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

Viet-An Nguyen[1], Jordan Boyd-Graber[2], and Philip Resnik[1,3,4]

[1]Computer Science, [3]Linguistics, [4]UMIACS, University of Maryland, College Park, MD [2]Computer Science, University of Colorado, Boulder, CO

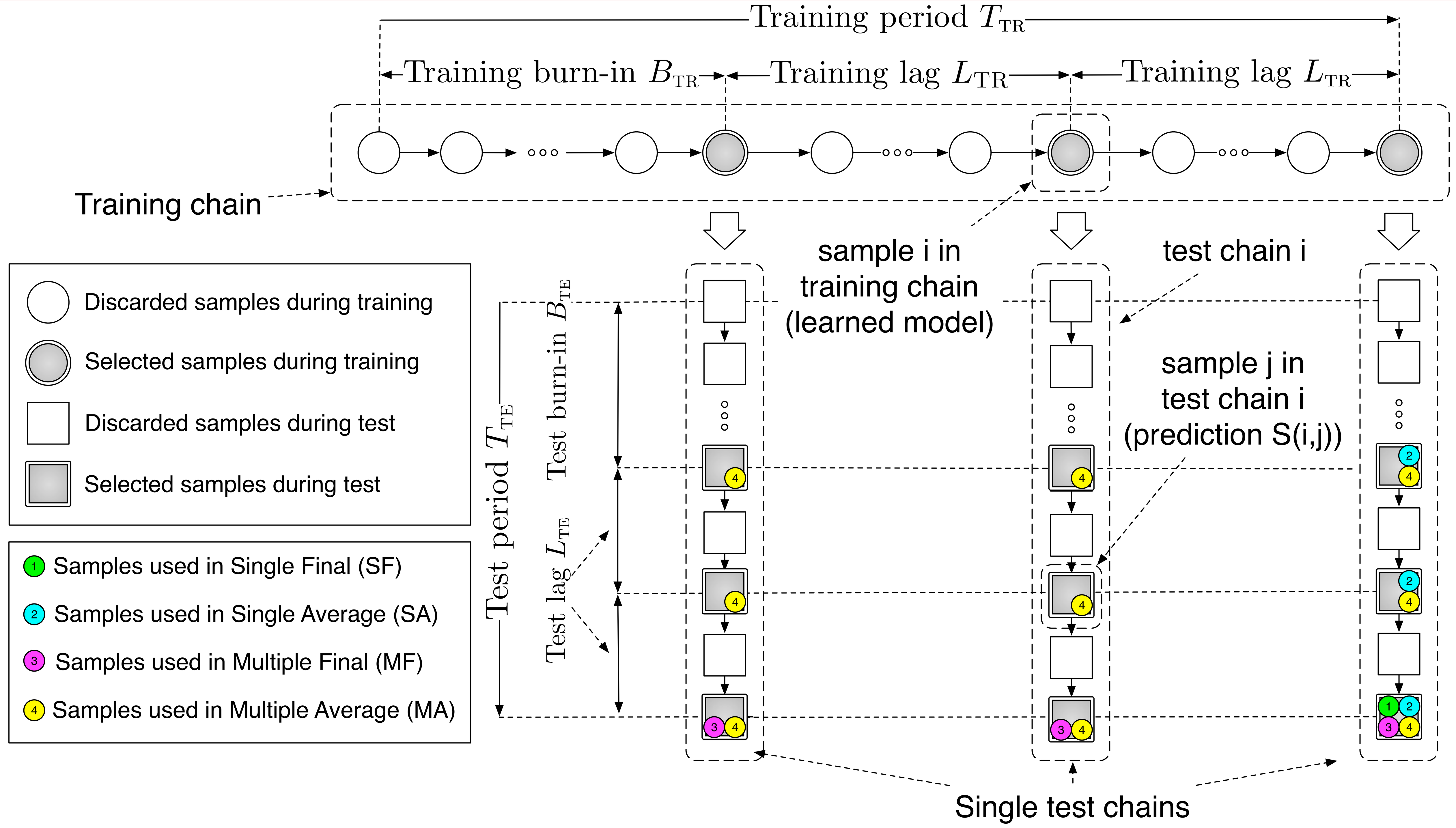
Take-away Messages

Averaging predicted values obtained from multiple test chains consistently improves performances in predicting held-out words (unsupervised topic models) and real-valued metadata of test documents (supervised topic models) across multiple datasets.

