Sparse Stochastic Inference for Latent Dirichlet Allocation

David Mimno\textsuperscript{1}, Matthew D. Hoffman\textsuperscript{2}, David M. Blei\textsuperscript{1}

\textsuperscript{1}Dept. of Computer Science, Princeton U.
\textsuperscript{2}Dept. of Statistics, Columbia U.

Presentation led by Miao Liu

March 7, 2013
1 Introduction

2 Hybrid Stochastic-MCMC Inference

3 Related Work

4 Empirical Results
• A hybrid algorithm for Bayesian topic models
  • sparse Gibbs sampling to estimate variational expectation of local variables (efficiency)
  • online VB to update the variational distribution of global variables (scalability)
• Applications: find large numbers of topics in massive collections of documents
  • 1.2 million books
  • 33 billion words
Latent Dirichlet Allocation (Blei et al., 2003)

- **Notations**
  - $\alpha$: the parameter of Dirichlet prior on the per-document topic distribution
  - $\beta$: the parameter of Dirichlet prior on the per-topic word distribution
  - $\theta_i$: the topic distribution for document $i$
  - $\phi_i$: the word distribution for topic $k$
  - $z_{ij}$: the topic for the $j$th word in document $i$
  - $w_{ij}$: the specific word

- **The generative process**

\[
\theta_i \sim \text{Dir}(\alpha), \quad \text{where} \quad i \in \{1, \cdots, M\} \quad (1)
\]

\[
\phi_k \sim \text{Dir}(\beta), \quad \text{where} \quad k \in \{1, \cdots, K\} \quad (2)
\]

For each of the words $w_{ij}$, where $j \in \{1, \cdots, N_i\}$

\[
z_{ij} \sim \text{Multinomial}(\theta_i), \quad w_{ij} \sim \text{Multinomial}(\phi_{z_{ij}}) \quad (4)
\]
Online LDA (Hoffman et al. (2010))

Algorithm 2 Online variational Bayes for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\lambda$ randomly.
for $t = 0$ to $\infty$ do
    $E$ step:
    Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
    repeat
        Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log \theta_{tk}] + \mathbb{E}_q[\log \beta_{kw}]\}$
        Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
    until $\frac{1}{K} \sum_k |\text{change in } \gamma_{tk}| < 0.00001$

    $M$ step:
    Compute $\tilde{\lambda}_{kw} = \eta + D n_{tw} \phi_{twk}$
    Set $\lambda = (1 - \rho_t) \lambda + \rho_t \tilde{\lambda}$.
end for

- pros
  - less memory
  - faster convergence
- does not scale to large number of topics
Hybrid Stochastic-MCMC Inference

- Sampling: a second source of stochasticity for the gradient
- Marginalize out the topic proportions $\theta_d$
- Corpus-level global variables: $K$ topic-word distributions $\beta_1, \cdots, \beta_K$ over the $V$-dimensional vocabulary.
- Document-level local variables: For a document $d$ of length $N_d$
  - $\theta_d$
  - $z_d = (z_{d1}, \cdots, z_{dN_d})$
- variational distribution
  $q(z_1, \cdots, z_D, \beta_1 \cdots, \beta_K) = \prod_d q(z_d) \prod_k q(\beta_k)$.
  - $q(z_d)$ is a single distribution over the $K^{N_d}$ possible topic configurations
- The VB lower bound

$$\log p(w | \alpha, \eta) \geq \sum_d \mathbb{E}_q \log \left[ p(z_d | \alpha) \prod_i \beta_{z_{di}w_{di}} \right] + \sum_k \mathbb{E}_q \log p(\beta_k | \eta) + \mathcal{H}(q)$$

(5)
• The optimal variational distribution over topic configurations for a document, holding all other variational distribution fixed

\[
q^*(z_d) \propto \exp \left\{ \mathbb{E}_{q(-z_d)} \left[ \log p(z_d|\alpha)p(w_d|z_d, \beta) \right] \right\}
\]

\[
= \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_d)} \prod_k \frac{\Gamma(\alpha + \sum_i I_{z_{di}=k})}{\Gamma(\alpha)}
\]

(6)

(7)

• The hyper parameter of the optimal variational distribution over topic-word distribution, holding the other distributions fixed

\[
\lambda_{kw} = \eta + \sum_d \sum_i \mathbb{E}_q[I_{z_{di}=k}I_{w_{di}=w}]
\]

(8)
Online Stochastic Inference for $\lambda_{kw}$

- Recast the variational LB as a summation over per-document terms $l_d$

$$l_d = \sum_w \left( \mathbb{E}_q[N_{dkw}] + \frac{1}{D} (\eta - \lambda) \right) \mathbb{E}_q[\log \beta_{kw}] + \frac{1}{D} \left( \log \Gamma(\sum_w \lambda_{kw}) - \sum_w \log \Gamma(\lambda_{kw}) \right), \quad (9)$$

- $\sum_d \frac{\partial}{\partial \lambda_k} l_d = \mathbb{E} \left[ \frac{D}{|B|} \sum_{d \in B} \frac{\partial}{\partial \lambda_k} l_d \right]$

- $\mathbb{E}_q[N_{dkw}] = \sum_i \mathbb{E}_q[l_{zi=k} l_{w_{di}=w}]$

- The natural gradient (Hoffman et al., 2010)

$$\mathbb{E}_q[N_{dkw}] + \frac{1}{D} (\eta - \lambda_{kw}) \quad (10)$$

- faster convergence
- cheaper computation

- Issue: the natural gradient cannot be directly evaluated (with a combinatorial number of topic configurations $z_d$).
MCMC within Stochastic Inference

