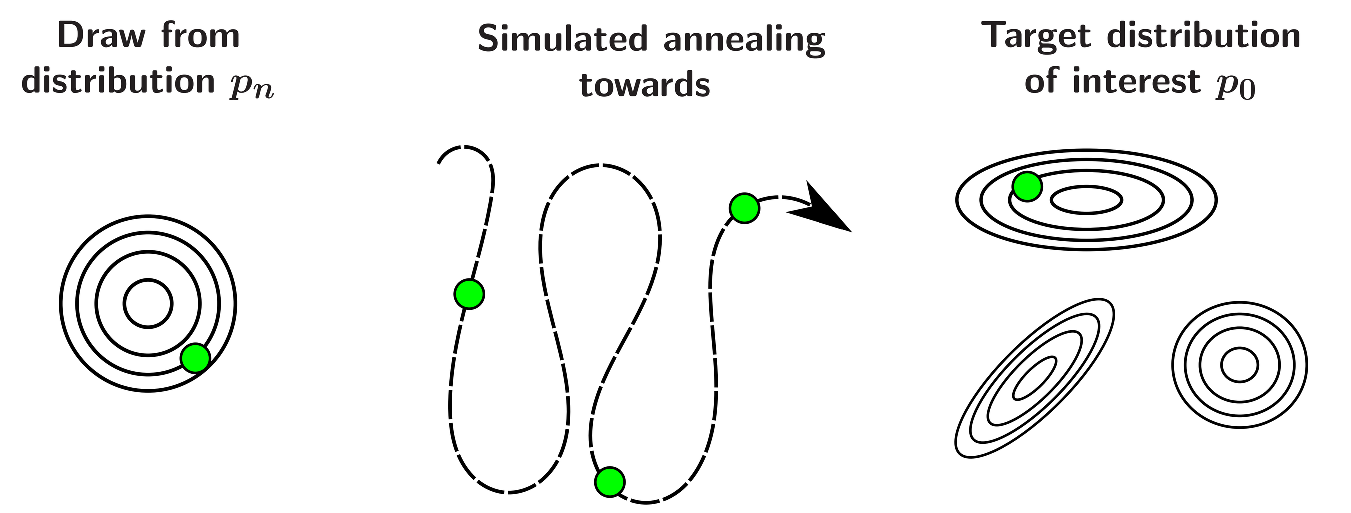
Robust Evaluation of Topic Models

James Foulds and Padhraic Smyth
University of California, Irvine

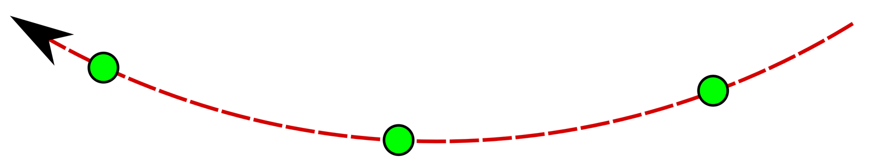
Abstract

- Despite recent advances in learning and inference algorithms, **evaluating** the predictive performance of topic models is still painfully slow and unreliable.
- We propose a new strategy for computing **relative log-likelihood** (or perplexity) scores of topic models, based on **annealed importance sampling**.
- The proposed method has **smaller Monte Carlo error** than previous approaches, leading to marked improvements in both **accuracy** and **computation time**.

Annealed Importance Sampling (Neal, 2001)



Use this as an **importance sampling proposal distribution** for:



Annealing in the reverse direction, from the **target** to the **source**.

The importance samples can be used to estimate the ratio of normalizing constants of $f_0 \propto p_0$ and $f_n \propto p_n$, via

$$\frac{\sum w^{(i)}}{N} \Rightarrow \frac{\int f_0(x)dx}{\int f_n(x)dx}$$

Wallach *et al.* (2009) show how to employ AIS in the context of topic models to estimate $Pr(w^{(d)}|\Phi, \alpha^{(d)})$:

