Scalable Deep Poisson Factor Analysis for Topic Modeling

Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, Lawrence Carin

Duke University

July 9th, 2015
**Problem of interest:** How to develop deep generative models for documents that are represented in bag-of-words form?

- **Directed Graphical Models:**
  - Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
  - Focused Topic Model (FTM) (Williamson et al., 2010)
  - Poisson Factor Analysis (PFA) (Zhou et al., 2012)

- **Going “Deep”?**
  - Hierarchical tree-structured topic models
  - nested Chinese Restaurant Process (nCRP) (Blei et al., 2004)
  - Hierarchical Dirichlet Process (HDP) (Teh et al., 2006)
  - nested Hierarchical Dirichlet Process (nHDP) (Paisley et al., 2015)

- How about we want to model general topic correlations?
Undirected Graphical Models:
- Replicated Softmax Model (RSM) (Salakhutdinov and Hinton, 2009b)
- One generalization of the Restricted Boltzmann Machine (RBM) (Hinton, 2002)

Going Deep?
- Deep Belief Networks (DBN) (Hinton et al., 2006; Hinton and Salakhutdinov, 2011)
- Deep Boltzmann Machines (DBM) (Salakhutdinov and Hinton, 2009a; Srivastava et al., 2013)

Topics are not defined “properly”.
Introduction

Main idea:

- Poisson Factor Analysis (PFA) + Deep Sigmoid Belief Network (SBN) or Restricted Boltzmann Machine (RBM).
- PFA is employed to interact with data at the bottom layer.
- Deep SBN or RBM serve as a flexible prior for revealing topic structure.

![Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a deep Boltzmann machine.](image)

**Figure:** Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a deep Boltzmann machine.
Poisson Factor Analysis: (Zhou et al., 2012)

- We represent a discrete matrix \( X \in \mathbb{Z}_+^{P \times N} \) containing counts from \( N \) documents and \( P \) words as

\[
X = \text{Pois}(\Phi(\Theta \circ H^{(1)})) .
\]  

1. Each column of \( \Phi, \phi_k \), encodes the relative importance of each word in topic \( k \).
2. Each column of \( \Theta, \theta_n \), contains relative topic intensities specific to document \( n \).
3. Each column of \( H^{(1)}, h_n^{(1)} \), defines a sparse set of topics associated with each document.
Poisson Factor Analysis: (Zhou et al., 2012)

- We construct PFAs by placing Dirichlet priors on $\phi_k$ and gamma priors on $\theta_n$.

$$x_{pn} = \sum_{k=1}^{K} x_{pkn}, \quad x_{pkn} \sim \text{Pois}(\phi_{pk}\theta_{kn}h_{kn}^{(1)}),$$

with priors specified as $\phi_k \sim \text{Dir}(a_{\phi}, \ldots, a_{\phi})$, $\theta_{kn} \sim \text{Gamma}(r_k, p_n/(1 - p_n))$, $r_k \sim \text{Gamma}(\gamma_0, 1/c_0)$, and $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$.

- Previously, a beta-Bernoulli process prior is defined on $h_{n}^{(1)}$, assuming topic independence (Zhou and Carin, 2015).

- The novelty in our models comes from the prior for $h_{n}^{(1)}$. 

Presented by David Carlson (Duke)
Structured Priors on the Latent Binary matrix:

- Assume $h_n^{(1)} \in \{0, 1\}^{K_1}$, we define another hidden set of units $h_n^{(2)} \in \{0, 1\}^{K_2}$ placed at a layer “above” $h_n^{(1)}$.

- **Modeling with the RBM**: (Undirected)

  $$- E(h_n^{(1)}, h_n^{(2)}) = (h_n^{(1)})^\top c^{(1)} + (h_n^{(1)})^\top W^{(1)} h_n^{(2)} + (h_n^{(2)})^\top c^{(2)}.$$  

- **Modeling with the SBN** (Neal, 1992): (Directed)

  $$p(h_{k_2n}^{(2)} = 1) = \sigma(c_{k_2}^{(2)}), 
  \quad (4)$$

  $$p(h_{k_1n}^{(1)} = 1| h_n^{(2)}) = \sigma \left( (w_{k_1}^{(1)})^\top h_n^{(1)} + c_{k_1}^{(1)} \right). \quad (5)$$
Model Formulation

Going Deep?
- Add multiple layers of SBNs or RBMs.

