Scalable Deep Poisson Factor Analysis for Topic Modeling

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

Duke University, Durham NC 27708, USA

INTRODUCTION

Problem of interest: Developing deep generative models for documents that are represented in bag-of-words form.

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 serves as a flexible prior for revealing topic structure.

Challenge: Designing scalable Bayesian inference algorithms.

Solutions: stochastic algorithms.

- Applying Bayesian conditional density filtering algorithm.
- Extending recently proposed work on stochastic gradient thermostats.

MODEL FORMULATION

Poisson Factor Analysis: We represent a discrete matrix $\mathbf{X} \in \mathbb{Z}_+^{P \times N}$ containing counts from N documents and P words as

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

- $\mathbf{\Phi}$: factor loadings; $\mathbf{\Theta}$: factor scores; $\mathbf{H}^{(1)}$: latent binary features.
- We construct PFAs by placing Dirichlet priors on $\boldsymbol{\phi}_k$ (one column of $\mathbf{\Phi}$), and gamma priors on $\boldsymbol{\theta}_n$ (one column of $\mathbf{\Theta}$), *i.e.*, $\boldsymbol{\phi}_k \sim \text{Dir}(a_\phi, \dots, a_\phi)$, $\theta_{kn} \sim \text{Ga}(r_k, p_n/(1-p_n))$, $r_k \sim \text{Ga}(\gamma_0, 1/c_0)$, and $\gamma_0 \sim \text{Ga}(e_0, 1/f_0)$.
- The **novelty** in our model comes from the prior for $\boldsymbol{h}_n^{(1)}$ (one column of $\mathbf{H}^{(1)}$).

Structured Priors on the Latent Binary Matrix

Modeling with the SBN: Assume $\boldsymbol{h}_n^{(1)} \in \{0,1\}^{K_1}$, we define $\boldsymbol{h}_n^{(2)} \in \{0,1\}^{K_2}$ placed at a layer "above" $\boldsymbol{h}_n^{(1)}$. An SBN model has the generative process,

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

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

Modeling with the RBM: An RBM is a Markov random field with the same bipartite structure as the SBN. The energy function of an RBM is defined as

$$E(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)}) = -(\boldsymbol{h}_n^{(1)})^\top \boldsymbol{c}^{(1)} - (\boldsymbol{h}_n^{(1)})^\top \mathbf{W}^{(1)} \boldsymbol{h}_n^{(2)} - (\boldsymbol{h}_n^{(2)})^\top \boldsymbol{c}^{(2)}\,.$$

Deep Architecture: we can add multiple layers of SBNs or RBMs to obtain a deep architecture, $p(\boldsymbol{h}_n^{(1)}, \dots, \boldsymbol{h}_n^{(L)}) = p(\boldsymbol{h}_n^{(L)}) \Pi_{\ell=2}^{L} p(\boldsymbol{h}_n^{(\ell-1)} | \boldsymbol{h}_n^{(\ell)})$.

SCALABLE INFERENCE

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.

Graphical Model

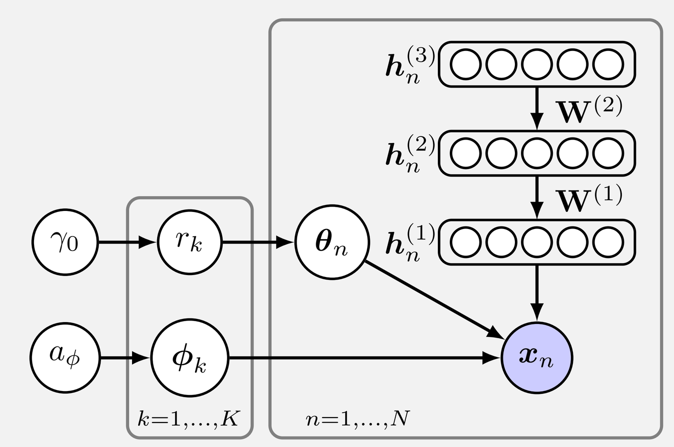

Figure: Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a *deep Boltzmann machine*.

Stochastic Gradient Nóse-Hoover Thermostats (SGNHT):

- Extending *Hamiltonian Monte Carlo* using stochastic gradient.
- Introducing thermostat to maintain system temperature.
- Adaptively absorbing stochastic gradient noise.

