

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. **Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling.** *Empirical Methods in Natural Language Processing*, 2014, 6 pages.

```
@inproceedings{Nguyen:Boyd-Graber:Resnik-2014,  
Author = {Viet-An Nguyen and Jordan Boyd-Graber and Philip Resnik},  
Url = {docs/2014_emnlp_howto_gibbs.pdf},  
Booktitle = {Empirical Methods in Natural Language Processing},  
Location = {Doha, Qatar},  
Year = {2014},  
Title = {Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Mo  
}
```

Downloaded from [http://cs.colorado.edu/~jbg/docs/2014\\_emnlp\\_howto\\_gibbs.pdf](http://cs.colorado.edu/~jbg/docs/2014_emnlp_howto_gibbs.pdf)

# Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling

**Viet-An Nguyen**  
Computer Science  
University of Maryland  
College Park, MD  
vietan@cs.umd.edu

**Jordan Boyd-Graber**  
Computer Science  
University of Colorado  
Boulder, CO  
Jordan.Boyd.Graber@colorado.edu

**Philip Resnik**  
Linguistics and UMIACS  
University of Maryland  
College Park, MD  
resnik@umd.edu

## Abstract

Markov chain Monte Carlo (MCMC) approximates the posterior distribution of latent variable models by generating many samples and averaging over them. In practice, however, it is often more convenient to cut corners, using only a single sample or following a suboptimal averaging strategy. We systematically study different strategies for averaging MCMC samples and show empirically that averaging properly leads to significant improvements in prediction.

## 1 Introduction

Probabilistic topic models are powerful methods to uncover hidden thematic structures in text by projecting each document into a low dimensional space spanned by a set of *topics*, each of which is a distribution over words. Topic models such as latent Dirichlet allocation (Blei et al., 2003, LDA) and its extensions discover these topics from text, which allows for effective exploration, analysis, and summarization of the otherwise unstructured corpora (Blei, 2012; Blei, 2014).

In addition to exploratory data analysis, a typical goal of topic models is prediction. Given a set of unannotated training data, *unsupervised topic models* try to learn good topics that can generalize to unseen text. *Supervised topic models* jointly capture both the text and associated *metadata* such as a continuous response variable (Blei and McAuliffe, 2007; Zhu et al., 2009; Nguyen et al., 2013), single label (Rosen-Zvi et al., 2004; Lacoste-Julien et al., 2008; Wang et al., 2009) or multiple labels (Ramage et al., 2009; Ramage et al., 2011) to predict metadata from text.

Probabilistic topic modeling requires estimating the posterior distribution. Exact computation of the posterior is often intractable, which motivates approximate inference techniques (Asuncion et al., 2009). One popular approach is Markov chain Monte Carlo (MCMC), a class of inference algorithms to approximate the target posterior distribution. To make prediction, MCMC algorithms generate samples on training data to estimate corpus-level latent variables, and use them to generate samples to estimate document-level latent variables for test data. The underlying theory requires averaging on both training and test samples, but in practice it is often convenient to cut corners: either skip averaging entirely by using just the values of the last sample or use a single training sample and average over test samples.

We systematically study non-averaging and averaging strategies when performing predictions using MCMC in topic modeling (Section 2). Using popular unsupervised (LDA in Section 3) and supervised (SLDA in Section 4) topic models via thorough experimentation, we show empirically that cutting corners on averaging leads to consistently poorer prediction.

## 2 Learning and Predicting with MCMC

While reviewing all of MCMC is beyond the scope of this paper, we need to briefly review key concepts.<sup>1</sup> To estimate a target density  $p(x)$  in a high-dimensional space  $\mathcal{X}$ , MCMC generates samples  $\{x_t\}_{t=1}^T$  while exploring  $\mathcal{X}$  using the Markov assumption. Under this assumption, sample  $x_{t+1}$  depends on sample  $x_t$  only, forming a *Markov chain*, which allows the sampler to spend more time in the most important regions of the density. Two concepts control sample collection:

**Burn-in  $B$ :** Depending on the initial value of the Markov chain, MCMC algorithms take time to reach the target distribution. Thus, in practice, samples before a burn-in period  $B$  are often discarded.

