Scalable Inference for Logistic-Normal Topic Models

Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng and Bo Zhang
Department of Computer Science and Technology, Tsinghua University

AllIMG Group Meeting, Tsinghua University, Nov. 13, 2013
Topic Modeling

• Given $D$ documents
• Vocabulary size is $V$
• bag-of-words
• Learn $K$ latent topics (code / feature)
• Documents are mixture of topics

```
<table>
<thead>
<tr>
<th>Document</th>
<th>Word</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D \times V$</td>
<td>$D \times K$</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Topic</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K \times V$</td>
</tr>
</tbody>
</table>
```
Topic Modeling

- **Motivation 1:** Dimensionality reduction
- **Motivation 2:** Semantics of topics / visualization
- **Motivation 3:** Feature learning

\[ D \times V = D \times K \times K \times V \]

**Diagram:**
- Document \( D \times V \)
- Topic \( D \times K \)
- Word \( K \times V \)
- Dictionary
- Feature Representation

\[ K \ll V \]
Topic Modeling

- Motivation 1: Dimensionality reduction
- Motivation 2: Semantics of topics / visualization
- Motivation 3: Feature learning

\[ D \times V = D \times K \times K \times V \]

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

Document, Word $\rightarrow$ Topics
Image, Bag of SIFT feature $\rightarrow$ Objects
User, User behavior $\rightarrow$ User characteristics
Topic Modeling

- Motivation 1: Dimensionality reduction
- Motivation 2: Semantics of topics / visualization
- **Motivation 3: Feature learning**

\[ D \times V = D \times K \times K \times V \]
Latent Dirichlet Allocation (LDA)

LDA generating process:
For each topic \( k \)
- Draw \( \phi_k \sim \text{Dir}(\beta) \), \( \phi_k \in \mathbb{R}^V \)
For each document \( d \)
- Draw \( \theta_d \sim \text{Dir}(\alpha) \), \( \theta_d \in \mathbb{R}^K \)
- For each position \( n \)
  - Draw \( z_{dn} \sim \text{Mult}(\theta_d) \)
  - Draw \( w_{dn} \sim \text{Mult}(\phi_{z_{dn}}) \)

\[ D \times V = D \times K \times K \times V \]

(Blei et al., JMLR 2003)
Latent Dirichlet Allocation (LDA)

Likelihood:

\[ l(\theta_d, \Phi | C_d^w) = \left( \sum_z \theta_dz \Phi_{zv} \right)^{C_d^w} \]

where \(C_d^w=\text{number of } \{w_{dn}=w\}\)

Maximum Likelihood Estimation:

\[
\min_{\Theta, \Phi} \text{loss}(C, \Theta \Phi)
\]

s.t. \(\sum_k \theta_{dk} = 1, \sum_w \Phi_{kw} = 1, \theta_{dk} \geq 0, \Phi_{kw} \geq 0\),

where \(\text{loss}(C, \Theta \Phi) = -\log \prod_{d,w} l(\Theta, \Phi | C_d^w)\)

\[ = - \sum_{d,w} C_d^w \log(\theta_d \Phi_{.,w}) \]

(Blei et al., JMLR 2003)
Latent Dirichlet Allocation (LDA)

- Usually tackled by approximate Bayesian Inference methods
  - Variational Inference
  - MCMC methods

```
\begin{align*}
\mathbf{C} & \quad \text{Word} \\
D \times V & \\
\mathbf{\Theta} & \quad \text{Topic} \\
D \times K & \\
\mathbf{\Phi} & \quad \text{Word} \\
K \times V & \\
\end{align*}
```

(blei et al., JMLR 2003)
Problem 1: Model flexibility

• How to model correlation of topics?
• Dirichlet: independent topic components

\[ \theta_d \sim \text{Dir}(\alpha) \propto \prod_{k} \theta_{dk}^{\alpha_k-1} \]

• Logistic normal: model correlation by \( \Sigma \)

\[ \theta_d = \text{softmax}(\eta_d) \]
where \( \eta_d \sim N(\mu, \Sigma) \)

\[ \theta_{dk} = \text{softmax}(\eta_d)_k = \frac{\exp(\eta_{dk})}{\sum_j \exp(\eta_{dj})} \]

mean parameter link natural parameter
Correlated Topic Models (CTM)