Introduction

- 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.
- Two parameters define sample collection control sample collection:
  - **Burn-in** (B): Samples are kept only after a burn-in period B to remove samples that are not converged.
  - **Sampler-lag** (L): All but every L samples are discarded to avoid auto-correlation.

Which Samples Should We Average?



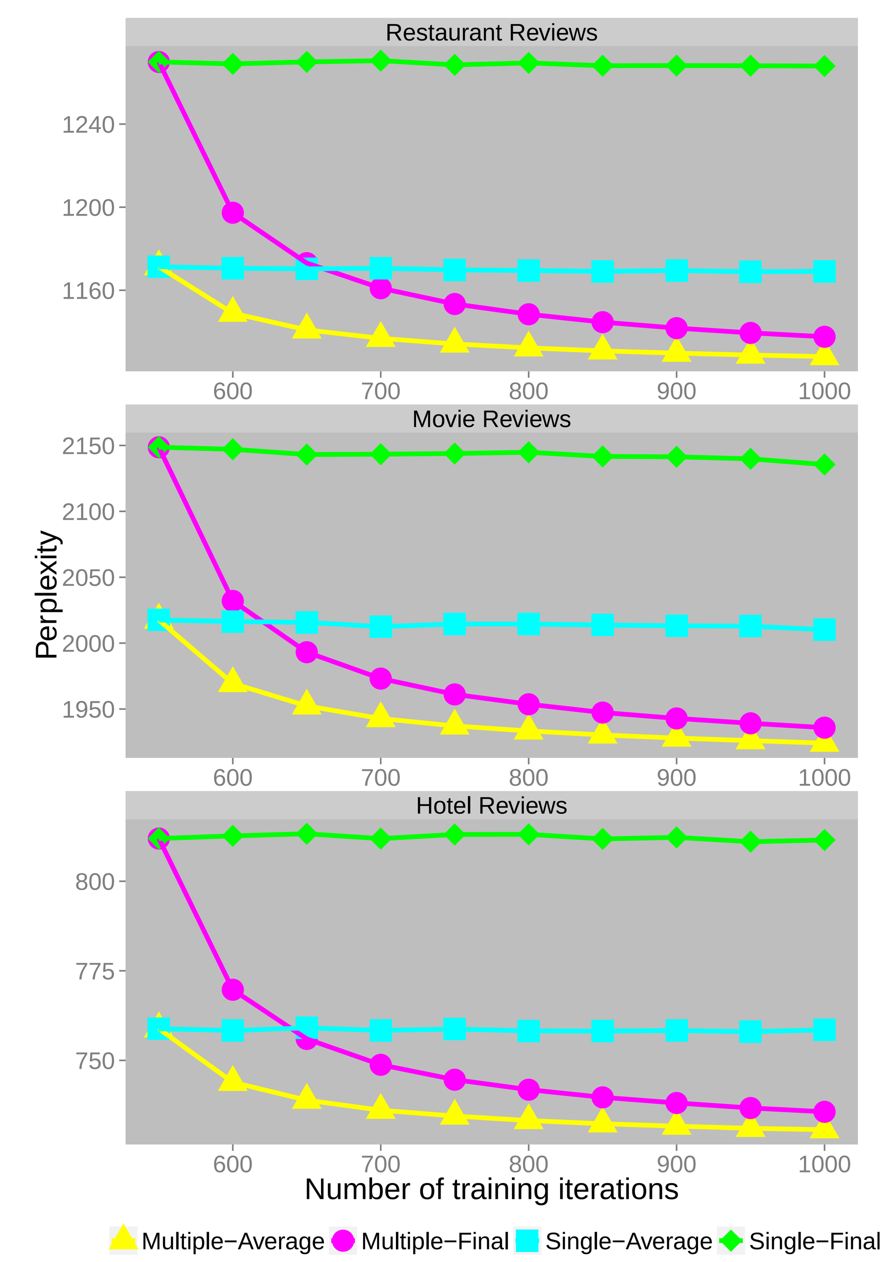
Final prediction: averaging over individual predicted values obtained using different samples $\mathcal{S}$

$$\hat{f} = E_p[f] \approx \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} f(s)$$