**Sample-lag  $L$ :** Averaging over samples to estimate the target distribution requires i.i.d. samples. However, future samples depend on the current samples (i.e., the Markov assumption). To avoid autocorrelation, we discard all but every  $L$  samples.

### 2.1 MCMC in Topic Modeling

As generative probabilistic models, topic models define a joint distribution over latent variables and observable evidence. In our setting, the latent variables consist of corpus-level *global* variables  $\mathbf{g}$  and document-level *local* variables  $\mathbf{l}$ ; while the evidence consists of words  $\mathbf{w}$  and additional metadata  $\mathbf{y}$ —the latter omitted in unsupervised models.

During training, MCMC estimates the posterior  $p(\mathbf{g}, \mathbf{l}^{\text{TR}} | \mathbf{w}^{\text{TR}}, \mathbf{y}^{\text{TR}})$  by generating a *training Markov chain* of  $T_{\text{TR}}$  samples.<sup>2</sup> Each training sample  $i$  provides a set of fully realized global latent variables  $\hat{\mathbf{g}}(i)$ , which can generate test data. During test time, given a

<sup>1</sup>For more details please refer to Neal (1993), Andrieu et al. (2003), Resnik and Hardisty (2010).

<sup>2</sup>We omit hyperparameters for clarity. We split data into training (TR) and testing (TE) folds, and denote the training iteration  $i$  and the testing iteration  $j$  within the corresponding Markov chains.

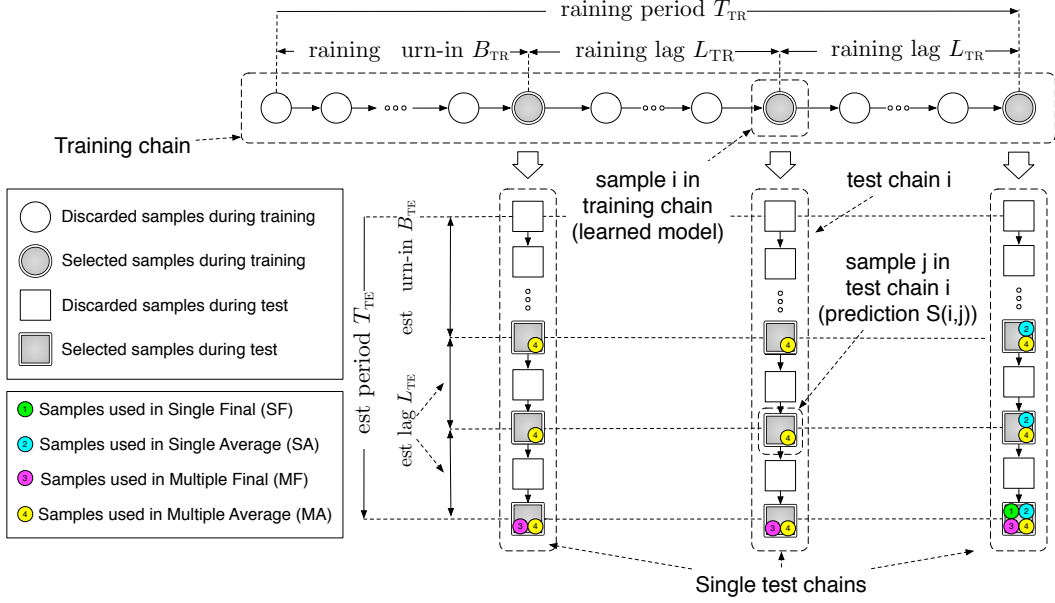


Figure 1: Illustration of training and test chains in MCMC, showing samples used in four prediction strategies studied in this paper: Single Final (SF), Single Average (SA), Multiple Final (MF), and Multiple Average (MA).

learned model from training sample  $i$ , we generate a *test Markov chain* of  $T_{TE}$  samples to estimate the local latent variables  $p(\mathbf{l}^{TE} | \mathbf{w}^{TE}, \hat{\mathbf{g}}(i))$  of test data. Each sample  $j$  of test chain  $i$  provides a fully estimated local latent variables  $\hat{\mathbf{l}}^{TE}(i, j)$  to make a prediction.