- **LDA generating process:**
  For each topic $k$
    - Draw $\phi_k \sim \text{Dir}(\beta)$
  For each document $d$
    - Draw $\theta_d \sim \text{Dir}(\alpha)$
    - For each position $n$
      - Draw $z_{dn} \sim \text{Mult}(\theta_d)$
      - Draw $w_{dn} \sim \text{Mult}(\phi_{z_{dn}})$

- **CTM generating process:**
  Draw $\mu, \Sigma \sim \text{NIW}(\mu_0, W, \rho, \kappa)$
  For each topic $k$
    - Draw $\phi_k \sim \text{Dir}(\beta)$
  For each document $d$
    - Draw $\eta_d \sim \text{N}(\mu, \Sigma)$, $\theta_d = \text{softmax}(\eta_d)$
    - For each position $n$
      - Draw $z_{dn} \sim \text{Mult}(\theta_d)$
      - Draw $w_{dn} \sim \text{Mult}(\phi_{z_{dn}})$

(Lafferty et al., NIPS 2005)
Problem 2: Speed

- **Big** data
- Wikipedia
  - 4.3M articles
- Facebook
  - 1.11B Users
- Google
  - 1 Trillion web items

- **Slow** algorithms (200 topics)
- VB CTM
  - 2K articles / hr
- Gibbs LDA
  - 36K articles / hr
- Online LDA:
  - 120k articles / hr
- Y! LDA (1000 machines)
  - 400M articles / hr
Gibbs Sampling

• 3-stage sampling
• \( p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \)
• \( p(\eta | \mu, \Sigma, Z) \)
• \( p(Z | W, \eta, \beta) \) integrate out \( \Phi \)

\( K \): number of topics
\( D \): number of documents
\( V \): size of vocabulary
\( \theta_d = \text{softmax}(\eta_d) \)
\( C_k = \{c_k^w\}_{w=1}^V \)
\( C_k^w \): number of \( (z_{dn} = k, w_{dn} = w) \)
\( C_d^w \): number of \( (w_{dn} = w) \)
\( C_d^k \): number of \( (z_{dn} = k) \)
\( \delta(\beta) = \frac{\prod_{i=1}^V \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^V \beta_i)} \)
Gibbs Sampling

\[ p(Z \mid W, \eta, \beta) \propto p(W \mid Z, \beta) p(Z \mid \eta) \]
\[ = p(Z \mid \eta) \int p(W \mid Z, \Phi) p(\Phi \mid \beta) \, d\Phi \]

\[ = \prod_{d=1}^{D} \left( \prod_{n=1}^{N_d} \frac{e^{\eta_d^{z_{dn}}}}{\sum_{j=1}^{K} e^{\eta_d^j}} \right) \prod_{k=1}^{K} \delta(C_k + \beta) \]

\[ p(z_{dn} = k \mid Z_{-dn}, W, \eta) \propto e^{\eta_d^k} \frac{C_{k,n}^{w_{dn}} + \beta_{w_{dn}}}{\sum_j^{V} C_{k,n}^j + \sum_j^{V} \beta_j} \]

prior, likelihood

\[ p(\mu, \Sigma \mid \eta, \mu_0, W_0, \rho, \kappa) \]
\[ p(\eta \mid \mu, \Sigma, Z) \]
\[ p(Z \mid W, \eta, \beta) \]

\( K \): number of topics
\( D \): number of documents
\( V \): size of vocabulary
\( \theta_d = \text{softmax}(\eta_d) \)
\( C_k = \{C_k^w\}_{w=1}^{V} \)
\( C_k^w \): number of \( z_{dn} = k, w_{dn} = w \)
\( C_d^w \): number of \( w_{dn} = w \)
\( C_d^k \): number of \( z_{dn} = k \)
\( \delta(\beta) = \frac{\prod_{i=1}^{V} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{V} \beta_i)} \)
Gibbs Sampling

\[ p(Z \mid W, \eta, \beta) \]
\[ \propto p(W \mid Z, \beta) p(Z \mid \eta) \]
\[ = p(Z \mid \eta) \int_{\Phi} \left( p(W \mid Z, \Phi) p(\Phi, \beta) \right) d\Phi \]

