Scalable Deep Poisson Factor Analysis for Topic Modeling

Zhe Gan ZHE.GAN@DUKE.EDU
Changyou Chen CHANGYOU.CHEN@DUKE.EDU
Ricardo Henao RICARDO.HENAO@DUKE.EDU
David Carlson DAVID.CARLSON@DUKE.EDU
Lawrence Carin LCARIN@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

Abstract

A new framework for topic modeling is developed, based on deep graphical models, where interactions between topics are inferred through deep latent binary hierarchies. The proposed multi-layer model employs a deep sigmoid belief network or restricted Boltzmann machine, the bottom binary layer of which selects topics for use in a Poisson factor analysis model. Under this setting, topics live on the bottom layer of the model, while the deep specification serves as a flexible prior for revealing topic structure. Scalable inference algorithms are derived by applying Bayesian conditional density filtering algorithm, in addition to extending recently proposed work on stochastic gradient thermostats. Experimental results on several corpora show that the proposed approach readily handles very large collections of text documents, infers structured topic representations, and obtains superior test perplexities when compared with related models.

1. Introduction

Considerable research effort has been devoted to developing probabilistic models for documents. In the context of topic modeling, a popular approach is latent Dirichlet allocation (LDA) (Blei et al., 2003), a directed graphical model that aims to discover latent topics (word distributions) in collections of documents that are represented in bag-of-words form. Recent work focuses on linking observed word counts in a document to latent nonnegative matrix factorization, via a Poisson distribution, termed Poisson factor analysis (PFA) (Zhou et al., 2012). Different choices of priors on the latent nonnegative matrix factorization can lead to equivalent marginal distributions to LDA, as well as to the Focused Topic Model (FTM) of Williamson et al. (2010).

Additionally, hierarchical ("deep") tree-structured topic models have been developed by using structured Bayesian nonparametric priors, including the nested Chinese restaurant process (nCRP) (Blei et al., 2004), and the recently proposed nested hierarchical Dirichlet process (nHDP) (Paisley et al., 2015). The nCRP is limited because it requires that each document select topics from a single path in a tree, while the nHDP allows each document to access the entire tree by defining priors over a *base tree*. However, the relationship between two paths in these models is only explicitly given on shared parent nodes.

Another alternative for topic modeling is to develop undirected graphical models, such as the Replicated Softmax Model (RSM) (Salakhutdinov & Hinton, 2009a), based on a generalization of the restricted Boltzmann machine (RBM) (Hinton, 2002). Also closely related to the RBM is the neural autoregressive density estimator (DocNADE) (Larochelle & Lauly, 2012), a neural-network-based method, that has been shown to outperform the RSM.

Deep models, such as the Deep Belief Network (DBN) (Hinton et al., 2006), the Deep Boltzmann Machine (DBM) (Salakhutdinov & Hinton, 2009b), and layered Bayesian networks (Kingma & Welling, 2014; Mnih & Gregor, 2014; Danilo et al., 2014; Gan et al., 2015) are becoming popular, as they consistently obtain state-of-the-art performances on a variety of machine learning tasks. A popular theme in this direction of work is to extend shallow topic models to deep counterparts. In such a setting, documents arise from a cascade of layers of latent variables. For instance, DBNs and DBMs have been generalized to model documents by utilizing the RBM as a building block (Hinton & Salakhutdinov, 2011; Srivastava et al., 2013).

Combining ideas from traditional Bayesian topic modeling and deep models, we propose a new deep generative model

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

for topic modeling, in which the Bayesian PFA is employed to interact with the data at the bottom layer, while the Sigmoid Belief Network (SBN) (Neal, 1992), a directed graphical model closely related to the RBM, is utilized to buildup binary hierarchies. Furthermore, our model is not necessarily restricted to SBN modules, and it is shown how an undirected model such as the RBM can be incorporated into the framework as well.

Compared with the original DBN and DBM, our proposed model: (*i*) tends to infer a more compact representation of the data, due to the "explaining away" effect described by Hinton et al. (2006); (*ii*) allows for more direct exploration of the effect of a single deep hidden node through ancestral sampling; and (*iii*) can be easily incorporated into larger probabilistic models in a modular fashion. Compared with the nCRP and nHDP, our proposed model only infers topics at the bottom layer, but defines a flexible prior to capture high-order relationships between topics via a deep binary hierarchical structure. In practice, this translates into better perplexities and very interesting topic correlations, although not in a tree representation as in nCRP or nHDP.

Another important contribution we present is to develop two scalable Bayesian learning algorithms for our model: one based on the recently proposed *Bayesian conditional density filtering* (BCDF) algorithm (Guhaniyogi et al., 2014), and the other based on the *stochastic gradient Nóse-Hoover thermostats* (SGNHT) algorithm (Ding et al., 2014). We extend the SGNHT by introducing additional *thermostat variables* into the dynamic system, increasing the stability and convergence when compared to the original SGNHT algorithm.

2. Model Formulation

2.1. Poisson Factor Analysis

Given a discrete matrix $\mathbf{X} \in \mathbb{Z}_+^{P\times N}$ containing counts from N documents and P words, Poisson factor analysis (Zhou et al., 2012) assumes the entries of $\mathbf{X}$ are summations of $K < \infty$ latent counts, each produced by a latent factor (in the case of topic modeling, a hidden topic). We represent $\mathbf{X}$ using the following factor model

$$\mathbf{X} = \text{Pois}(\mathbf{\Phi}(\mathbf{\Theta} \circ \mathbf{H}^{(1)}))\,, \tag{1}$$

where $\mathbf{\Phi}$ is the factor loading matrix. Each column of $\mathbf{\Phi}$, $\boldsymbol{\phi}_k \in \triangle_P$, encodes the relative importance of each word in topic k, with $\triangle_P$ representing the P-dimensional simplex. $\mathbf{\Theta} \in \mathbb{R}_+^{K\times N}$ is the factor score matrix. Each column, $\boldsymbol{\theta}_n$, contains relative topic intensities specific to document n. $\mathbf{H}^{(1)} \in \{0,1\}^{K\times N}$ is a latent binary feature matrix. Each column, $\boldsymbol{h}_n^{(1)}$, defines a sparse set of topics associated with each document. For the single-layer PFA, the use of the superscript (1) on $\boldsymbol{h}_n^{(1)}$ is unnecessary; we introduce this notation here in preparation for the subsequent deep model, for which $\boldsymbol{h}_n^{(1)}$ will correspond to the associated first-layer latent binary units. The symbol $\circ$ represents the Hadamard, or element-wise multiplication of two matrices. The factor scores for document n are $\boldsymbol{\theta}_n \circ \boldsymbol{h}_n^{(1)}$.