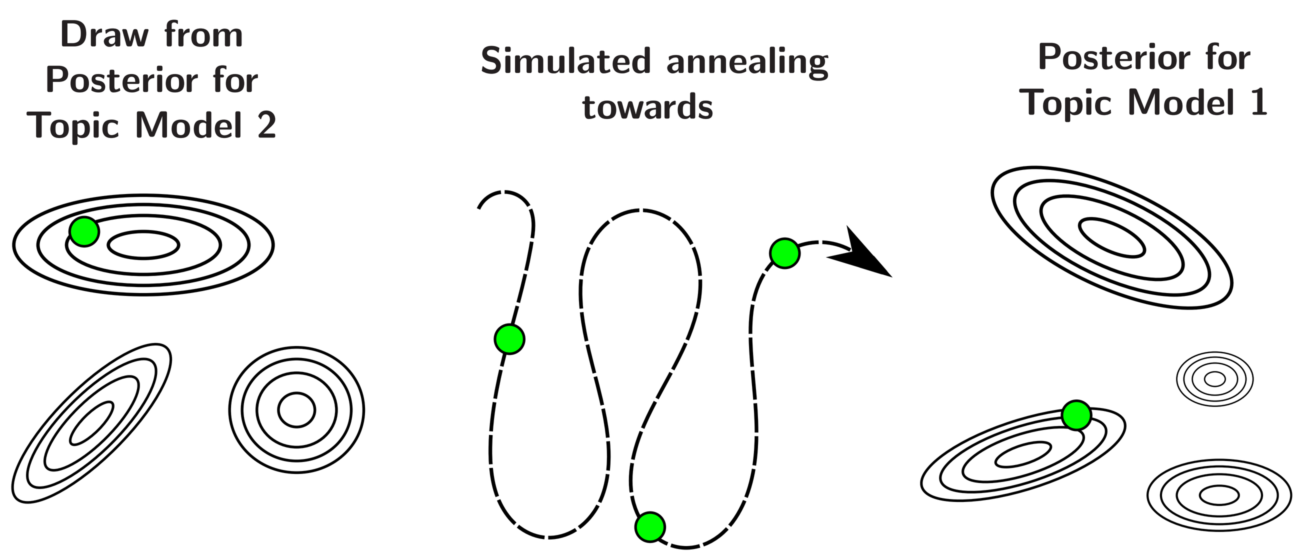
- Perform AIS on the topic assignments z.
- Anneal from the prior to the posterior.
- Estimate the likelihood by averaging the importance samples.

The Proposed Method

- Typically for evaluation we are interested in the **relative** performance of topic model 1 (e.g. a new model) and topic model 2 (e.g. vanilla LDA):

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

- This could be estimated by running AIS **once for each model**.
- However, AIS is already capable of computing ratios. We therefore propose to use AIS to **compute this ratio directly**. The procedure is:



Note that this approach avoids several sources of Monte Carlo error incurred by naively running AIS for each model separately. Specifically, the naive method:

- estimates the denominator of a ratio even though it is a constant $(=1)$,
- uses different z's for both models,
- and is run twice, introducing Monte Carlo noise each time.

Convergence check: Anneal in the reverse direction to compute the reciprocal.

Experimental Analysis: NIPS Corpus

- A corpus of 1740 NIPS articles from 1987 – 1999. We held out a test set of 130 articles.
- **Task**: compute the relative performance of **learned topics**, and **perturbed** versions of these topics (5 % random noise).

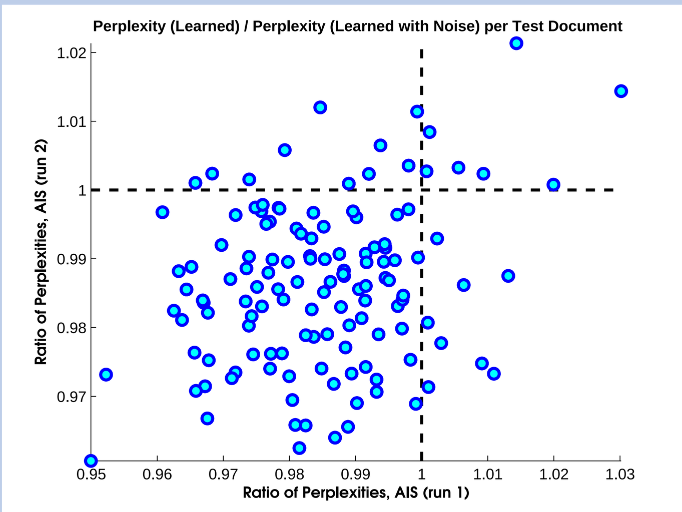
How to read these graphs

- Each dot represents a document
- Each axis shows, for the corresponding method, the estimated ratio,