\[ = \prod_{d=1}^{D} \left( \prod_{n=1}^{N_d} \frac{e^{\eta_d^{z_{dn}}}}{\sum_{j=1}^{K} e^{\eta_d^{j}}} \right) \prod_{k=1}^{K} \frac{\delta(C_k + \beta)}{\delta(\beta)} \]

\[ p(z_{dn} = k \mid Z_{-dn}, W, \eta) \]
\[ \propto e^{\eta_d^k} \frac{C_{w_{dn}}^{k} + \beta_{w_{dn}}}{\sum_{j=1}^{V} C_{k, -n}^{j} + \sum_{j=1}^{V} \beta_j} \]

\[ \Phi \]
\[ \beta \]
\[ Z \]
\[ W \]

\[ p(\mu, \Sigma \mid \eta, \mu_0, W_0, \rho, \kappa) \]
\[ p(\eta \mid \mu, \Sigma, Z) \]
\[ p(Z \mid W, \eta, \beta) \]

\[ \begin{align*}
K: & \text{ number of topics} \\
D: & \text{ number of documents} \\
V: & \text{ size of vocabulary} \\
\theta_d: = & \text{ softmax}(\eta_d) \\
C_k: = & \{C_k^w\}_{w=1}^V \\
C_k^w: & \text{ number of } (z_{dn} = k, w_{dn} = w) \\
C_d^w: & \text{ number of } (w_{dn} = w) \\
C_d^k: & \text{ number of } (z_{dn} = k) \\
\delta(\beta): = & \frac{\prod_{i=1}^{V} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{V} \beta_i)}
\end{align*} \]

\( C_{k, -n}^{w} \) is very sparse

prior \quad likelihood
Gibbs Sampling

\[
p(z_{dn} = k | Z_{-dn}, W, \eta) \\
\propto e^{\eta_d^k} \frac{C_{k,-n}^W + \beta_{wdn}}{\sum_j^V C_j^k \eta_j - \sum_j^V \beta_j} \\
+ e^{\eta_d^k} \frac{C_{k,-n}^V + e^{\eta_d^k} \beta_{wdn}}{e^{\eta_d^k} \beta_{wdn}} \\
\]

where \( P = \sum_j^V C_j^k \eta_j + \sum_j^V \beta_j \)

- Toss a coin, front:back = \( \sum_k e^{\eta_d^k} C_{k,-n}^w \cdot \sum_k e^{\eta_d^k} \beta_{wdn} \)
- If front
  - Sample from Mult(\( e^{\eta_d^k} C_{k,-n}^w \) / normalization)
- If back
  - Sample from Mult(\( e^{\eta_d^k} C_{k,-n}^w \) / normalization)

\( K \): number of topics
\( D \): number of documents
\( V \): size of vocabulary
\( \theta_d = \text{softmax}(\eta_d) \)
\( C_k = \{C_k^w\}_{w=1}^V \)
\( C_k^w \): number of \( (z_{dn} = k, w_{dn} = w) \)
\( C_d^w \): number of \( (w_{dn} = w) \)
\( C_d^k \): number of \( (z_{dn} = k) \)

(\text{Smola et al., VLDB 2010})
Gibbs Sampling

\[ p(\eta | \mu, \Sigma, Z) \propto p(Z | \eta) p(\eta | \mu, \Sigma) \]
\[ = \prod_{d=1}^{D} \left( \prod_{n=1}^{N_d} \frac{e^{\eta_{dn}^d}}{\sum_{j=1}^{K} e^{\eta_{dn}^j}} \right) N(\eta_d | \mu, \Sigma) \]

\[ p(\eta_d^k | \eta_d^{-k}, Z, W) \]
\[ \propto \prod_{n=1}^{N_d} \frac{e^{\eta_{dn}^d}}{\sum_{j=1}^{K} e^{\eta_{dn}^j}} N(\eta_d^k | \eta_d^{-k} \mu, \Sigma) \]

\[ \propto \frac{(e^{\eta_{dn}^d})_{C_k^d}}{\left( \sum_{j=1}^{K} e^{\eta_{dn}^j} \right)^{N_d}} N(\eta_d^k | \eta_d^{-k} \mu, \Sigma) \]

\[ \propto \frac{(e^{\rho_d^k})_{C_k^d}}{(1 + e^{\rho_d^k})^{N_d}} N(\eta_d^k | \mu_d^k, \sigma_d^k) \]