A wide variety of algorithms have been developed by constructing PFAs with different prior specifications (Zhou & Carin, 2015). If $\mathbf{H}^{(1)}$ is an all-ones matrix, LDA is recovered from (1) by employing Dirichlet priors on $\boldsymbol{\phi}_k$ and $\boldsymbol{\theta}_n$, for $k = 1,\dots,K$ and $n = 1,\dots,N$, respectively. This version of LDA is referred to as Dir-PFA by Zhou et al. (2012). For our proposed model, we construct PFAs by placing Dirichlet priors on $\boldsymbol{\phi}_k$ and gamma priors on $\boldsymbol{\theta}_n$. This is summarized as,

$$x_{pn} = \textstyle\sum_{k=1}^K x_{pnk}\,,\;\; x_{pnk} \sim \text{Pois}(\phi_{pk}\theta_{kn}h_{kn}^{(1)})\,, \tag{2}$$

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

The novelty in our model comes from the prior for the binary feature matrix $\mathbf{H}^{(1)}$. Previously, Zhou & Carin (2015) proposed a beta-Bernoulli process prior on the columns $\{\boldsymbol{h}_n^{(1)}\}_{n=1}^N$ with $p_n = 0.5$. This model was called NB-FTM, tightly related with the focused topic model (FTM) (Williamson et al., 2010). In the work presented here, we construct $\mathbf{H}^{(1)}$ from a deep structure based on the SBN (or RBM) with binary latent units.

2.2. Structured Priors on the Latent Binary Matrix

The second part of our model consists of a deep structure for a binary hierarchy. To this end, we employ the SBN (or RBM). In the following we start by describing a single-layer model with SBN (or RBM), and then we generalize it to a deep model.

Modeling with the SBN We assume the latent vector for document n, $\boldsymbol{h}_n^{(1)} \in \{0,1\}^{K_1}$. This matches most of the RBM and SBN literature, for which typically the *observed* data are binary. In our model, however, these binary variables are not observed; they are hidden and related to the data through the PFA in (2).

To construct a structured prior, we define another hidden set of units $\boldsymbol{h}_n^{(2)} \in \{0,1\}^{K_2}$ placed at a layer "above" $\boldsymbol{h}_n^{(1)}$. The layers are related through a set of weights defined by the matrix $\mathbf{W}^{(1)} = [\boldsymbol{w}_1^{(1)} \;\dots\; \boldsymbol{w}_{K_1}^{(1)}]^\top \in \mathbb{R}^{K_1\times K_2}$. An SBN model has the generative process,

$$p(h_{k_2n}^{(2)} = 1) = \sigma(c_{k_2}^{(2)})\,, \tag{3}$$

$$p(h_{k_1n}^{(1)} = 1|\boldsymbol{h}_n^{(2)}) = \sigma\left((\boldsymbol{w}_{k_1}^{(1)})^\top \boldsymbol{h}_n^{(1)} + c_{k_1}^{(1)}\right)\,, \tag{4}$$

where $h_{k_1n}^{(1)}$ and $h_{k_2n}^{(2)}$ are elements of $\boldsymbol{h}_n^{(1)}$ and $\boldsymbol{h}_n^{(2)}$, re-

spectively. The function $\sigma(x) \triangleq 1/(1+e^{-x})$ is the logistic function, and $c_{k_1}^{(1)}$ and $c_{k_2}^{(2)}$ are bias terms. The global parameters $\mathbf{W}^{(1)}$ are used to characterize the mapping from $\boldsymbol{h}_n^{(2)}$ to $\boldsymbol{h}_n^{(1)}$ for all documents.

Modeling with the RBM The SBN is closely related to the RBM, which is a Markov random field with the same bipartite structure as the SBN. The RBM defines a distribution over a binary vector that is proportional to the exponential of its *energy*, defined (using the same notation as in SBN) as $E(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)}) =$

$$-(\boldsymbol{h}_n^{(1)})^\top \boldsymbol{c}^{(1)} - (\boldsymbol{h}_n^{(1)})^\top \mathbf{W}^{(1)} \boldsymbol{h}_n^{(2)} - (\boldsymbol{h}_n^{(2)})^\top \boldsymbol{c}^{(2)} . \quad (5)$$

In the experiments we consider both the deep SBN and deep RBM for representation of the latent binary units, which are connected to topic usage in a given document.

Remark An important benefit of SBNs over RBMs is that in the former sparsity or shrinkage priors can be readily imposed on the global parameters $\mathbf{W}^{(1)}$, and fully Bayesian inference can be implemented as shown in Gan et al. (2015). The RBM relies on an approximation technique known as contrastive divergence (Hinton, 2002), for which prior specification for model parameters is limited.

2.3. Deep Architecture for Topic Modeling

Specifying a prior distribution on $\boldsymbol{h}_n^{(2)}$ as in (3) might be too restrictive in some cases. Alternatively, we can use another SBN prior for $\boldsymbol{h}_n^{(2)}$, in fact, we can add multiple layers as in Gan et al. (2015) to obtain a deep architecture,

$$p(\boldsymbol{h}_n^{(1)}, \ldots, \boldsymbol{h}_n^{(L)}) = p(\boldsymbol{h}_n^{(L)}) \prod_{\ell=2}^{L} p(\boldsymbol{h}_n^{(\ell-1)} | \boldsymbol{h}_n^{(\ell)}), \quad (6)$$

where L is the number of layers, $p(\boldsymbol{h}_n^{(L)})$ is the prior for the top layer defined as in (3), $p(\boldsymbol{h}_n^{(\ell-1)} | \boldsymbol{h}_n^{(\ell)})$ is defined as in (4), and the weights $\mathbf{W}^{(\ell)} \in \mathbb{R}^{K_\ell \times K_{\ell+1}}$ and biases $\boldsymbol{c}^{(\ell)} \in \mathbb{R}^{K_\ell}$ are omitted from the conditional distributions to keep notation uncluttered. A similar deep architecture may be designed for the RBM (Salakhutdinov & Hinton, 2009b).

Instead of employing the beta-Bernoulli specification for $\boldsymbol{h}_n^{(1)}$ as in the NB-FTM, which assumes independent topic usage probabilities, we propose using (6) instead as the prior for $\boldsymbol{h}_n^{(1)}$, thus

$$p(\boldsymbol{x}_n, \boldsymbol{h}_n) = p(\boldsymbol{x}_n | \boldsymbol{h}_n^{(1)}) p(\boldsymbol{h}_n^{(1)}, \ldots, \boldsymbol{h}_n^{(L)}) , \quad (7)$$

where $\boldsymbol{h}_n \triangleq \{\boldsymbol{h}_n^{(1)}, \ldots, \boldsymbol{h}_n^{(L)}\}$, and $p(\boldsymbol{x}_n | \boldsymbol{h}_n^{(1)})$ as in (2). The prior $p(\boldsymbol{h}_n^{(1)} | \boldsymbol{h}_n^{(2)} \ldots, \boldsymbol{h}_n^{(L)})$ can be seen as a flexible prior distribution over binary vectors that encodes high-order interactions across elements of $\boldsymbol{h}_n^{(1)}$. The graphical model for our model, Deep Poisson Factor Analysis (DPFA) is shown in Figure 1.