Different ways to collect samples

1. **Single Final (SF)** uses the last sample of the last test chain
2. **Single Average (SA)** uses multiple samples of the last test chain
3. **Multiple Final (MF)** uses the last samples of multiple test chains
4. **Multiple Average (MA)** uses multiple test chains, each has multiple samples

Consistent Prediction Results with LDA Across Datasets



Prediction task

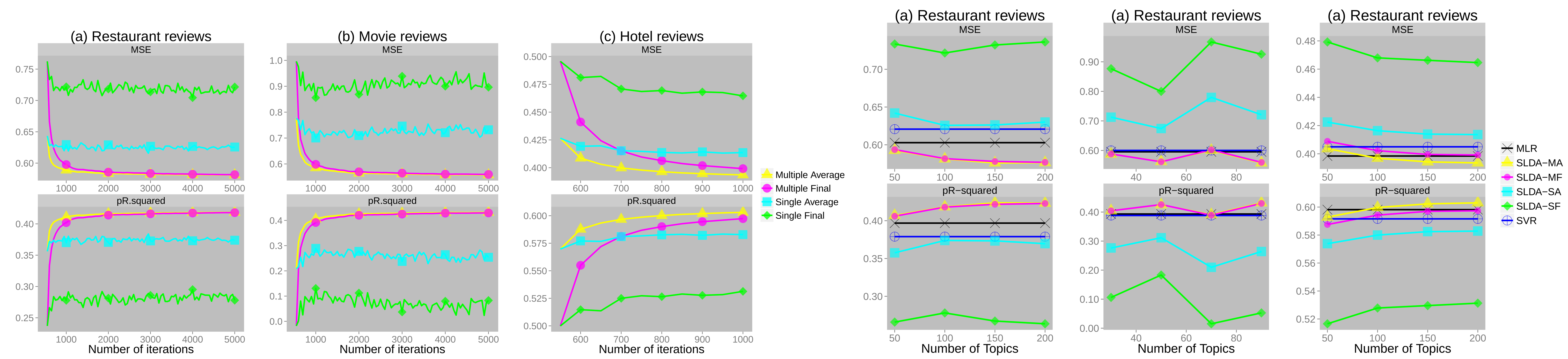
- **Task**: Predicting words in held-out documents
- **Evaluation**: Perplexity—computed using the *estimating θ method* (Wallach et al., 2009)

LDA

- Each document d is a multinomial over topics θ_d
- Each topic k is a multinomial over words ϕ_k
- **Train**: Estimate topics $\{\hat{\phi}_k(i)\}$ at each training iteration i.
- **Test**: Estimate the topic proportion $\hat{\theta}_{d,k}(i,j)$ for each test document d
- **Prediction**: Likelihood of each test token $w_{d,n}$

$$f(i,j) = \sum_{k=1}^{K} \hat{\theta}_{d,k}(i,j) \cdot \hat{\phi}_{k,w_{d,n}}(i)$$

Consistent Prediction Results with Supervised LDA Across Datasets



Prediction task

- **Task**: Predicting real-valued metadata of unseen document given the text
- **Evaluation**: Mean squared error (MSE) and predictive R-squared
- **Train**: Estimate topics $\{\hat{\phi}_k(i)\}$ and regression parameters $\{\hat{\eta}_k(i)\}$ at each training iteration i
- **Test**: For each test document, sample the topic assignments for all its tokens
- **Prediction**: Response variable for each test document

$$f(i,j) = \hat{\eta}(i)^T \bar{z}_d^{\text{TE}}(i,j)$$

SLDA

- Going beyond LDA, SLDA jointly captures the relationship between latent topics and document's real-valued metadata
- Given a set of documents, each is associated with a continuous response variable y_d, SLDA models

$$y_d \sim \mathcal{N}(\eta^T \bar{z}_d, \rho)$$

where $\bar{z}_{d,k} = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = k]$

vietan@cs.umd.edu, Jordan.Boyd.Graber@colorado.edu, resnik@umd.edu