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \]
\[ p(\eta | \mu, \Sigma, Z) \]
\[ p(Z | W, \eta, \beta) \]

\( K \): number of topics
\( D \): number of documents
\( V \): size of vocabulary
\( \theta_d = \text{softmax}(\eta_d) \)
\( C_k = \{C_{k}^{w}\}_{w=1}^{V} \)
\( C_{k}^{w} \): number of \((z_{dn} = k, w_{dn} = w)\)
\( C_{d}^{w} \): number of \((w_{dn} = w)\)
\( C_{d}^{k} \): number of \((z_{dn} = k)\)
\( \rho_d^k = \eta_d^k - \log(\sum_{j \neq k} e^{\eta_d^j}) \)
\( \mu_d^k = \mu_k - \Lambda_{kk}^{-1} \Lambda_{k-k}(\eta_d^{-k} - \mu^{-k}) \)
\( \sigma_d^k = \Lambda_{kk}^{-1} \)

\[ \beta \rightarrow \Phi \]
\[ K \]

\[ \mu_0, W_0, \beta \]

\[ \mu, \Sigma \rightarrow \eta \rightarrow Z \rightarrow W \]

\[ N_d \]

\[ D \]
Gibbs Sampling

\[ p(\eta_d^k | \eta_d^{-k}, Z, W) \]

\[ \alpha \frac{\left( e^{\rho_d^k} \right)^{C_d^k}}{(1 + e^{\rho_d^k})^{N_d}} N(\eta_d^k | \mu_d^k, \sigma_k^2) \]

• \( p(\eta_d^k | \eta_d^{-k}, Z, W, \lambda_d^k) = N(\gamma_d^k, (\tau_d^k)^2) \)

• \( p(\lambda_d^k | \eta_d) = PG(N_d, \rho_d^k) \)

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \]
\[ p(\eta | \mu, \Sigma, Z) \]
\[ p(Z | W, \eta, \beta) \]

\( K \): number of topics
\( D \): number of documents
\( V \): size of vocabulary
\( \theta_d = \text{softmax}(\eta_d) \)
\( C_k = \{C_w^w\}_{w=1}^V \)
\( C_k^w \): number of \( z_{dn} = k, w_{dn} = w \)
\( C_d^w \): number of \( w_{dn} = w \)
\( C_d^k \): number of \( z_{dn} = k \)
\( \rho_d^k = \eta_d^k - \log(\sum_{j \neq k} e^{\eta_d^j}) \)
\( \mu_d^k = \mu_k - \Lambda_{kk}^{-1}\Lambda_{kk}^{-1}(\eta_d^{-k} - \mu_{-k}) \)
\( \sigma_d^k = \Lambda_{kk}^{-1} \)

(Polson et al., arXiv 2013)
Gibbs Sampling

\[ \mu, \Sigma \sim NIW(\mu_0, W_0, \rho, \kappa) \]

\[ \Sigma \sim IW(W_0, \kappa) \]

- Draw \( x_1, \ldots, x_K \sim N(0, \frac{W_0}{\kappa}) \)
- \( \Sigma = \sum_i (x_i - \bar{x})(x_i - \bar{x})^T \)
- \( \mu \sim N(\mu_0, \frac{\Sigma}{\rho}) \)

\[ p(\mu, \Sigma \mid \eta, \mu_0, W_0, \rho, \kappa) \]
\[ p(\eta \mid \mu, \Sigma, Z) \]
\[ p(Z \mid W, \eta, \beta) \]

- \( K \): number of topics
- \( D \): number of documents
- \( V \): size of vocabulary
- \( \theta_d = \text{softmax}(\eta_d) \)
- \( C_k = \{C_k^w\}_{w=1}^V \)
- \( C_k^w \): number of \( z_{dn} = k, w_{dn} = w \)
- \( C_d^w \): number of \( w_{dn} = w \)
- \( C_d^k \): number of \( z_{dn} = k \)
- \( \rho_d^k = \eta_d^k - \log(\sum_{j \neq k} e^{\eta_d^j}) \)
- \( \mu_d^k = \mu_k - \Lambda_{kk}^{-1}(\eta_d^k - \mu_{-k}) \)
- \( \sigma_d^k = \Lambda_{kk}^{-1} \)

\[ p(\mu, \Sigma, \eta, W_0, \rho, \kappa) \]