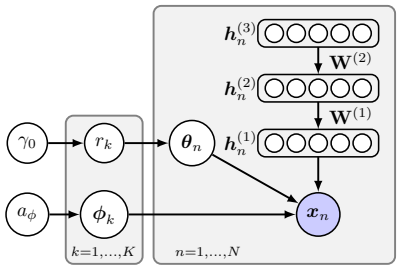

Figure 1. Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a *deep Boltzmann machine*.

3. Scalable Posterior Inference

We focus on learning our model with fully Bayesian algorithms, however, emerging large-scale corpora prohibit standard MCMC inference algorithms to be applied directly. For example, in the experiments, we consider the *RCV1-v2* and the *Wikipedia* corpora, which contain about 800K and 10M documents, respectively. Therefore, fast algorithms for big Bayesian learning are essential. While parallel algorithms based on distributed architectures such as the *parameter server* (Ho et al., 2013; Li et al., 2014) are popular choices, in the work presented here, we focus on another direction for scaling up inference by stochastic algorithms, where mini-batches instead of the whole dataset are utilized in each iteration of the algorithms. Specifically, we develop two stochastic Bayesian inference algorithms based on Bayesian conditional density filtering (Guhaniyogi et al., 2014) and stochastic gradient thermostats (Ding et al., 2014), both of which have theoretical guarantees in the sense of asymptotical convergence to the true posterior distribution.

3.1. Bayesian conditional density filtering

Bayesian conditional density filtering (BCDF) is a recently proposed stochastic algorithm for Bayesian online learning (Guhaniyogi et al., 2014), that extends Markov chain Monte Carlo (MCMC) sampling to streaming data. Sampling in BCDF proceeds by drawing from the conditional posterior distributions of model parameters, obtained by propagating surrogate conditional sufficient statistics (SCSS). In practice, we repeatedly update the SCSS using the current mini-batch and draw S samples from the conditional densities using, for example, a Gibbs sampler. This eliminates the need to load the entire dataset into memory, and provides computationally cheaper Gibbs updates. More importantly, it can be proved that BCDF leads to an approximation of the conditional distributions that produce samples from the correct target posterior asymptotically, once the entire dataset is seen (Guhaniyogi et al., 2014).

In the learning phase, we are interested in learning the global parameters $\boldsymbol{\Psi}_g = (\{\boldsymbol{\phi}_k\}, \{r_k\}, \gamma_0, \{\mathbf{W}^{(\ell)}, \boldsymbol{c}^{(\ell)}\})$.

Algorithm 1 BCDF algorithm for DPFA.

Input: text documents, *i.e.*, a count matrix $\mathbf{X}$.
Initialize $\mathbf{\Psi}_g^{(0)}$ randomly and set $\mathbf{S}_g^{(0)}$ all to zero.
for $t = 1$ **to** ∞ **do**
  Get one mini-batch $\mathbf{X}^{(t)}$.
  Initialize $\mathbf{\Psi}_g^{(t)} = \mathbf{\Psi}_g^{(t-1)}$, and $\mathbf{S}_g^{(t)} = \mathbf{S}_g^{(t-1)}$.
  Initialize $\mathbf{\Psi}_l^{(t)}$ randomly.
  for $s = 1$ **to** S **do**
    Gibbs sampling for DPFA on $\mathbf{X}^{(t)}$.
    Collect samples $\mathbf{\Psi}_g^{1:S}$, $\mathbf{\Psi}_l^{1:S}$ and $\mathbf{S}_g^{1:S}$.
  end for
  Set $\mathbf{\Psi}_g^{(t)} = \text{mean}(\mathbf{\Psi}_g^{1:S})$, and $\mathbf{S}_g^{(t)} = \text{mean}(\mathbf{S}_g^{1:S})$.
end for

Denote local variables as $\mathbf{\Psi}_l = (\mathbf{\Theta}, \mathbf{H}^{(\ell)})$, and let $\mathbf{S}_g$ represent the SCSS for $\mathbf{\Psi}_g$, the BCDF algorithm can be summarized in Algorithm 1. Specifically, we need to obtain the conditional densities, which can be readily derived granted the full local conjugacy of the proposed model. Using *dot notation* to represent marginal sums, *e.g.*, $x_{\cdot nk} \triangleq \sum_p x_{pnk}$, we can write the key conditional densities for (2) as (Zhou & Carin, 2015)

$$
\begin{aligned}
x_{pnk}|- &\sim \text{Multi}(x_{pn}; \zeta_{pn1}, \ldots, \zeta_{pnK}) \,, \\
\boldsymbol{\phi}_k|- &\sim \text{Dir}(a_\phi + x_{1\cdot k}, \ldots, a_\phi + x_{P\cdot k}) \,, \\
\theta_{kn}|- &\sim \text{Gamma}(r_k h_{kn}^{(1)} + x_{\cdot nk}, p_n) \,, \\
h_{kn}^{(1)}|- &\sim \delta(x_{\cdot nk} = 0)\text{Ber}\left(\tfrac{\tilde{\pi}_{kn}}{\tilde{\pi}_{kn}+(1-\pi_{kn})}\right) + \delta(x_{\cdot nk} > 0) \,,
\end{aligned}
$$

where $\tilde{\pi}_{kn} = \pi_{kn}(1-p_n)^{r_k}$, $\pi_{kn} = \sigma((\boldsymbol{w}_k^{(1)})^\top \boldsymbol{h}_n^{(2)} + c_k^{(1)})$, and $\zeta_{pnk} \propto \phi_{pk}\theta_{kn}$. Additional details are provided in the Supplementary Material. For the conditional distributions of $\mathbf{W}^{(\ell)}$ and $\mathbf{H}^{(\ell)}$, we use the same data augmentation technique as in Gan et al. (2015), where Pólya-Gamma (PG) variables $\gamma_{k_\ell n}^{(\ell)}$ (Polson et al., 2013) are introduced for hidden unit k_ℓ in layer ℓ corresponding to observation $\boldsymbol{v}_n$. Specifically, each $\gamma_{k_\ell n}^{(\ell)}$ has conditional posterior $\text{PG}(1, (\boldsymbol{w}_{k_\ell}^{(\ell)})^\top \boldsymbol{h}_n^{(\ell+1)} + c_{k_\ell}^{(\ell)})$. If we place a Gaussian prior $N(0, \sigma^2\mathbf{I})$ on $\boldsymbol{w}_{k_\ell}^{(\ell)}$, the posterior will still be Gaussian with covariance matrix $\mathbf{\Sigma}_{k_\ell}^{(\ell)} = [\sum_n \gamma_{k_\ell n}^{(\ell)} \boldsymbol{h}_n^{(\ell+1)} (\boldsymbol{h}_n^{(\ell+1)})^\top + \sigma^{-2}\mathbf{I}]^{-1}$ and mean $\boldsymbol{\mu}_{k_\ell}^{(\ell)} = \mathbf{\Sigma}_{k_\ell}^{(\ell)}[\sum_n (h_{k_\ell n}^{(\ell)} - 1/2 - c_{k_\ell}^{(\ell)}\gamma_{k_\ell n}^{(\ell)})\boldsymbol{h}_n^{(\ell+1)}]$. Furthermore, for $\ell > 1$, the conditional distribution of $h_{k_\ell n}^{(\ell)}$ can be obtained as[1]