Figure 1 shows an overview. To reduce the effects of unconverged and autocorrelated samples, during training we use a burn-in period of  $B_{TR}$  and a sample-lag of  $L_{TR}$  iterations. We use  $\mathcal{T}_{TR} = \{i | i \in (B_{TR}, T_{TR}] \wedge (i - B_{TR}) \bmod L_{TR} = 0\}$  to denote the set of indices of the selected models. Similarly,  $B_{TE}$  and  $L_{TE}$  are the test burn-in and sample-lag. The set of indices of selected samples in test chains is  $\mathcal{T}_{TE} = \{j | j \in (B_{TE}, T_{TE}] \wedge (j - B_{TE}) \bmod L_{TE} = 0\}$ .

## 2.2 Averaging Strategies

We use  $S(i, j)$  to denote the prediction obtained from sample  $j$  of the test chain  $i$ . We now discuss different strategies to obtain the final prediction:

- **Single Final (SF)** uses the last sample of last test chain to obtain the predicted value,

$$S_{SF} = S(T_{TR}, T_{TE}). \quad (1)$$

- **Single Average (SA)** averages over multiple samples in the last test chain

$$S_{SA} = \frac{1}{|\mathcal{T}_{TE}|} \sum_{j \in \mathcal{T}_{TE}} S(T_{TR}, j). \quad (2)$$

This is a common averaging strategy in which we obtain a point estimate of the global latent variables at the end of the training chain. Then, a single test chain is generated on the test data and multiple samples of this test chain are averaged to obtain the final prediction (Chang, 2012; Singh et al., 2012; Jiang et al., 2012; Zhu et al., 2014).

- **Multiple Final (MF)** averages over the last samples of multiple test chains from multiple models

$$S_{MF} = \frac{1}{|\mathcal{T}_{TR}|} \sum_{i \in \mathcal{T}_{TR}} S(i, T_{TE}). \quad (3)$$

- **Multiple Average (MA)** averages over all samples of multiple test chains for distinct models,

$$S_{MA} = \frac{1}{|\mathcal{T}_{TR}|} \frac{1}{|\mathcal{T}_{TE}|} \sum_{i \in \mathcal{T}_{TR}} \sum_{j \in \mathcal{T}_{TE}} S(i, j), \quad (4)$$

## 3 Unsupervised Topic Models

We evaluate the predictive performance of the unsupervised topic model LDA using different averaging strategies in Section 2.

**LDA:** Proposed by Blei et al. in 2003, LDA posits that each document  $d$  is a multinomial distribution  $\theta_d$  over  $K$  topics, each of which is a multinomial distribution  $\phi_k$  over the vocabulary. LDA’s generative process is:

1. For each topic  $k \in [1, K]$ 
  - (a) Draw word distribution  $\phi_k \sim \text{Dir}(\beta)$
2. For each document  $d \in [1, D]$ 
  - (a) Draw topic distribution  $\theta_d \sim \text{Dir}(\alpha)$
  - (b) For each word  $n \in [1, N_d]$ 
    - i. Draw topic  $z_{d,n} \sim \text{Mult}(\theta_d)$
    - ii. Draw word  $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n}})$

In LDA, the global latent variables are topics  $\{\phi_k\}_{k=1}^K$  and the local latent variables for each document  $d$  are topic proportions  $\theta_d$ .

**Train:** During training, we use collapsed Gibbs sampling to assign each token in the training data with a topic (Steyvers and Griffiths, 2006). The probability of

assigning token  $n$  of training document  $d$  to topic  $k$  is  $p(z_{d,n}^{\text{TR}} = k | z_{-d,n}^{\text{TR}}, w_{-d,n}^{\text{TR}}, w_{d,n}^{\text{TR}} = v) \propto$

$$\frac{N_{\text{TR},d,k}^{-d,n} + \alpha}{N_{\text{TR},d,\cdot}^{-d,n} + K\alpha} \cdot \frac{N_{\text{TR},k,v}^{-d,n} + \beta}{N_{\text{TR},k,\cdot}^{-d,n} + V\beta}, \quad (5)$$

where  $N_{\text{TR},d,k}$  is the number of tokens in the training document  $d$  assigned to topic  $k$ , and  $N_{\text{TR},k,v}$  is the number of times word type  $v$  assigned to topic  $k$ . Marginal counts are denoted by  $\cdot$ , and  $^{-d,n}$  denotes the count excluding the assignment of token  $n$  in document  $d$ .