Diagram:
- \( \mu_0, W_0, \rho, \kappa \)
- \( \mu, \Sigma \)
- \( \eta \)
- \( Z \)
- \( W \)
- \( \beta \to \Phi \)
- \( \lambda \)
Gibbs Sampling

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \sim \text{NIW}(\mu', W', \rho', \kappa') \]

\[ \mu' = \frac{\rho}{\rho + D} \mu_0 + \frac{D}{\rho + D} \bar{\eta} \]

\[ W' = W + \sum_d (\eta_d - \bar{\eta})(\eta_d - \bar{\eta})^T \]

\[ + \frac{\rho D}{\rho + D} (\bar{\eta} - \mu_0)(\bar{\eta} - \mu_0)^T \]

\[ \rho' = \rho + D \]

\[ \kappa' = \kappa + D \]

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \]

\[ p(\eta | \mu, \Sigma, Z) \]

\[ p(Z | W, \eta, \beta) \]

\[ K: \text{number of topics} \]

\[ D: \text{number of documents} \]

\[ V: \text{size of vocabulary} \]

\[ \theta_d = \text{softmax}(\eta_d) \]

\[ C_k = \{C_k^w\}_{w=1}^V \]

\[ C_k^w: \text{number of } (z_{dn} = k, w_{dn} = w) \]

\[ C_d^w: \text{number of } (w_{dn} = w) \]

\[ C_d^k: \text{number of } (z_{dn} = k) \]

\[ \rho_d^k = \eta_d^k - \log(\sum_{j \neq k} e^{\eta_d^j}) \]

\[ \mu_d^k = \mu_k - \Lambda_{kk}^{-1} \Lambda_{k\cdot \cdot} (\eta_d^k - \mu_d^k) \]

\[ \sigma_d^k = \Lambda_{kk}^{-1} \]

\[ K \]

\[ \beta \rightarrow \Phi \]

\[ \mu_0, W_0, \rho, \kappa \]

\[ \mu, \Sigma \rightarrow \eta \rightarrow Z \rightarrow W \]

\[ N_d \]

\[ D \]

\[ \lambda \]
Gibbs Sampling

for every iteration

for every document d

for every position n

\[ O(\bar{N}DK) - O(\bar{N}Ds(K)) \]

\[ z_{dn} \sim p(z_{dn} = k \mid Z_{-dn}, W, \eta) \]

for every document d

for every topic k

\[ O(DK) \]

\[ \eta_d^k \sim N\left(\gamma_d^k, (\tau_d^k)^2\right) \]

\[ \mu_d^k = \mu_k - \Lambda_{kk}^{-1} \Lambda_{k-k} (\eta_d^k - \mu_{-k}) \]

\[ \sigma_d^k = \Lambda_{kk}^{-1} \]

\[ \lambda_d^k \sim PG(N_d, \rho_d^k) \]

\[ \mu, \Sigma \sim NIW(\mu', W', \rho', \kappa') \]

Typically setting:

\[ D \sim \text{Millions} \]

\[ K = 500 - 2000 \]

\[ \bar{N} = 100 - 1000 \]

\[ s(\bar{K}) = 10 - 100 \]
Gibbs Sampling

for every iteration
  for every document d
    for every position n
      \( z_{dn} \sim p(z_{dn} = k \mid Z_{-dn}, W, \eta) \)

\( O(\bar{N}DK) - O(\bar{N}Ds(K)) \)

Typically setting:
- \( D \sim \text{Millions} \)
- \( K = 500 - 2000 \)
- \( \bar{N} = 100 - 1000 \)
- \( s(\bar{K}) = 10 - 100 \)

\( O(DK^2) \)

\( \eta_d^k \sim N \left( \gamma_d^k, (\tau_d^k)^2 \right) \)

\( \mu_d^k = \mu_k - \Lambda_{kk}^{-1} \Lambda_{k-k} (\eta_{-k}^k - \mu_{-k}) \)

\( \sigma_d^k = \Lambda_{kk}^{-1} \)

\( \lambda_d^k \sim PG \left( N_d, \rho_d^k \right) \)

\( \mu, \Sigma \sim NIW \left( \mu', W', \rho', \kappa' \right) \)