Figure: Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a deep Boltzmann machine.
Challenge: Designing scalable Bayesian inference algorithms.
Solutions: Scaling up inference by stochastic algorithms.
- Applying Bayesian conditional density filtering algorithm (Guhaniyogi et al., 2014).
- Extending recently proposed work on stochastic gradient thermostats (Ding et al., 2014).
Bayesian conditional density filtering (BCDF):

- Repeatedly updating the surrogate conditional sufficient statistics (SCSS) using the current mini-batch.
- Drawing samples from the conditional posterior distributions of model parameters, based on SCSS.
- "stochastic Gibbs-style" updates.

Input: text documents, i.e., a count matrix $X$.
Initialize $\psi_g^{(0)}$ randomly and set $S_g^{(0)}$ all to zero.

for $t = 1$ to $\infty$ do

Get one mini-batch $X^{(t)}$.
Initialize $\psi_g^{(t)} = \psi_g^{(t-1)}$, and $S_g^{(t)} = S_g^{(t-1)}$.
Initialize $\psi_l^{(t)}$ randomly.

for $s = 1$ to $S$ do

Gibbs sampling for DPFA on $X^{(t)}$.
Collect samples $\psi_g^{1:S}, \psi_l^{1:S}$ and $S_g^{1:S}$.
end for

Set $\psi_g^{(t)} = \text{mean}(\psi_g^{1:S})$, and $S_g^{(t)} = \text{mean}(S_g^{1:S})$.
end for

$\psi_g$: global parameters
$\psi_l$: local hidden variables
$S_g$: SCSS for $\psi_g$
Stochastic Gradient Nöse-Hoover Thermostats (SGNHT):

- Extending *Hamiltonian Monte Carlo* using stochastic gradient.
- Introducing *thermostat* to maintain system temperature.
- Adaptively *absorbing* stochastic gradient noise.
- The motion of the particles in the system are defined by the stochastic differential equations (SDE)

\[
\begin{align*}
    d\Psi_g &= \mathbf{v} \, dt, \\
    d\mathbf{v} &= \tilde{f}(\Psi_g) \, dt - \xi \mathbf{v} \, dt + \sqrt{D} \, d\mathcal{W}, \\
    d\xi &= \left(\frac{1}{M} \mathbf{v}^T \mathbf{v} - 1\right) \, dt,
\end{align*}
\]

where $\Psi_g \in \mathbb{R}^M$ are model parameters, $\mathbf{v} \in \mathbb{R}^M$ are the momentum variables, $\tilde{f}(\Psi_g) \triangleq -\nabla_{\Psi_g} \tilde{U}(\Psi_g)$, and $\tilde{U}(\Psi_g)$ is the negative log-posterior.
Extension:

- Extending the SGNHT by introducing multiple thermostat variables \((\xi_1, \ldots, \xi_M)\) into the system such that each \(\xi_i\) controls one degree of the particle momentum.

- The proposed SGNHT is defined by the following SDEs

\[
\begin{align*}
\mathrm{d}\Psi_g &= \mathbf{v}\mathrm{d}t, \\
\mathrm{d}\mathbf{v} &= \tilde{f}(\Psi_g)\mathrm{d}t - \Xi\mathbf{v}\mathrm{d}t + \sqrt{D}\mathrm{d}\mathcal{W}, \\
\mathrm{d}\Xi &= (\mathbf{q} - \mathbf{I})\mathrm{d}t,
\end{align*}
\]

where \(\Xi = \text{diag}(\xi_1, \xi_2, \ldots, \xi_M)\), \(\mathbf{q} = \text{diag}(v_1^2, \ldots, v_M^2)\)

**Theorem**

The equilibrium distribution of the SDE system in (7) is

\[
p(\Psi_g, \mathbf{v}, \Xi) \propto \exp\left(-\frac{1}{2} \mathbf{v}^\top \mathbf{v} - U(\Psi_g) - \frac{1}{2} \text{tr}\left\{ (\Xi - D)^\top (\Xi - D) \right\} \right).
\]
Stochastic Gradient Nöse-Hoover Thermostats (SGNHT):

**Input:** text documents, *i.e.*, a count matrix $X$.

**Random Initialization.**

```
for $t = 1$ to $\infty$ do
    $\Psi_{g}^{(t+1)} = \Psi_{g}^{(t)} + \nu^{(t)} h$.

    $\nu^{(t+1)} = \tilde{f}(\Psi_{g}^{(t+1)}) h - \Xi^{(t)} \nu^{(t)} h + \sqrt{2Dh} \mathcal{N}(0, I)$.

    $\Xi^{(t+1)} = \Xi^{(t)} + (q^{(t+1)} - I)h$, where $q = \text{diag}(v_{1}^{2}, \ldots, v_{M}^{2})$.

end for
```
Stochastic Gradient Nőse-Hoover Thermostats (SGNHT):