At each training iteration  $i$ , we estimate the distribution over words  $\hat{\phi}_k(i)$  of topic  $k$  as

$$\hat{\phi}_{k,v}(i) = \frac{N_{\text{TR},k,v}(i) + \beta}{N_{\text{TR},k,\cdot}(i) + V\beta} \quad (6)$$

where the counts  $N_{\text{TR},k,v}(i)$  and  $N_{\text{TR},k,\cdot}(i)$  are taken at training iteration  $i$ .

**Test:** Because we lack explicit topic annotations for these data (c.f. Nguyen et al. (2012)), we use *perplexity*—a widely-used metric to measure the predictive power of topic models on held-old documents. To compute perplexity, we follow the *estimating  $\theta$*  method (Walach et al., 2009, Section 5.1) and evenly split each test document  $d$  into  $w_d^{\text{TE}_1}$  and  $w_d^{\text{TE}_2}$ . We first run Gibbs sampling on  $w_d^{\text{TE}_1}$  to estimate the topic proportion  $\hat{\theta}_d^{\text{TE}}$  of test document  $d$ . The probability of assigning topic  $k$  to token  $n$  in  $w_d^{\text{TE}_1}$  is  $p(z_{d,n}^{\text{TE}_1} = k | z_{-d,n}^{\text{TE}_1}, w^{\text{TE}_1}, \hat{\phi}(i)) \propto$

$$\frac{N_{\text{TE}_1,d,k}^{-d,n} + \alpha}{N_{\text{TE}_1,d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE}_1}(i)} \quad (7)$$

where  $N_{\text{TE}_1,d,k}$  is the number of tokens in  $w_d^{\text{TE}_1}$  assigned to topic  $k$ . At each iteration  $j$  in test chain  $i$ , we can estimate the topic proportion vector  $\hat{\theta}_d^{\text{TE}}(i, j)$  for test document  $d$  as

$$\hat{\theta}_{d,k}^{\text{TE}}(i, j) = \frac{N_{\text{TE}_1,d,k}(i, j) + \alpha}{N_{\text{TE}_1,d,\cdot}(i, j) + K\alpha} \quad (8)$$

where both the counts  $N_{\text{TE}_1,d,k}(i, j)$  and  $N_{\text{TE}_1,d,\cdot}(i, j)$  are taken using sample  $j$  of test chain  $i$ .

**Prediction:** Given  $\hat{\theta}_d^{\text{TE}}(i, j)$  and  $\hat{\phi}(i)$  at sample  $j$  of test chain  $i$ , we compute the predicted likelihood for each unseen token  $w_{d,n}^{\text{TE}_2}$  as  $S(i, j) \equiv p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}(i, j), \hat{\phi}(i)) = \sum_{k=1}^K \hat{\theta}_{d,k}^{\text{TE}}(i, j) \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE}_2}(i)}$ .

Using different strategies described in Section 2, we obtain the final predicted likelihood for each unseen token  $p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}, \hat{\phi})$  and compute the perplexity as  $\exp\left(-\left(\sum_d \sum_n \log(p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}, \hat{\phi}))\right) / N^{\text{TE}_2}\right)$  where  $N^{\text{TE}_2}$  is the number of tokens in  $w^{\text{TE}_2}$ .

**Setup:** We use three Internet review datasets in our experiment. For all datasets, we preprocess by tokenizing, removing stopwords, stemming, adding bigrams to

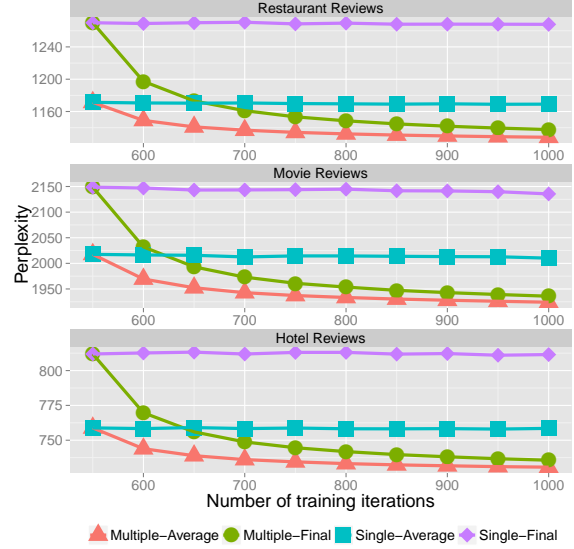


Figure 2: Perplexity of LDA using different averaging strategies with different number of training iterations  $T_{\text{TR}}$ . Perplexity generally decreases with additional training iterations, but the drop is more pronounced with multiple test chains.