\( O(DK^2) \)
A random variable $X$ has a Polya-Gamma distribution with parameters $a > 0$ and $c \in \mathbb{R}$, if

$$X \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

where $g_k \sim Ga(a, 1)$ are gamma random variables. By computing the truncated sum of Eq. 1, we can obtain an approximate sampler

$$X_{truncated} = \frac{1}{2\pi^2} \sum_{k=1}^{K} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

(Polson et al., arXiv 2012)
Sample from PG distribution

\[ p(\eta_d^k | \eta_d^{-k}, \mathbf{z}, \mathbf{w}, \lambda_d^k) \sim N(\gamma_d^k, (\tau_d^k)^2) \]
\[ p(\lambda_d^k | \eta_d ) \sim \text{PG}(N_d, \rho_d^k) \]

however, this approximation sampler is biased. [1] proposed a sampler which corrects the bias by multiplying a constant

\[ X_{\text{truncated}} = \frac{\mathbb{E}[X]}{\mathbb{E}[X_{\text{truncated}}]} \]  (3)

where \( \mathbb{E}[X] = \frac{a}{2c} \tanh(\frac{c}{2}) \) and \( \mathbb{E}[X_{\text{truncated}}] = \frac{1}{2\pi^2} \sum_{k=1}^{K} \frac{a}{(k-1/2)^2 + c^2/(4\pi^2)} \), according to [3, 1]. Denote this approach as \text{truncated}_K.

(Dunson et al., ICML 2012)
Sample from PG distribution

\[ p(\eta^k_d | \eta^{-k}_d, \textbf{z}, \textbf{w}, \lambda^k_d) \sim N \left( \gamma^k_d, (\tau^k_d)^2 \right) \]

\[ p(\lambda^k_d | \eta_d) \sim PG(N_d, \rho^k_d) \]

[4] proposed a precise sampling algorithm for Polya-Gamma distributions

\[ X_{\text{precise}} \overset{D}{=} \sum_{n=1}^{a} X_n \] (4)

where \( X_n \sim PG(1, c) \) are i.i.d. samples. Denote this approach of precise. Draw samples from \( PG(1, c) \) can be done in \( O(1) \).[4]. However, \( a \) is document length \( N_d \) in logistic-normal topic models, since \( N_d \) is quite large, \( O(N_d) \) sampler is too slow. In this paper we draw \( K < a \) samples instead. Denote this approach as \( \text{pg1}_K \), note that \( \text{pg1}_K = \text{precise} \).

(Polson et al., arXiv 2013)
Sample from PG distribution

\[ p(\eta_d^k|\eta_{-d}^k, Z, W, \lambda_d^k) \sim N(\gamma_d^k, (\tau_d^k)^2) \]
\[ p(\lambda_d^k|\eta_d) \sim PG(N_d, \rho_d^k) \]

Notes that \( a = N_d \) is large, \( X \) is sum of i.i.d. random variables. There is another approximation by the central limit theorem

\[ X_{\text{gaussian}} \sim \mathcal{N}(\mu, \sigma^2) \quad (5) \]

<table>
<thead>
<tr>
<th>method</th>
<th>precise distribution?</th>
<th>precise mean?</th>
<th>precise variance?</th>
<th>time complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>truncated_K</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>( O(K) )</td>
</tr>
<tr>
<td>precise</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>( O(a) )</td>
</tr>
<tr>
<td>pg1_K</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>( O(K) )</td>
</tr>
<tr>
<td>gaussian</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>( O(1) )</td>
</tr>
</tbody>
</table>
Sample from PG distribution

\[ p(\eta_d^k|\eta_{d^-}^k, Z, W, \lambda_d^k) \sim N(\gamma_d^k, (\tau_d^k)^2) \]

\[ p(\lambda_d^k|\eta_d) \sim PG(N_d, \rho_d^k) \]

Table 3: Comparison for different PG samplers. Parameters are same as Fig. 1 in the paper.