$$
h_{k_\ell n}^{(\ell)} \sim \text{Bernoulli}\left(\sigma(d_{k_\ell n})\right) \,, \tag{8}
$$

where

$$
\begin{aligned}
d_{k_\ell n} &= (\boldsymbol{w}_{\cdot,k_\ell}^{(\ell-1)})^\top \boldsymbol{h}_n^{(\ell-1)} + (\boldsymbol{w}_{k_\ell}^{(\ell)})^\top \boldsymbol{h}_n^{(\ell+1)} + c_{k_\ell}^{(\ell)} \\
&- \frac{1}{2}\sum_{k_{\ell-1}} \left(w_{k_{\ell-1}k_\ell}^{(\ell-1)} + \gamma_{k_{\ell-1}n}^{(\ell-1)} (2\psi_{k_{\ell-1}n}^{\backslash k_\ell} w_{k_{\ell-1}k_\ell}^{(\ell-1)} + (w_{k_{\ell-1}k_\ell}^{(\ell-1)})^2) \right) \,,
\end{aligned}
$$

and $\psi_{k_{\ell-1}n}^{\backslash k_\ell} = \sum_{k'_\ell \neq k_\ell} w_{k_{\ell-1}k'_\ell}^{(\ell-1)} h_{k'_\ell n}^{(\ell)} + c_{k_{\ell-1}}^{(\ell-1)}$. Note that $\boldsymbol{w}_{\cdot,k_{\ell+1}}^{(\ell)}$ and $\boldsymbol{w}_{k_\ell}^{(\ell)}$ represents the $k_{\ell+1}$th column and the transpose of the k_ℓth row of $\mathbf{W}^{(\ell)}$, respectively. As can be seen, the conditional posterior distribution of $h_{k_\ell n}^{(\ell)}$ is both related to $\boldsymbol{h}_n^{(\ell-1)}$ and $\boldsymbol{h}_n^{(\ell+1)}$.

[1]Here and in the rest of the paper, whenever $\ell > L$, $\boldsymbol{h}_n^{(\ell)}$ is defined as a zero vector, for conciseness.

3.2. Stochastic gradient thermostats

Our second learning algorithm adopts the recently proposed SGNHT for large scale Bayesian sampling (Ding et al., 2014), which is more scalable and accurate than the previous BCDF algorithm. SGNHT generalizes the *stochastic gradient Langevin dynamics* (SGLD) (Welling & Teh, 2011) and the *stochastic gradient Hamiltonian Monte Carlo* (SGHMC) (Chen et al., 2014) by introducing momentum variables into the system, which is adaptively damped using a thermostat. The thermostat exchanges energy with the target system (*e.g.*, a Bayesian model) to maintain a constant temperature; this has the potential advantage of making the system jump out of local modes easier and reach the equilibrium state faster (Ding et al., 2014).

Specifically, let $\mathbf{\Psi}_g \in \mathbb{R}^M$ be model parameters[2] which corresponds to the location of particles in a physical system, $\boldsymbol{v} \in \mathbb{R}^M$ be the momentum of these particles, which are driven by stochastic forces $\tilde{f}$ defined as the negative stochastic gradient (evaluated on a subset of data) of a Bayesian posterior, *e.g.*, $\tilde{f}(\mathbf{\Psi}_g) \triangleq -\nabla_{\mathbf{\Psi}_g}\tilde{U}(\mathbf{\Psi}_g)$, where $\tilde{U}(\mathbf{\Psi}_g)$ is the negative log-posterior of a Bayesian model. The motion of the particles in the system are then defined by the following stochastic differential equations:

$$
\begin{aligned}
\mathrm{d}\mathbf{\Psi}_g = \boldsymbol{v}\mathrm{d}t \,, \qquad & \mathrm{d}\boldsymbol{v} = \tilde{f}(\mathbf{\Psi}_g)\mathrm{d}t - \xi\boldsymbol{v}\mathrm{d}t + \sqrt{D}\mathrm{d}\mathcal{W} \,, \\
& \mathrm{d}\xi = \left(\tfrac{1}{M}\boldsymbol{v}^T\boldsymbol{v} - 1\right)\mathrm{d}t \,,
\end{aligned} \tag{9}
$$

where t indexes time, $\mathcal{W}$ is the standard Wiener process, ξ is called the thermostat variable which ensures the system temperature to be constant, and D is the variance of the total noise injected into the system and is assumed to be constant.

It can be shown that under certain assumptions, the equilibrium distribution of system (9) corresponds to the model posterior (Ding et al., 2014). As a result, the SDE (9) can be solved by using the Euler-Maruyama scheme (Tuckerman, 2010), where a mini-batch of the whole data is used to evaluate the stochastic gradient $\tilde{f}$. Note only one thermostat variable ξ is used in the SDE system (9); this is not robust enough to control the system temperature well because of the high dimensionality of $\mathbf{\Psi}_g$. Based on the techniques in (Ding et al., 2014), we extend the SGNHT by introducing

[2]With a little abuse of notation but for conciseness, we use $\mathbf{\Psi}_g$ to denote the reparameterized version of the parameters (such that $\mathbf{\Psi}_g \in \mathbb{R}^M$) if any, required in SGNHT.

multiple thermostat variables $(\xi_1, \cdots, \xi_M)$ into the system such that each ξ_i controls one degree of the particle momentum. Intuitively, this allows energy to be exchanged between particles and thermostats more efficiently, thus driving the system to equilibrium states more rapidly. Empirically we have also verified the superiority of the proposed modification over the original SGNHT. Formally, let $\boldsymbol{\Xi} = \text{diag}(\xi_1, \xi_2, \cdots, \xi_M)$, $\mathbf{q} = \text{diag}(v_1^2, \cdots, v_M^2)$, we define our proposed SGNHT using the following SDEs

$$\begin{aligned} \mathrm{d}\boldsymbol{\Psi}_g = \boldsymbol{v}\mathrm{d}t\,, \qquad & \mathrm{d}\boldsymbol{v} = \tilde{f}(\boldsymbol{\Psi}_g)\mathrm{d}t - \boldsymbol{\Xi}\boldsymbol{v}\mathrm{d}t + \sqrt{D}\mathrm{d}\mathcal{W}\,, \\ & \mathrm{d}\boldsymbol{\Xi} = (\mathbf{q} - \mathbf{I})\,\mathrm{d}t\,, \end{aligned} \tag{10}$$

where $\mathbf{I}$ is the identity matrix. Interestingly, we are still able to prove that the equilibrium distribution of the above system corresponds to the model posterior.