\[ \text{Input: } \text{text documents, i.e., a count matrix } X. \]

\[ \text{Random Initialization.} \]

\[ \text{for } t = 1 \text{ to } \infty \text{ do} \]

\[ \Psi_g^{(t+1)} = \Psi_g^{(t)} + \nu^{(t)} h. \]

\[ \nu^{(t+1)} = \tilde{f}(\Psi_g^{(t+1)}) h - \Xi^{(t)} \nu^{(t)} h + \sqrt{2D} h \mathcal{N}(0, I). \]

\[ \Xi^{(t+1)} = \Xi^{(t)} + (q^{(t+1)} - I) h, \text{ where } q = \text{diag}(v_1^2, \ldots, v_M^2). \]

\[ \text{end for} \]

Discussion:

- **BCDF**: ease of implementation, but prefers the conditional densities for all the parameters.

- **SGNHT**: more general and robust, fast convergence.
Datasets:

- **20 Newsgroups**: 20K documents with a vocabulary size of 2K.
- **RCV1-v2**: 800K documents with a vocabulary size of 10K.
- **Wikipedia**: 10M documents with a vocabulary size of 8K.
Quantitative Evaluation:

**Table:** 20 Newsgroups.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>METHOD</th>
<th>Dim</th>
<th>PERP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPFA-SBN-t</td>
<td>GIBBS</td>
<td>128-64-32</td>
<td>827</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>GIBBS</td>
<td>128-64-32</td>
<td>846</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>SGNHT</td>
<td>128-64-32</td>
<td>846</td>
</tr>
<tr>
<td>DPFA-RBM</td>
<td>SGNHT</td>
<td>128-64-32</td>
<td>896</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>BCDF</td>
<td>128-64-32</td>
<td>905</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>GIBBS</td>
<td>128-64</td>
<td>851</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>SGNHT</td>
<td>128-64</td>
<td>850</td>
</tr>
<tr>
<td>DPFA-RBM</td>
<td>SGNHT</td>
<td>128-64</td>
<td>893</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>BCDF</td>
<td>128-64</td>
<td>896</td>
</tr>
<tr>
<td>LDA</td>
<td>GIBBS</td>
<td>128</td>
<td>893</td>
</tr>
<tr>
<td>NB-FTM</td>
<td>GIBBS</td>
<td>128</td>
<td>887</td>
</tr>
<tr>
<td>RSM</td>
<td>CD5</td>
<td>128</td>
<td>877</td>
</tr>
<tr>
<td>NHDP</td>
<td>sVB</td>
<td>(10,10,5)◊</td>
<td>889</td>
</tr>
</tbody>
</table>

**Table:** RCV1-v2 & Wikipedia.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>METHOD</th>
<th>Dim</th>
<th>RCV</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPFA-SBN</td>
<td>SGNHT</td>
<td>1024-512-256</td>
<td>964</td>
<td>770</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>SGNHT</td>
<td>512-256-128</td>
<td>1073</td>
<td>799</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>SGNHT</td>
<td>128-64-32</td>
<td>1143</td>
<td>876</td>
</tr>
<tr>
<td>DPFA-RBM</td>
<td>SGNHT</td>
<td>128-64-32</td>
<td>920</td>
<td>942</td>
</tr>
<tr>
<td>DPFA-SBN</td>
<td>BCDF</td>
<td>128-64-32</td>
<td>1149</td>
<td>986</td>
</tr>
<tr>
<td>LDA</td>
<td>BCDF</td>
<td>128</td>
<td>1179</td>
<td>1059</td>
</tr>
<tr>
<td>NB-FTM</td>
<td>BCDF</td>
<td>128</td>
<td>1155</td>
<td>991</td>
</tr>
<tr>
<td>RSM</td>
<td>CD5</td>
<td>128</td>
<td>1171</td>
<td>1001</td>
</tr>
<tr>
<td>NHDP</td>
<td>sVB</td>
<td>(10,5,5)◊</td>
<td>1041</td>
<td>932</td>
</tr>
</tbody>
</table>
Experiments

Quantitative Evaluation:

![Graph showing Perplexity vs Iteration Number](image)

Sensitivity Analysis:

![Graph showing Perplexity vs #Documents Seen](image)

Figure: Perplexities. (Left) 20 News. (Middle) RCV1-v2. (Right) Wikipedia.
## Experiments