the vocabulary, and we filter using TF-IDF to obtain a vocabulary of 10,000 words.<sup>3</sup> The three datasets are:

- HOTEL: 240,060 reviews of hotels from TripAdvisor (Wang et al., 2010).
- RESTAURANT: 25,459 reviews of restaurants from Yelp (Jo and Oh, 2011).
- MOVIE: 5,006 reviews of movies from Rotten Tomatoes (Pang and Lee, 2005)

We report cross-validated average performance over five folds, and use  $K = 50$  topics for all datasets. To update the hyperparameters, we use slice sampling (Walach, 2008, p. 62).<sup>4</sup>

**Results:** Figure 2 shows the perplexity of the four averaging methods, computed with different number of training iterations  $T_{\text{TR}}$ . SA outperforms SF, showing the benefits of averaging over multiple test samples from a single test chain. However, both multiple chain methods (MF and MA) significantly outperform these two methods.

This result is consistent with Asuncion et al. (2009), who run multiple training chains but a single test chain for each training chain and average over them. This is more costly since training chains are usually significantly longer than test chains. In addition, multiple training chains are sensitive to their initialization.

<sup>3</sup>To find bigrams, we begin with bigram candidates that occur at least 10 times in the corpus and use a  $\chi^2$  test to filter out those having a  $\chi^2$  value less than 5. We then treat selected bigrams as single word types and add them to the vocabulary.

<sup>4</sup>MCMC setup:  $T_{\text{TR}} = 1,000$ ,  $B_{\text{TR}} = 500$ ,  $L_{\text{TR}} = 50$ ,  $T_{\text{TE}} = 100$ ,  $B_{\text{TE}} = 50$  and  $L_{\text{TE}} = 5$ .

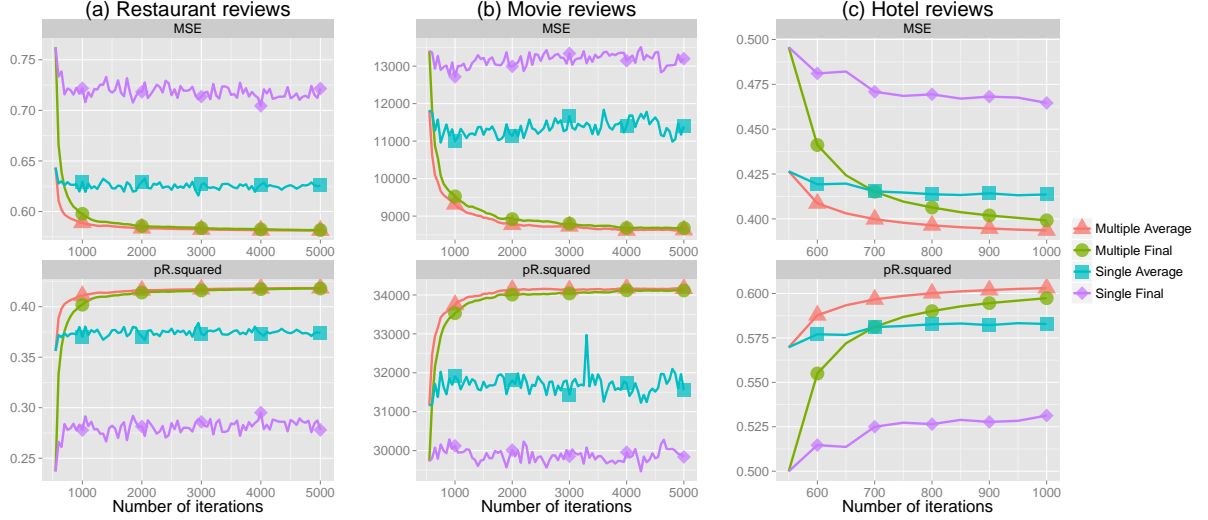


Figure 3: Performance of SLDA using different averaging strategies computed at each training iteration.