<table>
<thead>
<tr>
<th>method</th>
<th>m</th>
<th>samples/second</th>
<th>\text{Var}[\lambda]</th>
<th>KS(\lambda)</th>
<th>\text{E}[\eta]</th>
<th>KS(\eta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>precise</td>
<td>-</td>
<td>1,602</td>
<td>6.65</td>
<td>-</td>
<td>1.0459</td>
<td>-</td>
</tr>
<tr>
<td>pg1</td>
<td>1</td>
<td>1,449,280</td>
<td>6.63</td>
<td>0.1146</td>
<td>1.0450</td>
<td>0.0146</td>
</tr>
<tr>
<td>pg1</td>
<td>2</td>
<td>757,576</td>
<td>6.66</td>
<td>0.0810</td>
<td>1.0467</td>
<td>0.0088</td>
</tr>
<tr>
<td>pg1</td>
<td>4</td>
<td>400,000</td>
<td>6.65</td>
<td>0.0562</td>
<td>1.0454</td>
<td>0.0080</td>
</tr>
<tr>
<td>pg1</td>
<td>8</td>
<td>215,517</td>
<td>6.67</td>
<td>0.0391</td>
<td>1.0463</td>
<td>0.0051</td>
</tr>
<tr>
<td>pg1</td>
<td>16</td>
<td>111,139</td>
<td>6.67</td>
<td>0.0259</td>
<td>1.0461</td>
<td>0.0041</td>
</tr>
<tr>
<td>pg1</td>
<td>32</td>
<td>56,721</td>
<td>6.66</td>
<td>0.0176</td>
<td>1.0450</td>
<td>0.0055</td>
</tr>
<tr>
<td>pg1</td>
<td>64</td>
<td>28,769</td>
<td>6.65</td>
<td>0.0123</td>
<td>1.0450</td>
<td>0.0049</td>
</tr>
<tr>
<td>truncated</td>
<td>1</td>
<td>3,846,150</td>
<td>15.49</td>
<td>0.1024</td>
<td>1.0241</td>
<td>0.0732</td>
</tr>
<tr>
<td>truncated</td>
<td>2</td>
<td>2,127,660</td>
<td>10.45</td>
<td>0.0558</td>
<td>1.0371</td>
<td>0.0350</td>
</tr>
<tr>
<td>truncated</td>
<td>4</td>
<td>1,111,110</td>
<td>8.37</td>
<td>0.0281</td>
<td>1.0415</td>
<td>0.0174</td>
</tr>
<tr>
<td>truncated</td>
<td>8</td>
<td>578,035</td>
<td>7.44</td>
<td>0.0140</td>
<td>1.0429</td>
<td>0.0087</td>
</tr>
<tr>
<td>truncated</td>
<td>16</td>
<td>313,480</td>
<td>7.04</td>
<td>0.0076</td>
<td>1.0441</td>
<td>0.0044</td>
</tr>
<tr>
<td>truncated</td>
<td>32</td>
<td>165,289</td>
<td>6.84</td>
<td>0.0039</td>
<td>1.0437</td>
<td>0.0043</td>
</tr>
<tr>
<td>truncated</td>
<td>64</td>
<td>84,962</td>
<td>6.76</td>
<td>0.0027</td>
<td>1.0449</td>
<td>0.0026</td>
</tr>
<tr>
<td>gaussian</td>
<td>-</td>
<td>6,250,000</td>
<td>6.66</td>
<td>0.0036</td>
<td>1.0458</td>
<td>0.0024</td>
</tr>
</tbody>
</table>
Parallelization

for every iteration
for every document $d$
for every position $n$

$$z_{dn} \sim p(z_{dn} = k \mid Z_{-dn}, W, \eta)$$

for every document $d$
for every topic $k$

$$\eta^k_d \sim N \left( \gamma^k_d, \tau^k_d \right)$$

$$\mu^k_d = \mu_k - \Lambda^{-1}_{kk} \Lambda_{k-k} (\eta^k_d - \mu_k)$$

$$\sigma^k_d = \Lambda^{-1}_{kk}$$

$$\lambda^k_d \sim PG (N_d, \rho^k_d)$$

$$\mu, \Sigma \sim NIW (\mu', W', \rho', \kappa')$$
Parallelization

\[ p(z_{dn} = k \mid Z_{-dn}, W, \eta) \propto e^{\eta_d} \frac{C_{w_{dn}}^w + \beta_{w_{dn}}}{\sum_{j=1}^{V} C_{k,-n}^j + \sum_{j=1}^{V} \beta_j} \]