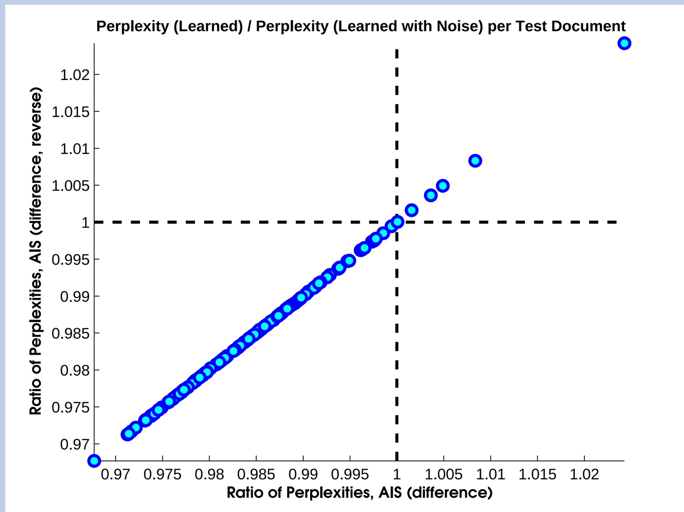
$$\frac{\text{perplexity of the learned topics}}{\text{perplexity of the perturbed learned topics}}$$

- **Dots below 1**: Unperturbed topics are better (likely **correct**)
- **Dots above 1**: Perturbed topics are better (likely **incorrect**)
- **Dots on the diagonal**: The two methods **agree** on the perplexity ratio

Naive AIS

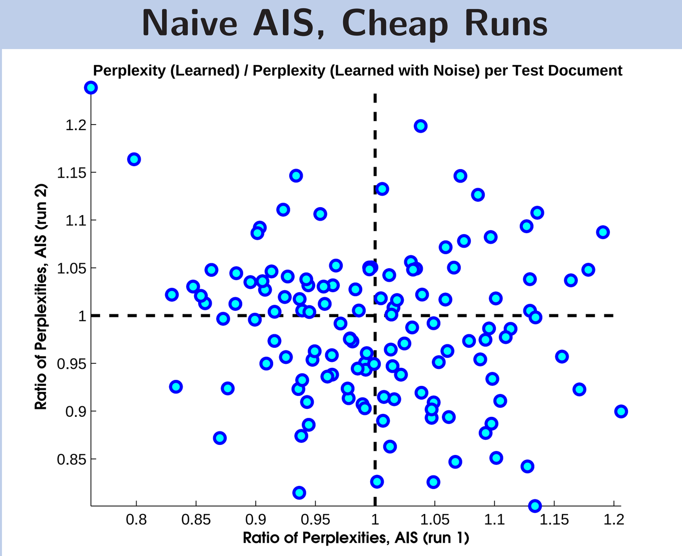


The Proposed Method

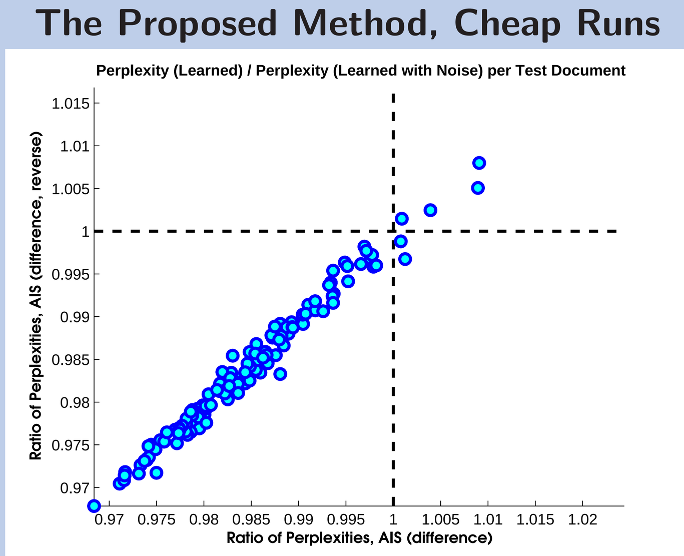


The proposed method was much more **consistent between runs**, in both directions of annealing. It also was much more reliable at determining the **direction of the difference** between models correctly.