## 4 Supervised Topic Models

We evaluate the performance of different prediction methods using supervised latent Dirichlet allocation (SLDA) (Blei and McAuliffe, 2007) for sentiment analysis: predicting review ratings given review text. Each review text is the document  $\mathbf{w}_d$  and the metadata  $\mathbf{y}_d$  is the associated rating.

**SLDA:** Going beyond LDA, SLDA captures the relationship between latent topics and metadata by modeling each document’s continuous response variable using a normal linear model, whose covariates are the document’s empirical distribution of topics:  $\mathbf{y}_d \sim \mathcal{N}(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \rho)$  where  $\boldsymbol{\eta}$  is the regression parameter vector and  $\bar{\mathbf{z}}_d$  is the empirical distribution over topics of document  $d$ . The generative process of SLDA is:

1. For each topic  $k \in [1, K]$ 
  - (a) Draw word distribution  $\phi_k \sim \text{Dir}(\beta)$
  - (b) Draw parameter  $\eta_k \sim \mathcal{N}(\mu, \sigma)$
2. For each document  $d \in [1, D]$ 
  - (a) Draw topic distribution  $\theta_d \sim \text{Dir}(\alpha)$
  - (b) For each word  $n \in [1, N_d]$ 
    - i. Draw topic  $z_{d,n} \sim \text{Mult}(\theta_d)$
    - ii. Draw word  $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n}})$
  - (c) Draw response  $y_d \sim \mathcal{N}(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \rho)$  where  $\bar{z}_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = k]$

where  $\mathbb{I}[x] = 1$  if  $x$  is true, and 0 otherwise.

In SLDA, in addition to the  $K$  multinomials  $\{\phi_k\}_{k=1}^K$ , the global latent variables also contain the regression parameter  $\eta_k$  for each topic  $k$ . The local latent variables of SLDA resembles LDA’s: the topic proportion vector  $\theta_d$  for each document  $d$ .

**Train:** For posterior inference during training, following Boyd-Graber and Resnik (2010), we use stochastic EM, which alternates between (1) a Gibbs sampling

step to assign a topic to each token, and (2) optimizing the regression parameters. The probability of assigning topic  $k$  to token  $n$  in the training document  $d$  is

$$p(z_{d,n}^{\text{TR}} = k | \mathbf{z}_{-d,n}^{\text{TR}}, \mathbf{w}_{-d,n}^{\text{TR}}, w_{d,n}^{\text{TR}} = v) \propto$$

$$\mathcal{N}(y_d; \mu_{d,n}, \rho) \cdot \frac{N_{\text{TR},d,k}^{-d,n} + \alpha}{N_{\text{TR},d,\cdot}^{-d,n} + K\alpha} \cdot \frac{N_{\text{TR},k,v}^{-d,n} + \beta}{N_{\text{TR},k,\cdot}^{-d,n} + V\beta} \quad (9)$$

where  $\mu_{d,n} = (\sum_{k'=1}^K \eta_{k'} N_{\text{TR},d,k'}^{-d,n} + \eta_k) / N_{\text{TR},d}$  is the mean of the Gaussian generating  $y_d$  if  $z_{d,n}^{\text{TR}} = k$ . Here,  $N_{\text{TR},d,k}$  is the number of times topic  $k$  is assigned to tokens in the training document  $d$ ;  $N_{\text{TR},k,v}$  is the number of times word type  $v$  is assigned to topic  $k$ ;  $\cdot$  represents marginal counts and  $^{-d,n}$  indicates counts excluding the assignment of token  $n$  in document  $d$ .

We optimize the regression parameters  $\boldsymbol{\eta}$  using L-BFGS (Liu and Nocedal, 1989) via the likelihood

$$\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{2\rho} \sum_{d=1}^D (y_d^{\text{TR}} - \boldsymbol{\eta}^T \bar{\mathbf{z}}_d^{\text{TR}})^2 - \frac{1}{2\sigma} \sum_{k=1}^K (\eta_k - \mu)^2 \quad (10)$$

At each iteration  $i$  in the training chain, the estimated global latent variables include the a multinomial  $\hat{\phi}_k(i)$  and a regression parameter  $\hat{\eta}_k(i)$  for each topic  $k$ .