Synchronize \( C_k^w \) between nodes

\( C_k^w \) changes slowly

(Ahmed et al., WSDM 2012)
Parallelization

\[ p(\eta \mid \mu, \Sigma, Z) = \prod_d p(\eta_d \mid \mu, \Sigma, Z_d) \]

- Independent
- Just distribute document across nodes
Parallelization

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \sim \text{NIW}(\mu', W', \rho', \kappa') \]

\[ \mu' = \frac{\rho}{\rho + D} \mu_0 + \frac{D}{\rho + D} \bar{\eta} \]

\[ W' = W + \sum_d (\eta_d - \bar{\eta})(\eta_d - \bar{\eta})^T \]

\[ + \frac{\rho D}{\rho + D} (\bar{\eta} - \mu_0)(\bar{\eta} - \mu_0)^T \]

\[ \rho' = \rho + D \]

\[ \kappa' = \kappa + D \]

\[ p(\mu, \Sigma | \eta, \mu_0, W_0, \rho, \kappa) \]

\[ p(\eta | \mu, \Sigma, Z) \]

\[ p(Z | W, \eta, \beta) \]

- \( K \): number of topics
- \( D \): number of documents
- \( V \): size of vocabulary
- \( \theta_d = \text{softmax}(\eta_d) \)
- \( C_k = \{C_k^w\}_{w=1}^V \)
- \( C_k^w \): number of \( (z_{dn} = k, w_{dn} = w) \)
- \( C_k^w \): number of \( (w_{dn} = w) \)
- \( C_k \): number of \( (z_{dn} = k) \)
- \( \rho_d^k = \eta_d^k - \log(\sum_{j \neq k} e^{\eta_d^j}) \)
- \( \mu_d^k = \mu_k - \Lambda_{kk}^{-1} \Lambda_{k-k} (\eta_{d}^{-k} - \mu_{-k}) \)
- \( \sigma_d^k = \Lambda_{kk}^{-1} \)

Diagram:

- \( \beta \)
- \( \Phi \)
- \( K \)
- \( \mu_0, W_0, \rho, \kappa \)
- \( \mu, \Sigma \)
- \( \eta \)
- \( Z \)
- \( W \)
- \( \lambda \)
- \( N_d \)
- \( D \)
# Experiments

<table>
<thead>
<tr>
<th>Data set</th>
<th>D</th>
<th>K</th>
<th>vCTM</th>
<th>gCTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS</td>
<td>1.2K</td>
<td>100</td>
<td>1.9hr</td>
<td>8.9 min</td>
</tr>
<tr>
<td>20NG</td>
<td>11K</td>
<td>200</td>
<td>16hr</td>
<td>9 min</td>
</tr>
<tr>
<td>NYTimes</td>
<td>285K</td>
<td>400</td>
<td>-</td>
<td>0.5 hr</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>6M</td>
<td>1,000</td>
<td>-</td>
<td>17 hr</td>
</tr>
</tbody>
</table>
Experiments

Figure 5: (a)(b): Perplexity and training time of vCTM, single-core gCTM, and multi-core gCTM on the NIPS data set; (c)(d): Perplexity and training time of single-machine gCTM, multi-machine gCTM, and multi-machine Y!LDA on the NYTimes data set.
Experiments

Figure 7: Scalability analysis. We set $M = 8, 16, 24, 32, 40$ so that each machine processes 150K documents.
Problem 2: Speed

- Big data
- Wikipedia
  4.3M articles
- Facebook
  1.11B Users
- Google
  1 Trillion web items
- Slow ML algorithms (200 topics)
- VB CTM
  0.5K articles / hr
- Gibbs LDA
  36K articles / hr
- Online LDA:
  120k articles / hr
- Gibbs CTM (40 machines)
  2.7M articles / hr
- Y! LDA (1000 machines)
  400M articles / hr
Conclusion

• CTM
  • More flexible than LDA

• Inference
  • Much faster
  • Scalable
Future Work

• Speed
  • Online
  • Low-rank approximation
  • Adaptive updating?
  • More sparsity?

• Extension
  • DTM
  • Infinite-CTM
  • ...


Selected References

• A. Ahmed et al., Scalable inference in latent variable models. WSDM 2012
• D. Blei et al., Correlated topic models. NIPS 2006
• D. Blei et al., Latent Dirichlet allocation. JMLR 2003
• N. Polson et al., Bayesian inference for logistic models using Polya-Gamma latent variables