Theorem 1 *The equilibrium distribution of the SDE system in (10) is* $p(\boldsymbol{\Psi}_g, \boldsymbol{v}, \boldsymbol{\Xi})$

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{v}^\top\boldsymbol{v} - U(\boldsymbol{\Psi}_g) - \frac{1}{2}tr\left\{(\boldsymbol{\Xi} - D)^\top(\boldsymbol{\Xi} - D)\right\}\right).$$

The proof of the theorem is provided in the Supplementary Material. By Theorem 1, it is straightforward to see that the marginal distribution $p(\boldsymbol{\Psi}_g)$ of $p(\boldsymbol{\Psi}_g, \boldsymbol{v}, \boldsymbol{\Xi})$ is exactly the posterior of our Bayesian model. As a result, again we can generate approximate samples from $p(\boldsymbol{\Psi}_g, \boldsymbol{v}, \boldsymbol{\Xi})$ using the Euler-Maruyama scheme and discard the auxiliary variables $\boldsymbol{v}$ and $\boldsymbol{\Xi}$.

Learning for the SBN-based model Our SBN-based model is illustrated in Figure 1. In the learning phase we are interested in learning the global parameters $\boldsymbol{\Psi}_g$, the same as in BCDF. The constraints inside the parameters $\{\boldsymbol{\phi}_k\}$, *i.e.*, $\sum_p \phi_{pk} = 1$, prevent the SGNHT from being applied directly. Although we can overcome this problem by using re-parameterization methods as in Patterson & Teh (2013), we find it converges better when considering information geometry for these parameters. As a result, we use stochastic gradient Riemannian Langevin dynamics (SGRLD) (Patterson & Teh, 2013) to sample the topic-word distributions $\{\boldsymbol{\phi}_k\}$, and use the SGNHT to sample the remaining parameters. Based on the data augmentation for x_{pn} above, Section 3.1 shows that the posteriors of $\{\boldsymbol{\phi}_k\}$'s are Dirichlet distributions. This enables us to apply the same scheme as the SGRLD for LDA (Patterson & Teh, 2013) to sample $\{\boldsymbol{\phi}_k\}$'s. More details are provided in the Supplementary Material.

The rest of the parameters can be straightforwardly sampled using the SGNHT algorithm. Specifically we need to calculate the stochastic gradients of $\mathbf{W}^{(\ell)}$ and $\boldsymbol{c}^{(\ell)}$ evaluated on a mini-batch of data (denote $\mathcal{D}$ as the index set of a mini-batch). Based on the model definition in (6), these can be calculated as

$$\begin{aligned} \frac{\partial\tilde{U}}{\partial\boldsymbol{w}_{k\ell}^{(\ell)}} &= \frac{N}{|\mathcal{D}|}\sum_{n\in\mathcal{D}}\mathbb{E}_{\boldsymbol{h}_n^{(\ell)},\boldsymbol{h}_n^{(\ell+1)}}\left[\left(\tilde{\sigma}_{k\ell n}^{(\ell)} - h_{k\ell n}^{(\ell)}\right)\boldsymbol{h}_n^{(\ell+1)}\right], \\ \frac{\partial\tilde{U}}{\partial c_{k\ell}^{(\ell)}} &= \frac{N}{|\mathcal{D}|}\sum_{n\in\mathcal{D}}\mathbb{E}_{\boldsymbol{h}_n^{(\ell)},\boldsymbol{h}_n^{(\ell+1)}}\left[\tilde{\sigma}_{k\ell n}^{(\ell)} - h_{k\ell n}^{(\ell)}\right], \end{aligned}$$

where $\tilde{\sigma}_{k\ell n}^{(\ell)} = \sigma((\boldsymbol{w}_{k\ell}^{(\ell)})^\top\boldsymbol{h}_n^{(\ell+1)} + c_{k\ell}^{(\ell)})$, and the expectation is taken over posteriors. As in the case of LDA (Patterson & Teh, 2013), no closed-form integrations can be obtained for the above gradients, we thus use Monte Carlo integration to approximate the quantity. Specifically, given $\{\boldsymbol{w}_{k\ell}^{(\ell)}, c_{k\ell}^{(\ell)}\}$, we are able to collect samples of the local variables $(\boldsymbol{h}_n^{(\ell)})_{n\in\mathcal{D}}$ by running a few Gibbs steps and then using these samples to approximate the intractable integrations. Exact conditional distributions for $h_{k\ell n}^{(\ell)}$ exist without variable augmentation, however, we found that this approach does not mix well due to the highly correlated structure of hidden variables. Instead, we sample $h_{k\ell n}^{(\ell)}$ based on the same augmentation used in BCDF, given in (8).

Learning for the RBM-based model As mentioned above, our RBM-based model is recovered when replacing the SBN with the RBM in Figure 1. Despite minor changes in the construction, the intractable normalizer which consists of model parameters (*e.g.*, $\mathbf{W}^{(\ell)}$) prohibits exact MCMC sampling from being applied. As a result, we develop an approximate learning algorithm that alternates between sampling $(\{\boldsymbol{\phi}_k\}, \{\gamma_k\}, \gamma_0\})$ and $(\{\mathbf{W}^{(\ell)}, \boldsymbol{c}^{(\ell)}\})$. Specifically, we use the same conditional posteriors as in the SBN-based model to sample the former, but use the *contrastive divergence* algorithm (CD-1) (Hinton, 2002) for the latter. One main difference of our CD-1 algorithm *w.r.t* the original one is that the inputs (*i.e.*, $\boldsymbol{h}_n^{(1)}$) are hidden variables. To make the CD-1 work, conditioned on other model parameters, we first sample $\boldsymbol{h}_n^{(1)}$ using the posterior given in Section 3.1, then conditioned on $\boldsymbol{h}_n^{(1)}$, we apply the original CD-1 algorithm to calculate the approximate gradients for $(\{\mathbf{W}^{(\ell)}, \boldsymbol{c}^{(\ell)}\})$, which are then used for a gradient descent step in SGNHT. In fact, the CD-1 is also a stochastic approximate algorithm, discussed in Yuille (2005), making it naturally fit into our SGNHT framework.