**Test:** Like LDA, at test time we sample the topic assignments for all tokens in the test data

$$p(z_{d,n}^{\text{TE}} = k | \mathbf{z}_{-d,n}^{\text{TE}}, \mathbf{w}^{\text{TE}}) \propto \frac{N_{\text{TE},d,k}^{-d,n} + \alpha}{N_{\text{TE},d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE}}} \quad (11)$$

**Prediction:** The predicted value  $S(i, j)$  in this case is the estimated value of the metadata review rating

$$S(i, j) \equiv \hat{y}_d^{\text{TE}}(i, j) = \hat{\boldsymbol{\eta}}(i)^T \bar{\mathbf{z}}_d^{\text{TE}}(i, j), \quad (12)$$

where the empirical topic distribution of test document  $d$  is  $\bar{z}_{d,k}^{\text{TE}}(i, j) \equiv \frac{1}{N_{\text{TE},d}} \sum_{n=1}^{N_{\text{TE},d}} \mathbb{I}[z_{d,n}^{\text{TE}}(i, j) = k]$ .

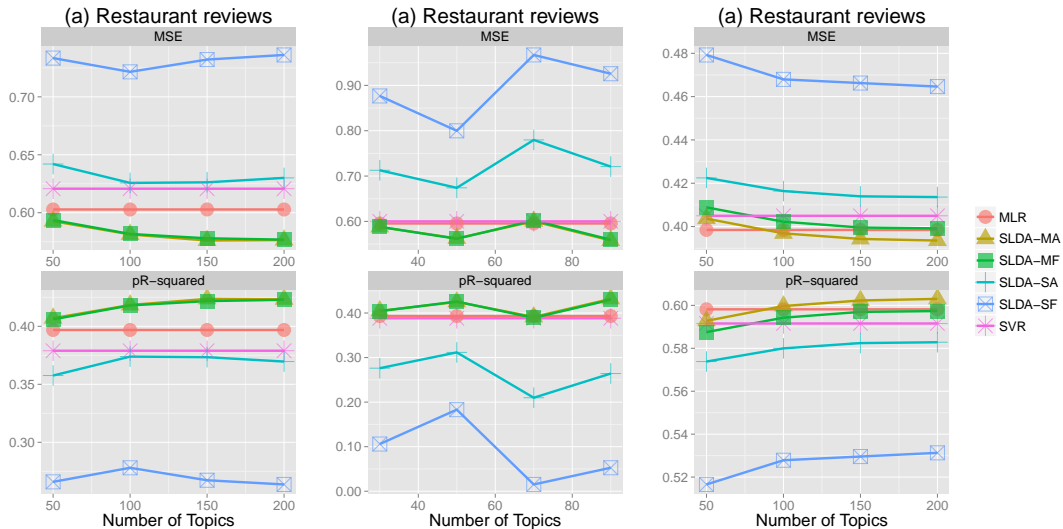


Figure 4: Performance of SLDA using different averaging strategies computed at the final training iteration  $T_{\text{TR}}$ , compared with two baselines MLR and SVR. Methods using multiple test chains (MF and MA) perform as well as or better than the two baselines, whereas methods using a single test chain (SF and SA) perform significantly worse.

**Experimental setup:** We use the same data as in Section 3. For all datasets, the metadata are the review rating, ranging from 1 to 5 stars, which is standardized using  $z$ -normalization. We use two evaluation metrics: mean squared error (MSE) and predictive R-squared (Blei and McAuliffe, 2007).

For comparison, we consider two baselines: (1) multiple linear regression (MLR), which models the metadata as a linear function of the features, and (2) support vector regression (Joachims, 1999, SVR). Both baselines use the normalized frequencies of unigrams and bigrams as features. As in the unsupervised case, we report average performance over five cross-validated folds. For all models, we use a development set to tune their parameter(s) and use the set of parameters that gives best results on the development data at test.<sup>5</sup>

**Results:** Figure 3 shows SLDA prediction results with different averaging strategies, computed at different training iterations.<sup>6</sup> Consistent with the unsupervised results in Section 3, SA outperforms SF, but both are outperformed significantly by the two methods using multiple test chains (MF and MA).