Naive AIS, Cheap Runs

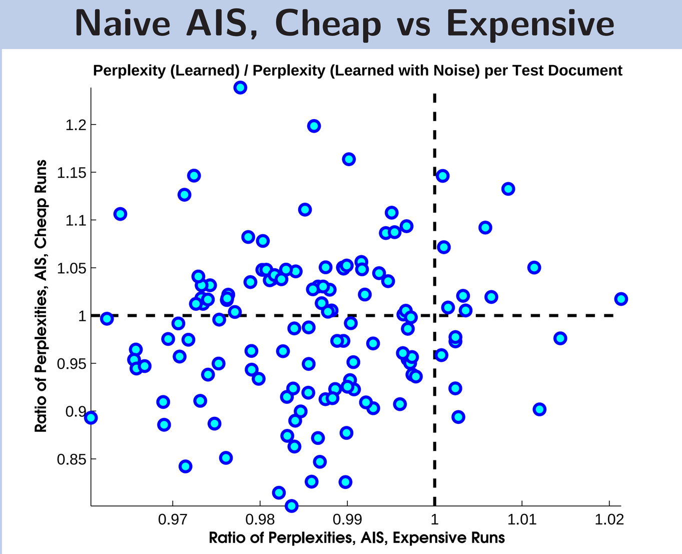


The Proposed Method, Cheap Runs

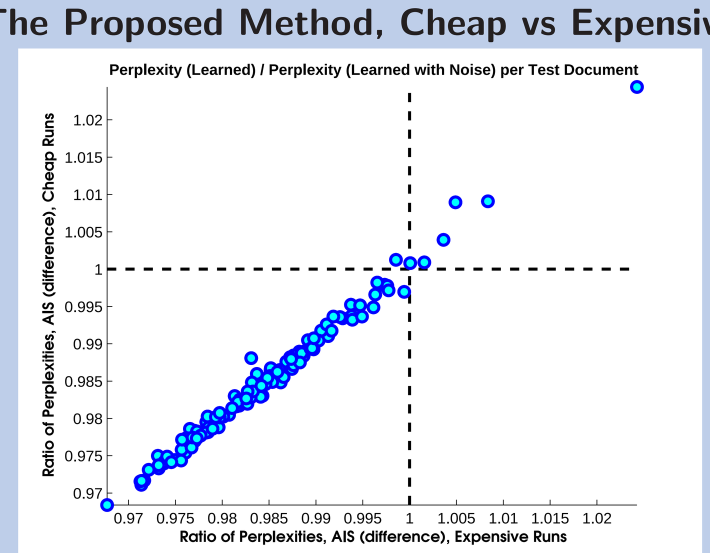


These advantages were most pronounced with a **small computational budget** per document.

Naive AIS, Cheap vs Expensive



The Proposed Method, Cheap vs Expensive

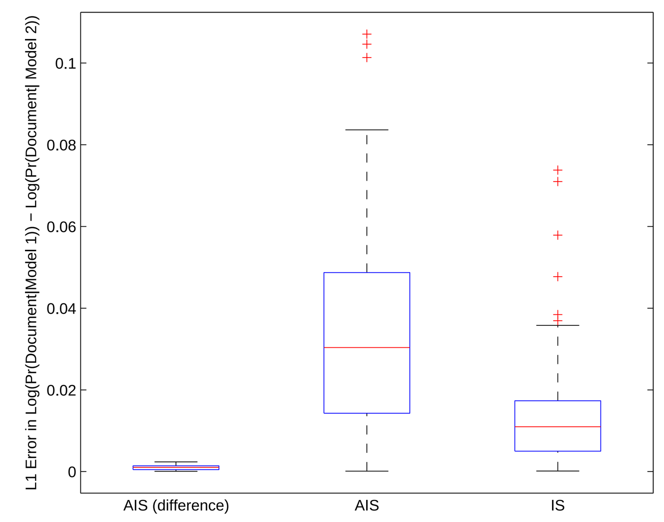


On a computational budget, accurate results were obtained, **similar to those of more expensive runs**.