### Topics we learned on 20 Newsgroups:

<table>
<thead>
<tr>
<th>T1</th>
<th>T3</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T14</th>
<th>T15</th>
<th>T19</th>
<th>T21</th>
<th>T24</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>people</td>
<td>group</td>
<td>world</td>
<td>evidence</td>
<td>game</td>
<td>israel</td>
<td>software</td>
<td>files</td>
<td>team</td>
</tr>
<tr>
<td>hit</td>
<td>real</td>
<td>groups</td>
<td>country</td>
<td>claim</td>
<td>games</td>
<td>israeli</td>
<td>modem</td>
<td>file</td>
<td>players</td>
</tr>
<tr>
<td>runs</td>
<td>simply</td>
<td>reading</td>
<td>countries</td>
<td>people</td>
<td>win</td>
<td>jews</td>
<td>port</td>
<td>ftp</td>
<td>player</td>
</tr>
<tr>
<td>good</td>
<td>world</td>
<td>newsgroup</td>
<td>germany</td>
<td>argument</td>
<td>cup</td>
<td>arab</td>
<td>mac</td>
<td>program</td>
<td>play</td>
</tr>
<tr>
<td>season</td>
<td>things</td>
<td>pro</td>
<td>nazi</td>
<td>agree</td>
<td>hockey</td>
<td>jewish</td>
<td>serial</td>
<td>format</td>
<td>teams</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T25</th>
<th>T26</th>
<th>T29</th>
<th>T40</th>
<th>T41</th>
<th>T43</th>
<th>T50</th>
<th>T54</th>
<th>T55</th>
<th>T64</th>
</tr>
</thead>
<tbody>
<tr>
<td>god</td>
<td>fire</td>
<td>people</td>
<td>wrong</td>
<td>image</td>
<td>boston</td>
<td>problem</td>
<td>card</td>
<td>windows</td>
<td>turkish</td>
</tr>
<tr>
<td>exist</td>
<td>fbi</td>
<td>life</td>
<td>doesn't</td>
<td>program</td>
<td>toronto</td>
<td>work</td>
<td>video</td>
<td>dos</td>
<td>armenian</td>
</tr>
<tr>
<td>exist</td>
<td>koresh</td>
<td>death</td>
<td>jim</td>
<td>application</td>
<td>montreal</td>
<td>problems</td>
<td>memory</td>
<td>file</td>
<td>armenians</td>
</tr>
<tr>
<td>human</td>
<td>children</td>
<td>kill</td>
<td>agree</td>
<td>widget</td>
<td>chicago</td>
<td>system</td>
<td>mhz</td>
<td>win</td>
<td>turks</td>
</tr>
<tr>
<td>atheism</td>
<td>batf</td>
<td>killing</td>
<td>quote</td>
<td>color</td>
<td>pittsburgh</td>
<td>fine</td>
<td>bit</td>
<td>ms</td>
<td>armenia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T65</th>
<th>T69</th>
<th>T78</th>
<th>T81</th>
<th>T91</th>
<th>T94</th>
<th>T112</th>
<th>T118</th>
<th>T120</th>
<th>T126</th>
</tr>
</thead>
<tbody>
<tr>
<td>truth</td>
<td>window</td>
<td>drive</td>
<td>makes</td>
<td>question</td>
<td>code</td>
<td>children</td>
<td>people</td>
<td>men</td>
<td>sex</td>
</tr>
<tr>
<td>true</td>
<td>server</td>
<td>disk</td>
<td>power</td>
<td>answer</td>
<td>mit</td>
<td>father</td>
<td>make</td>
<td>women</td>
<td>sexual</td>
</tr>
<tr>
<td>point</td>
<td>display</td>
<td>scsi</td>
<td>make</td>
<td>means</td>
<td>comp</td>
<td>child</td>
<td>person</td>
<td>man</td>
<td>cramers</td>
</tr>
<tr>
<td>fact</td>
<td>manager</td>
<td>hard</td>
<td>doesn't</td>
<td>true</td>
<td>unix</td>
<td>mother</td>
<td>things</td>
<td>hand</td>
<td>gay</td>
</tr>
<tr>
<td>body</td>
<td>client</td>
<td>drives</td>
<td>part</td>
<td>people</td>
<td>source</td>
<td>son</td>
<td>feel</td>
<td>world</td>
<td>homosexual</td>
</tr>
</tbody>
</table>
Visualization:
Sports, Computers, and Politics/Law.

Figure: Graphs induced by the correlation structure learned by DPFA-SBN for the 20 Newsgroups.
**Model:** Deep Poisson Factor Analysis
- PFA is employed to interact with data at the bottom layer.
- Deep SBN or RBM serve as a flexible prior for revealing topic structure.

**Scalable Inference:**
- Bayesian conditional density filtering.
- Stochastic gradient thermostats.

https://github.com/zhegan27/dpfa_icml2015
Questions?