3.3. Discussion

Both the BCDF and SGNHT are stochastic inference algorithms, allowing the models to be applied to large-scale data. In terms of ease of implementation, BCDF beats SGNHT in most cases, especially when the model is conjugate and the domain of parameters is constrained (*e.g.*, variables on a simplex). However, in general BCDF is more restrictive than SGNHT. For example, BCDF prefers the conditional densities for all the parameters, which is unavail-

able in some cases. Furthermore, BCDF has the limitation of being unable to deal with some *big models* where the number of model parameters is large, for instance, when the dimension of the hidden variables from the SBN in our model is huge. Finally, the conditions for BCDF to converge to the true posterior are more restricted. Altogether, these reasons make SGNHT more robust than BCDF.

4. Related Work

In traditional Bayesian topic models, topic correlations are typically modeled with shallow structures, *e.g.*, the correlated topic model (Blei & Lafferty, 2007) with correlation between topic proportions imposed via the logistic normal distribution. There exist also some work on hierarchical ("deep") correlation modeling, *e.g.*, the hierarchical Dirichlet process (Teh et al., 2006), which models topic proportions hierarchically via a stack of DPs. The nested Chinese restaurant process (Blei et al., 2004) (nCRP) models topic hierarchies by defining a tree structure prior based on the Chinese restaurant process, and the nested hierarchical Dirichlet process (Paisley et al., 2015) extends the nCRP by allowing each document to be able to access all the paths in the tree. One major difference between these models and ours is that they focus on discovering topic hierarchies instead of modeling general topic correlations.

In the deep learning community, topic models are mostly built using the RBM as a building block. For example, Hinton & Salakhutdinov (2011) and Maaloe et al. (2015) extended the DBN for topic modeling, while a deep version of the RSM was proposed by Srivastava et al. (2013). More recent work focuses on employing deep directed generative models for topic modeling, *e.g.*, deep exponential families (Ranganath et al., 2015), a class of latent variable models extending the DBN by defining the distribution of hidden variables in each layer using the exponential family, instead of the restricted Bernoulli distribution.

In terms of learning and inference algorithms, most of existing Bayesian topic models rely on MCMC methods or variational Bayes algorithms, which are impractical when dealing with large scale data. Therefore, stochastic variational inference algorithms have been developed (Hoffman et al., 2010; Mimno et al., 2012; Wang & Blei, 2012; Hoffman et al., 2013). Although scalable and usually fast converging, one unfavorable shortcoming of stochastic variational inference algorithms is the mean-field assumption on the approximate posterior.

Another direction for scalable Bayesian learning relies on the theory from stochastic differential equations (SDE). Specifically, Welling & Teh (2011) proposed the first stochastic MCMC algorithm, called *stochastic gradient Langevin dynamics* (SGLD), for large scale Bayesian learning. In order to make the learning faster, Patterson & Teh (2013) generalized SGLD by considering information geometry (Girolami & Calderhead, 2011; Byrne & Girolami, 2013) of model posteriors. Furthermore, Chen et al. (2014) generalized the SGLD by a second-order Langevin dynamic, called *stochastic gradient Hamiltonian Monte Carlo* (SGHMC). This is the stochastic version of the well known Hamiltonian MCMC sampler. One problem with SGHMC is that the unknown stochastic noise needs to be estimated to make the sampler correct, which is impractical. *Stochastic gradient thermostats* algorithms (SGNHT) overcome this problem by introducing the thermostat into the algorithm, such that the unknown stochastic noise could be adaptively absorbed into the thermostat, making the sampler asymptotically exact. Given the advantages of the SGNHT, in this paper we extend it to a multiple thermostats setting, where each thermostat exchanges energy with a degree of freedom of the system. Empirically we show that our extension improves on the original algorithm.

5. Experiments

We present experimental results on three publicly available corpora: a relatively small, *20 Newsgroups*, a moderately large, Reuters Corpus Volume I (*RCV1-v2*), and a large one, *Wikipedia*. The first two corpora are the same as those used in Srivastava et al. (2013). Specifically, the *20 Newsgroups* corpus contains 18,845 documents with a total of 0.7M words and a vocabulary size of 2K. The data was partitioned chronologically into 11,314 training and 7,531 test documents. The *RCV1-v2* corpus contains 804,414 newswire articles. There are 103 topics that form a tree hierarchy. After preprocessing, we are left with about 75M words, with a vocabulary size of 10K. We randomly select 794,414 documents for training and 10,000 for testing. Finally, we downloaded 10M random documents from *Wikipedia* using scripts provided in Hoffman et al. (2010) and randomly selected 1K documents for testing. As in Hoffman et al. (2010); Patterson & Teh (2013), a vocabulary size of 7,702 was taken from the top 10K words in Project Gutenberg texts.

The DPFA model consisting of SBN is denoted as DPFA-SBN, while its RBM counterpart is denoted DPFA-RBM. The performance of DPFA is compared to that of the following models: LDA (Blei et al., 2003), NB-FTM (Zhou & Carin, 2015), nHDP (Paisley et al., 2015) and RSM (Salakhutdinov & Hinton, 2009a).

For all the models considered, we calculate the predictive perplexities on the test set as follows: holding the global model parameters fixed, for each test document we randomly partition the words into a 80/20% split. We learn document-specific "local" parameters using the 80% portion, and then calculate the predictive perplexities on the

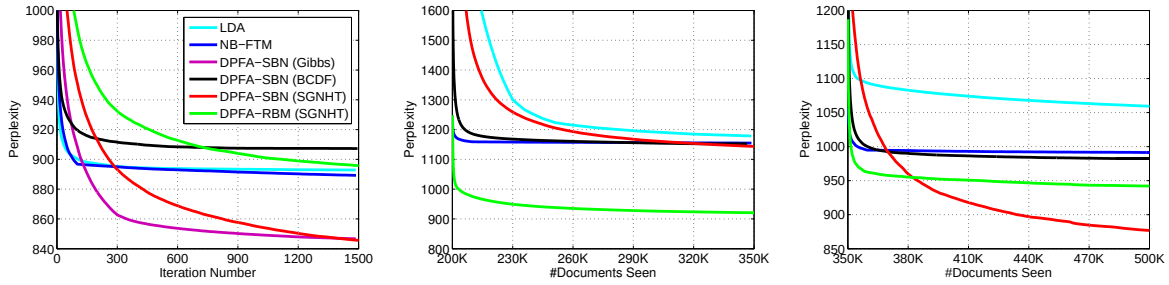

Figure 2. Predictive perplexities on a held-out test set as a function of training documents seen. The number of hidden units in each layer is 128, 64, 32, respectively. (Left) *20 Newsgroups*. (Middle) *RCV1-v2*. (Right) *Wikipedia*.

remaining 20% subset. Evaluation details are provided in the Supplementary Material.