We also compare the performance of the four prediction methods obtained at the final iteration  $T_{\text{TR}}$  of the training chain with the two baselines. The results in Figure 4 show that the two baselines (MLR and SVR) outperform significantly the SLDA using only a single test

chains (SF and SA). Methods using multiple test chains (MF and MA), on the other hand, match the baseline<sup>7</sup> (HOTEL) or do better (RESTAURANT and MOVIE).

## 5 Discussion and Conclusion

MCMC relies on averaging multiple samples to approximate target densities. When used for prediction, MCMC needs to generate and average over both training samples to learn from training data and test samples to make prediction. We have shown that simple averaging—not more aggressive, *ad hoc* approximations like taking the final sample (either training or test)—is not just a question of theoretical aesthetics, but an important factor in obtaining good prediction performance.

Compared with SVR and MLR baselines, SLDA using multiple test chains (MF and MA) performs as well as or better, while SLDA using a single test chain (SF and SA) falters. This simple experimental setup choice can determine whether a model improves over reasonable baselines. In addition, better prediction with shorter training is possible with multiple test chains. Thus, we conclude that averaging using multiple chains produces above-average results.

## Acknowledgments

We thank Jonathan Chang, Ke Zhai and Mohit Iyyer for helpful discussions, and thank the anonymous reviewers for insightful comments. This research was supported in part by NSF under grant #1211153 (Resnik) and #1018625 (Boyd-Graber and Resnik). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

<sup>5</sup>For MLR we use a Gaussian prior  $\mathcal{N}(0, 1/\lambda)$  with  $\lambda = a \cdot 10^b$  where  $a \in [1, 9]$  and  $b \in [1, 4]$ ; for SVR, we use  $\text{SVM}^{\text{light}}$  (Joachims, 1999) and vary  $C \in [1, 50]$ , which trades off between training error and margin; for SLDA, we fix  $\sigma = 10$  and vary  $\rho \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$ , which trades off between the likelihood of words and response variable.

<sup>6</sup>MCMC setup:  $T_{\text{TR}} = 5,000$  for RESTAURANT and MOVIE and 1,000 for HOTEL; for all datasets  $B_{\text{TR}} = 500$ ,  $L_{\text{TR}} = 50$ ,  $T_{\text{TE}} = 100$ ,  $B_{\text{TE}} = 20$  and  $L_{\text{TE}} = 5$ .

<sup>7</sup>This gap is because SLDA has not converged after 1,000 training iterations (Figure 3).

## References

- Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43.
- Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *UAI*.
- David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS*.
- David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3.
- David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April.
- David M. Blei. 2014. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1):203–232.
- Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *EMNLP*.
- Jonathan Chang. 2012. lda: Collapsed Gibbs sampling methods for topic models. <http://cran.r-project.org/web/packages/lda/index.html>. [Online; accessed 02-June-2014].
- Qixia Jiang, Jun Zhu, Maosong Sun, and Eric P. Xing. 2012. Monte Carlo methods for maximum margin supervised topic models. In *NIPS*.
- Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *WSDM*.
- Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, chapter 11. Cambridge, MA.
- Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*.
- D. Liu and J. Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Math. Prog.*
- Radford M. Neal. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.
- Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2012. SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In *ACL*.
- Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Neural Information Processing Systems*.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.
- Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.
- Daniel Ramage, Christopher D. Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *KDD*, pages 457–465.
- Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical Report UMIACS-TR-2010-04, University of Maryland. <http://drum.lib.umd.edu/handle/1903/10058>.
- Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI*.
- Sameer Singh, Michael Wick, and Andrew McCallum. 2012. Monte Carlo MCMC: Efficient inference by approximate sampling. In *EMNLP*, pages 1104–1113.
- Mark Steyvers and Tom Griffiths. 2006. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Leon Bottou and Michael Littman, editors, *ICML*.
- Hanna M Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.
- Chong Wang, David Blei, and Li Fei-Fei. 2009. Simultaneous image classification and annotation. In *CVPR*.
- Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *SIGKDD*, pages 783–792.
- Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *ICML*.
- Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2014. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, 15:1073–1110.