Overall Results

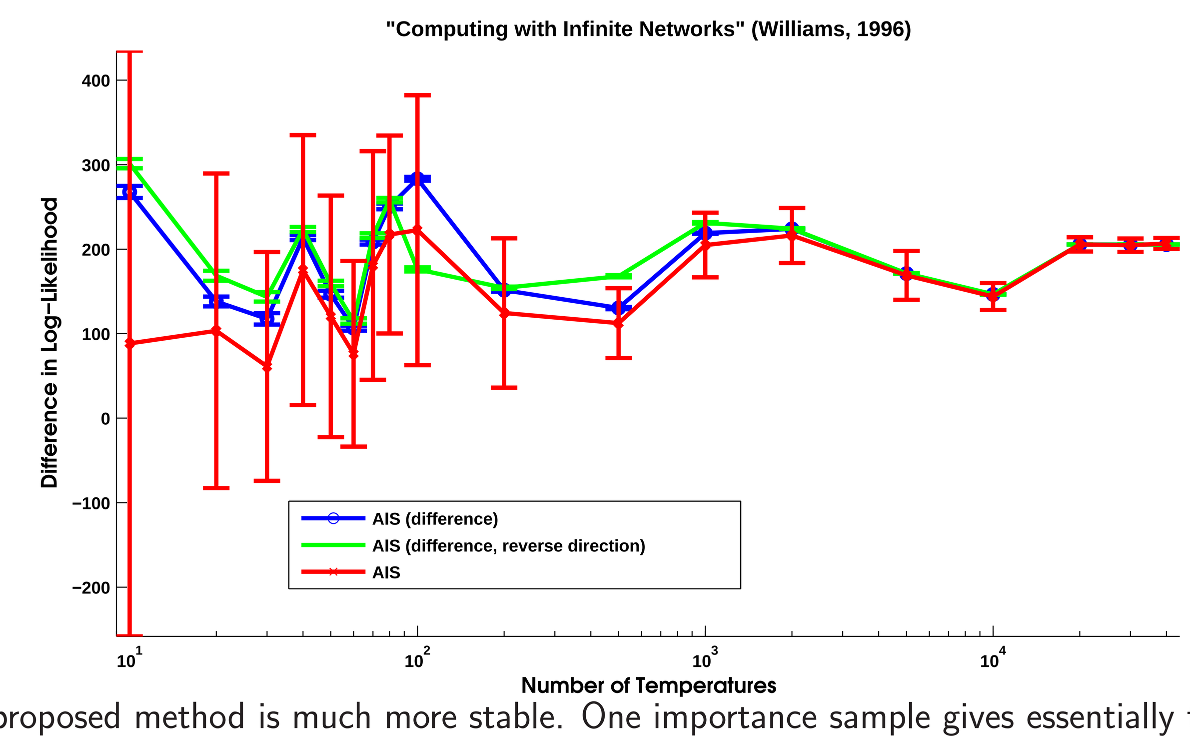
Method	Percent of Documents with Correct Evaluation (I.e., the Unperturbed Topics Win vs the Perturbed Topics)
Expensive runs:	
AIS	88 %
AIS (difference)	95 %
AIS (difference, reverse)	95 %
Cheap runs:	
AIS	52 %
AIS (difference)	95 %
AIS (difference, reverse)	96 %

Comparison to Ground Truth on Very Small Problems



- In this graph, lower values are better.
- Note: in this regime (4 topics, 8 words per document), importance sampling is better than the naive AIS method. This does not hold in general.

Varying the Number of Temperatures



- The proposed method is much more stable. One importance sample gives essentially the same answer as 100 importance samples.
- The **number of temperatures**, which controls the amount of the space explored, is important for all methods.
- **Recommendation**: use the proposed method, with **one importance sample**, and as **many temperatures** as time permits.

Mathematical Details

The standard AIS method for topic models (Wallach *et al.* , 2009)

- AIS on topic assignments $z^{(d)}$, collapsing $\theta^{(d)}$.
- Draw initial state from the prior over z,
  $f_n = Pr(z^{(d)}|\alpha^{(d)})$
- Anneal towards a distribution proportional to the posterior,
  $f_0 = Pr(w^{(d)}, z^{(d)}|\phi, \alpha^{(d)})$
- Estimate the likelihood via:

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

The proposed AIS scheme

- Set the initial and final distributions proportional to the posteriors for the two models
  - $f_0 = Pr(w^{(d)}, z^{(d)}|\phi^{(1)}, \alpha^{(d,1)})$
  - $f_n = Pr(w^{(d)}, z^{(d)}|\phi^{(2)}, \alpha^{(d,2)})$

A similar argument to the above gives us

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

which is what we wanted. We have importance weights

$$\log w^{(i)} = \frac{1}{n} \sum_{s=0}^{n-1} \log \frac{Pr(w^{(d)}, z_s^{(d)}|\phi^{(1)}, \alpha^{(d,1)})}{Pr(w^{(d)}, z_s^{(d)}|\phi^{(2)}, \alpha^{(d,2)})} .$$

References

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

Wallach, H.M., Murray, I., Salakhutdinov, R., & Mimno, D. 2009. Evaluation methods for topic models. *ICML*.

{jfoulds, smyth}@ics.uci.edu