For *20 Newsgroups* and *RCV1-v2* corpora, we use 2,000 mini-batches for burn-in followed by 1,500 collection samples to calculate test perplexities; while for the *Wikipedia* dataset, 3,500 mini-batches are used for burn-in. The mini-batch size for all stochastic algorithms is set to 100. To choose good parameters for SGNHT, *e.g.*, the step size and the variance of the injected noise, we randomly choose about 10% documents from the training data as validation set. For BCDF, 100 MCMC iterations are evaluated for each mini-batch, with the first 60 samples discarded. We set the hyperparameters of DPFA as $a_\phi = 1.01, c_0 = e_0 = 1, f_0 = 0.01$, and $p_n = 0.5$. The RSM is trained using convergence-divergence with step size 5 and a maximum of 10,000 iterations. For nHDP, we use the publicly available code from Paisley et al. (2015), in which stochastic variational Bayes (sVB) inference is implemented.

20 Newsgroups The results for the *20 Newsgroups* corpus are shown in Table 1. Perplexities are reported for our implementation of Gibbs sampling, BCDF and SGNHT, and the four considered competing methods. First, we examine the performance of different inference algorithms. As can be seen, for the same size model, *e.g.*, 128-64-32 (128 topics and 32 binary nodes on the top of the three-layer model), SGNHT can achieve essentially the same performance as Gibbs sampling, while BCDF is more likely to get trapped in a local mode. Next, we explore the advantage of employing deep models. Using three layers instead of two gives performance improvements in almost all the algorithms. In Gibbs sampling, there is an improvement of 36 units for the DPFA-SBN model, when a second layer is learned (NB-FTM is the one-hidden-layer DPFA). Adding the third hidden layer further improves the test perplexity.

Adding a sparsity-encouraging prior on $\mathbf{W}^{(\ell)}$ acts as a more stringent regularization that prevents overfitting, when compared with the commonly used L_2 norm (Gaussian prior). Furthermore, shrinkage priors have the effect of being able to effectively switch off the elements of $\mathbf{W}^{(\ell)}$, which benefits interpretability and helps to infer the number of units needed to represent the data. In our experiment, we observe that the DPFA-SBN model with the Student's t prior on $\mathbf{W}^{(\ell)}$ achieves a better test perplexity when compared with its counterpart without shrinkage.

RCV1-v2 & Wiki We present results for the *RCV1-v2* and *Wikipedia* corpora in Table 3. Direct Gibbs sampling in such a (big-data) setting is prohibitive, and is thus not discussed. First, we explore the effect of utilizing a larger deep network. For our DPFA-SBN model using the SGNHT algorithm, we observe that making the network 8 time larger in each hidden layer decreases the test perplexities by 155 and 84 units on *RCV1-v2* and *Wikipedia*, respectively. This demonstrates the ability of our stochastic inference algorithm to scale up both in terms of model and corpus size.

Both SBN and RBM can be utilized as the building block in our deep specification. For the *RCV1-v2* corpus, our best result is obtained by utilizing a three-layer deep Boltzmann machine. However, for the *20 Newsgroups* and *Wikipedia* corpora, with the same size model, we found empirically that the deep SBN achieves better performance.

Compared with nHDP, our DPFA models define a more flexible prior on topic interactions, and therefore in practice we also consistently achieve better perplexity results.

Table 1. Test perplexities for *20 Newsgroups*. "Dim" represents the number of hidden units in each layer, starting from the bottom. DPFA-SBN-*t* represents the DPFA-SBN model with Student's t prior on $\mathbf{W}^{(\ell)}$. ($\diamond$) represents the *base tree* size in nHDP.

MODEL	METHOD	DIM	PERP.
DPFA-SBN-*t*	GIBBS	128-64-32	**827**
DPFA-SBN	GIBBS	128-64-32	**846**
DPFA-SBN	SGNHT	128-64-32	**846**
DPFA-RBM	SGNHT	128-64-32	896
DPFA-SBN	BCDF	128-64-32	905
DPFA-SBN	GIBBS	128-64	851
DPFA-SBN	SGNHT	128-64	850
DPFA-RBM	SGNHT	128-64	893
DPFA-SBN	BCDF	128-64	896
LDA	GIBBS	128	893
NB-FTM	GIBBS	128	887
RSM	CD5	128	877
NHDP	SVB	$(10,10,5)^\diamond$	889

Table 2. Top words from the 30 topics corresponding to the graph in Figure 3, learned by DPFA-SBN from the *20Newsgroup* corpus.

T1	T3	T8	T9	T10	T14	T15	T19	T21	T24
year	people	group	world	evidence	game	israel	software	files	team
hit	real	groups	country	claim	games	israeli	modem	file	players
runs	simply	reading	countries	people	win	jews	port	ftp	player
good	world	newsgroup	germany	argument	cup	arab	mac	program	play
season	things	pro	nazi	agree	hockey	jewish	serial	format	teams
T25	**T26**	**T29**	**T40**	**T41**	**T43**	**T50**	**T54**	**T55**	**T64**
god	fire	people	wrong	image	boston	problem	card	windows	turkish
existence	fbi	life	doesn	program	toronto	work	video	dos	armenian
exist	koresh	death	jim	application	montreal	problems	memory	file	armenians
human	children	kill	agree	widget	chicago	system	mhz	win	turks
atheism	batf	killing	quote	color	pittsburgh	fine	bit	ms	armenia
T65	**T69**	**T78**	**T81**	**T91**	**T94**	**T112**	**T118**	**T120**	**T126**
truth	window	drive	makes	question	code	children	people	men	sex
true	server	disk	power	answer	mit	father	make	women	sexual
point	display	scsi	make	means	comp	child	person	man	cramer
fact	manager	hard	doesn	true	unix	mother	things	hand	gay
body	client	drives	part	people	source	son	feel	world	homosexual

We further show test perplexities as a function of documents processed during model learning in Figure 2. As can be seen, performance smoothly improves as the amount of data processed increases.

Table 3. Test perplexities on *RCV1-v2* and *Wikipedia*. "Dim" represents the number of hidden units in each layer, starting from the bottom. ($\diamond$) represents the *base tree* size in nHDP.

MODEL	METHOD	DIM	RCV	WIKI
DPFA-SBN	SGNHT	1024-512-256	964	**770**
DPFA-SBN	SGNHT	512-256-128	1073	799
DPFA-SBN	SGNHT	128-64-32	1143	876
DPFA-RBM	SGNHT	128-64-32	**920**	942
DPFA-SBN	BCDF	128-64-32	1149	986
LDA	BCDF	128	1179	1059
NB-FTM	BCDF	128	1155	991
RSM	CD5	128	1171	1001
NHDP	SVB	(10,5,5)$^\diamond$	1041	932

Sensitivity analysis We examined the sensitivity of the model performance with respect to batch sizes in SGNHT on the three corpora considered. We found that overall performance, both convergence speed and test perplexity, suffer considerably when the batch size is smaller than 10 documents. However, for batch sizes larger than 50 (100 for *RCV1-v2*) we obtain performances comparable to those shown in Tables 1 and 3. Additional details including test perplexity traces as a function of documents seen by the model are presented in the Supplementary Material.

