so stochastic noise is introduced due to the denominator, even though this is a constant. The prior may also in many cases be very different from the posterior, leading to high variance.

Furthermore, the most common 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. Thus, the real quantity of interest is the *relative* log-likelihood scores of the model and the baseline:

$$\log Pr(w^{(d)}|\phi^{(1)}, \alpha^{(d,1)}) - \log Pr(w^{(d)}|\phi^{(2)}, \alpha^{(d,2)}) = \log \frac{Pr(w^{(d)}|\phi^{(1)}, \alpha^{(d,1)})}{Pr(w^{(d)}|\phi^{(2)}, \alpha^{(d,2)})} \,. \tag{4}$$

To compute this, we must perform the AIS procedure once for each model, incurring the stochastic error twice. Given that the procedure is already designed to compute a ratio, to avoid these issues we propose to instead use AIS to compute Equation 4 directly. Let $f_0(z^{(d)}) = Pr(w^{(d)}, z^{(d)}|\phi^{(1)}, \alpha^{(d,1)})$, $f_n(z^{(d)}) = Pr(w^{(d)}, z^{(d)}|\phi^{(2)}, \alpha^{(d,2)})$, $f_s(z^{(d)}) = f_0(z^{(d)})^{\beta_s} f_n(z^{(d)})^{1-\beta_s}$ and the transition operator be the Gibbs sampler. We have importance weights

$$\begin{aligned} w^{(i)} &= \prod_{s=0}^{n-1} \frac{Pr(w^{(d)}, z_s^{(d)}|\phi^{(1)}, \alpha^{(d,1)})^{\beta_s} Pr(w^{(d)}, z_s^{(d)}|\phi^{(2)}, \alpha^{(d,2)})^{1-\beta_s}}{Pr(w^{(d)}, z_s^{(d)}|\phi^{(1)}, \alpha^{(d,1)})^{\beta_{s+1}} Pr(w^{(d)}, z_s^{(d)}|\phi^{(2)}, \alpha^{(d,2)})^{1-\beta_{s+1}}} \\ &= \prod_{s=0}^{n-1} \frac{Pr(w^{(d)}, z_s^{(d)}|\phi^{(1)}, \alpha^{(d,1)})^{\tau}}{Pr(w^{(d)}, z_s^{(d)}|\phi^{(2)}, \alpha^{(d,2)})^{\tau}} \,, \end{aligned} \tag{5}$$

assuming $\beta_s - \beta_{s+1} = \tau \; \forall s$. 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. Finally, the desired quantity can be estimated via

$$\frac{Pr(w^{(d)}|\phi^{(1)}, \alpha^{(d,1)})}{Pr(w^{(d)}|\phi^{(2)}, \alpha^{(d,2)})} \approx \sum \frac{w^{(i)}}{N} \,. \tag{6}$$

To implement this method we need to draw initially from $f_n(z^{(d)})$, which we accomplish via Gibbs sampling. Note that the sampler is still correct if the initial Gibbs sampler fails to converge, although the variance will be higher. Furthermore, 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 time wasted due to running the annealing on correlated samples.

3.1 Document Completion

Suppose we would instead like to compare the performance of the models on a document completion task, where we observe some portion of a document $w^{(d,a)}$ and the goal is to predict the remainder of the document $w^{(d,b)}$. In this case, we let $f_0(z^{(d)}) = Pr(w^{(d,b)}, z^{(d)}|w^{(d,a)}, \phi^{(1)}, \alpha^{(d,1)})$ and $f_n(z^{(d)}) = Pr(w^{(d,b)}, z^{(d)}|w^{(d,a)}, \phi^{(2)}, \alpha^{(d,2)})$. By a similar argument, we can estimate

$$\frac{Pr(w^{(d,b)}|w^{(d,a)}, \phi^{(1)}, \alpha^{(d,1)})}{Pr(w^{(d,b)}|w^{(d,a)}, \phi^{(2)}, \alpha^{(d,2)})} \approx \sum \frac{w^{(i)}}{N} \text{ , where} \tag{7}$$