- Solution: use MCMC to sample $z_d$ from $q^*(z_d)$ and the empirical average to estimate $\mathbb{E}_q[N_{dkw}]$
- $q^*(z_{di} = k|z\setminus i) \propto (\alpha + \sum_{j \neq i} I_{z_j=k}) \exp\{\mathbb{E}_q[\log \beta_{kw_{di}}]\}$
- $\mathbb{E}_q[N_{dkw}] \approx \hat{N}_{kw} = \frac{1}{S} \sum_s \sum_{d \in B} \sum_i [I_{z^s_{di}=k} I_{w_{di}=w}]$
- Impacts from MCMC
  1. add stochasticity to the gradient
  2. allows using collapsed objective function that does not represent $\theta_d$
  3. provides a sparse estimate of the gradient (for many words and topics, the estimate of $\mathbb{E}_q[N_{dkw}]$ will be zero)
Previous variational methods lead to dense update to KV topic parameters, which is expensive when $K$ and $V$ are large.

Algorithm 1 exploit the sparsity exhibited by samples from $q^*$. 

Algorithm 1: Algorithm for hybrid stochastic variational-Gibbs inference.

for $t \in 1, \ldots, \infty$ do
\[
\rho_t \leftarrow \left( \frac{1}{t_0 + t} \right)^{\kappa}
\]
sample minibatch $\mathcal{B}$

for $d \in \mathcal{B}$ do
\[
\text{initialize } z_d^0
\]
discard $B$ burn-in sweeps

for sample $s \in 1, \ldots, S$ do
\[
\text{for token } i \in 1, \ldots, N_d \text{ do}
\]
sample $z_{di}^s \propto (\alpha + N_{di}) e^{\mathbb{E}_q[\log \beta_{kw}]}$
\end{for}
\endfor

end for
\endfor
\[
\lambda_{kw}^t \leftarrow (1 - \rho_t)\lambda_{kw}^{t-1} + \rho_t \left( \eta + \frac{D}{|\mathcal{B}|} \hat{N}_{kw} \right)
\]
end for
Two sources of zero-mean noise in constructing an approximate gradient for VB
- Subsampling the data (SS)
- Monte Carlo inference (MC)

SAEM (Delyon et al., 1999): EM + MC

Kuhn & Lavielle (2004): extends SAEM to MCMC estimates

Online EM (Cappé & Moulines, 2009): EM + SS

Collapsed VB (Teh et al. 2006) also analytically marginalizes over $\theta_d$, but still maintains a fully factorized distribution over $z_d$

The prosed method does not restrict to such factored distributions, hence reduces bias

Parallelization
Evaluation

- Held-out probability: calculates the marginal likelihood for held out documents using the "left-to-right sequential sampling" (Wallach et al. 2009; Buntine, 2009)
- Topic coherence
  - measures the semantic quality of a topic by approximating the experience of a user viewing the $W$ most probable words for the topic (Mimno et al., 2011)
  - $D(w)$: the number of document containing one or more token of type $w$
  - $D(w_i, w_j)$: the number of documents containing at least one token of $w_i$ and $w_j$
  - $C(W) = \sum_{i=2}^{W} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}$
  - $C(W)$ is related to point-wise mutual information (Newman et al., 2010)
- Wallclock time
Dataset

- Science/Nature/PNAS articles
  - 350,000 research articles
  - Vocabulary size: 19,000
  - 90% articles for training, 10% for testing

- Pre-1922 books
  - 1.2 million books
  - 33 billion words
Comparison to Online VB (Hoffman et al. 2010)

- Each iteration consists of a mini-batch of 100 documents
- the number of coordinate ascent steps in VB is equal to the number of Gibbs sweeps
- both methods use the same learning schedule
- Standard online VB takes time linear in $K$

Figure 1. Comparison of seconds per mini-batch between online variational Bayes (Hoffman et al., 2010) and sampled online inference (this paper). Online VB is linear in $K$, while sampled inference takes advantage of sparsity.
Comparison to Online VB (Hoffman et al. 2010)

The entropy of the topic distributions:
- the proposed method: $6.8 \pm 0.46$
- online VB: $6.0 \pm 0.58$

This result could indicate that coordinate ascent over the local variables for online LDA is not converging?
Comparison to Sequential Monte Carlo (Ahmed et al 2012)

- The SMC sampler sweeps through each document the same number of times as the sampled online algorithm.
- The learning rate schedule allows sampled online inference to "forget" its initial topics.
- SMC weights all documents equally.

*Figure 3. Sampled online inference performs better than one pass of sequential Monte Carlo, after processing a comparable number of documents with $K = 200$.***
Effect of parameter settings

- Number of samples: $B + S$
- Topic-word smoothing: $\eta$
- Forgetting factors
  - learning rate
  - Size of corpus $D$

*Figure 4.* Topic quality is lowest for large values of $t_0$, but only in small corpora. Panels represent the proportion of training data used. Each panel shows coherence values for five $K = 100$ topic models with varying learning rates.
Table 1. Randomly selected topics from a 2000-topic model trained on a library of 1.2 million out-of-copyright books.

- killed, wounded, sword, slain, arms, military, rifle, wounds, loss, human, Plato, Socrates, universe, philosophical, minds, ethics, inflammation, affected, abdomen, ulcer, circulation, heart, ships, fleet, sea, shore, Admiral, vessels, land, boats, admiral, sister, child, tears, pleasure, daughters, loves, wont, sigh, warm, sentence, clause, syllable, singular, examples, clauses, syllables, provinces, princes, nations, imperial, possessions, invasion, women, Quebec, Women, Iroquois, husbands, thirty, whom, steam, engines, power, piston, boilers, plant, supplied, chimney, lines, points, direction, planes, Lines, scale, sections, extending