Visualization We can obtain a visual representation of the topic structure implied by the deep component of our DPFA model by computing correlations between topics using the weight matrices, $\mathbf{W}^{(\ell)}$, learned by DPFA-SBN, *i.e*, we evaluate the covariance $\mathbf{W}^{(1)}\mathbf{W}^{(2)}(\mathbf{W}^{(1)}\mathbf{W}^{(2)})^\top$, then scale it accordingly. Figure 3 shows a graph for a subset of 30 topics (nodes), where edge thickness encodes correlation coefficients and we have chosen, to ease visualization, to show only coefficients larger than 0.85. In addition, Table 2 shows the top words for each topic depicted in Figure 3. We see three very interesting subgraphs representing different categories, namely, sports, computers and politics/law. Complete tables of the most probable words in the learned topics, and graphs for the three corpora considered are presented in the Supplementary Material.

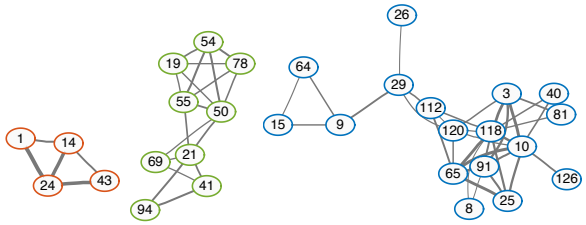

Figure 3. Graphs induced by the correlation structure learned by DPFA-SBN for the *20 Newsgroups*. Each node represents a topic with top words shown in Table 2.

6. Conclusion

We have presented the Deep Poisson Factor Analysis model, an extension of PFA, that models the high-order interactions between topics, via a deep binary hierarchical structure, employing SBNs and RBMs. To address large-scale datasets, two stochastic Bayesian learning algorithms were developed. Experimental results on several corpora show that the proposed approach obtains superior test perplexities and reveals interesting topic structures.

While this work has focused on unsupervised topic modeling, one can extend the model into a supervised version by joint modeling of the text with associated labels via latent binary features as in Zhang & Carin (2012). Furthermore, as mentioned in Section 5, *global-local* shrinkage priors (Polson & Scott, 2012) will encourage a large proportion of the elements of $\mathbf{W}^{(\ell)}$ to be shrunk close to zero. By setting the number of hidden units to a reasonably large value, this provides a natural way to let the model select automatically the number of features actually needed.

Acknowledgements

This research was supported in part by ARO, DARPA, DOE, NGA and ONR.

References

Blei, D. M. and Lafferty, J. D. A correlated topic model of science. *The Annals of Applied Statistics*, 2007.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *JMLR*, 2003.

Blei, D. M., Griffiths, T., Jordan, M. I., and Tenenbaum, J. B. Hierarchical topic models and the nested Chinese restaurant process. *NIPS*, 2004.

Byrne, S. and Girolami, M. Geodesic Monte Carlo on embedded manifolds. *Scandinavian J. Statist*, 2013.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian Monte Carlo. *ICML*, 2014.

Danilo, J. R., Shakir, M., and Daan, W. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. Bayesian sampling using stochastic gradient thermostats. *NIPS*, 2014.

Gan, Z., Henao, R., Carlson, D., and Carin, L. Learning deep sigmoid belief networks with data augmentation. *AISTATS*, 2015.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B*, 2011.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. Bayesian conditional density filtering. *arXiv:1401.3632*, 2014.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.

Hinton, G. E. and Salakhutdinov, R. Discovering binary codes for documents by learning deep generative models. *Topics in Cognitive Science*, 2011.

Hinton, G. E., Osindero, S., and Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural computation*, 2006.

Ho, Q., Cipar, J., Cui, H., Kim, J. K., Lee, S., Gibbons, P. B., Gibbons, G. A., Ganger, G. R., and Xing, E. P. More effective distributed ml via a stale synchronous parallel parameter server. *NIPS*, 2013.

Hoffman, M. D., Blei, D. M., and Bach, F. Online learning for latent Dirichlet allocation. *NIPS*, 2010.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *JMLR*, 2013.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *ICLR*, 2014.

Larochelle, H. and Lauly, S. A neural autoregressive topic model. *NIPS*, 2012.

Li, M., Andersen, D., Smola, A., and Yu, K. Communication efficient distributed machine learning with the parameter server. *NIPS*, 2014.

Maaloe, L., Arngren, M., and Winther, O. Deep belief nets for topic modeling. *arXiv:1501.04325*, 2015.

Mimno, D., Hoffman, M. D., and Blei, D. M. Sparse stochastic inference for latent Dirichlet allocation. *ICML*, 2012.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. *ICML*, 2014.

Neal, R. M. Connectionist learning of belief networks. *Artificial Intelligence*, 1992.

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. Nested hierarchical Dirichlet processes. *PAMI*, 2015.

Patterson, S. and Teh, Y. W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. *NIPS*, 2013.

Polson, N. G. and Scott, J. G. Local shrinkage rules, Lévy processes and regularized regression. *J. R. Statist. Soc. B*, 2012.

Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using Pólya-Gamma latent variables. *JASA*, 2013.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. Deep exponential families. *AISTATS*, 2015.

Salakhutdinov, R. and Hinton, G. E. Replicated softmax: an undirected topic model. *NIPS*, 2009a.

Salakhutdinov, R. and Hinton, G. E. Deep Boltzmann machines. *AISTATS*, 2009b.

Srivastava, N., Salakhutdinov, R., and Hinton, G. E. Modeling documents with deep Boltzmann machines. *UAI*, 2013.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *JASA*, 2006.

Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

Wang, C. and Blei, D. M. Truncation-free stochastic variational inference for Bayesian nonparametric models. *NIPS*, 2012.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. *ICML*, 2011.

Williamson, S., Wang, C., Heller, K., and Blei, D. M. The IBP compound Dirichlet process and its application to focused topic modeling. *ICML*, 2010.

Yuille, A. The convergence of contrastive divergences. *NIPS*, 2005.

Zhang, X. and Carin, L. Joint modeling of a matrix with associated text via latent binary features. *NIPS*, 2012.

Zhou, M. and Carin, L. Negative binomial process count and mixture modeling. *PAMI*, 2015.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. Beta-negative binomial process and Poisson factor analysis. *AISTATS*, 2012.