$$w^{(i)} = \prod_{s=0}^{n-1} \frac{Pr(w^{(d)}, z_s^{(d)}|\phi^{(1)}, \alpha^{(d,1)})^{\tau}}{Pr(w^{(d)}, z_s^{(d)}|\phi^{(2)}, \alpha^{(d,2)})^{\tau}} \times \frac{Pr(w^{(d,a)}|\phi^{(2)}, \alpha^{(d,2)})}{Pr(w^{(d,a)}|\phi^{(1)}, \alpha^{(d,1)})} \,. \tag{8}$$

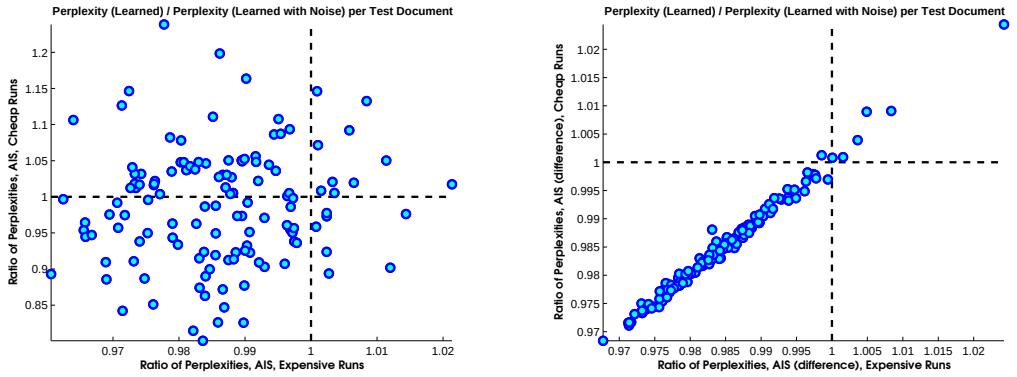

Figure 1: Experimental results

These importance weights consist of two terms: (1) the importance weights for the ratio of likelihoods of the entire document, computed exactly as in the previous section, and (2) the reciprocal ratio, computed on just the observed portion of the document. Note that (1) is independent of which portion of the document is observed, which means that the corresponding samples need only be computed once if we vary the observed portion. The term (2) can be computed using the procedure above, by reversing models 1 and 2 and only executing the sampler on the observed portion $w^{(d,a)}$.

3.2 Detecting Convergence Failures

AIS can fail if the annealing fails to converge to a high-probability state. 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.

4 Experiments

To evaluate the techniques, we performed a held-out prediction experiment on a corpus of 1370 articles from the NIPS conference, holding out 130 documents for testing. We learned topics via Gibbs sampling, and then created perturbed versions of them by taking a convex combination of the learned topics Φ and random topics, with 5% of the weight going to the random topics. Ratios of the perplexities for the two models were computed with both cheap (1 importance sample, 100 temperatures) and expensive (100 importance samples, 1000 temperatures) runs. The proposed method (Figure 1, right) remained accurate when on a budget, predicting that the unperturbed topics were best for 95% of the documents, compared to 52% for the standard method (Figure 1, left).

5 Conclusions

We have introduced a new method for evaluating topic models. The results are aligned with theoretical intuition, showing that the method is much more robust than previous methods at comparing the relative performance of the models. We are currently performing more comprehensive experiments, including exploring the trade-offs with respect to the particle filtering approaches of [4] and [5].

References

[1] W. Buntine. Estimating likelihoods for topic models. In *Advances in Machine Learning*, pages 51–64. Springer, 2009.

[2] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, pages 411–418, 2008.

[3] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[4] J.G. Scott and J. Baldridge. A recursive estimate for the predictive likelihood in a topic model. In *International Conference on Artificial Intelligence and Statistics*, pages 527–535, 2013.

[5] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *International Conference on Machine Learning*, pages 1105–1112. ACM, 2009.