Algorithms: (*Left*) BCDF. (*Right*) SGNHT.

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 $\mathbf{\Psi}_g^{(t)} = \text{mean}(\mathbf{\Psi}_g^{1:S})$, $\mathbf{S}_g^{(t)} = \text{mean}(\mathbf{S}_g^{1:S})$.
end for

Input: text documents,*i.e.*,a count matrix $\mathbf{X}$.
Random Initialization.
for $t = 1$ **to** ∞ **do**
  $\mathbf{\Psi}_g^{(t+1)} = \mathbf{\Psi}_g^{(t)} + \boldsymbol{v}^{(t)} h$.
  $\boldsymbol{v}^{(t+1)} = \tilde{f}(\mathbf{\Psi}_g^{(t+1)})h - \mathbf{\Xi}^{(t)} \boldsymbol{v}^{(t)} h + \sqrt{2Dh}\mathcal{N}(0, \mathbf{I})$.
  $\mathbf{\Xi}^{(t+1)} = \mathbf{\Xi}^{(t)} + (\boldsymbol{q}^{(t+1)} - \mathbf{I})h$, where $\boldsymbol{q} = \text{diag}(v_1^2, \dots, v_M^2)$.
end for

Discussion:

- BCDF: ease of implementation, but prefers the conditional densities for all the parameters.
- SGNHT: more general and robust, fast convergence.

EXPERIMENTS

Datasets:

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

Quantitative Evaluation

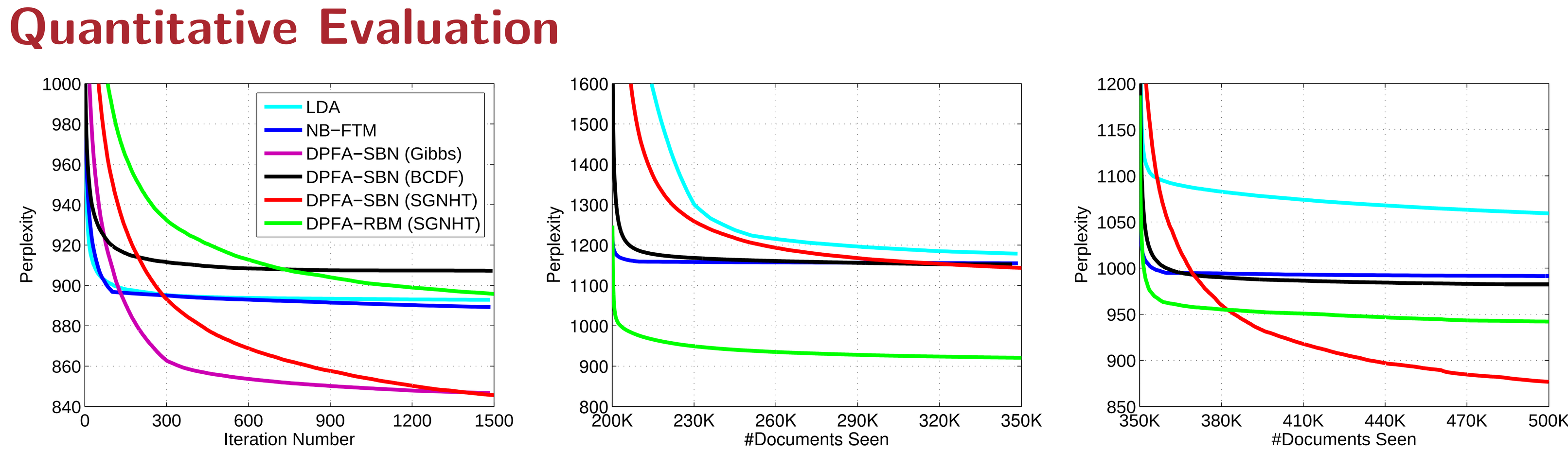

Figure: (Left) *20 Newsgroups*. (Middle) *RCV1-v2*. (Right) *Wikipedia*.

Table: 20 Newsgroups.

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
LDA	Gibbs	128	893
NB-FTM	Gibbs	128	887
RSM	CD5	128	877
nHDP	sVB	(10,10,5)$^\diamond$	889

Table: RCV1-v2 & Wikipedia.

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

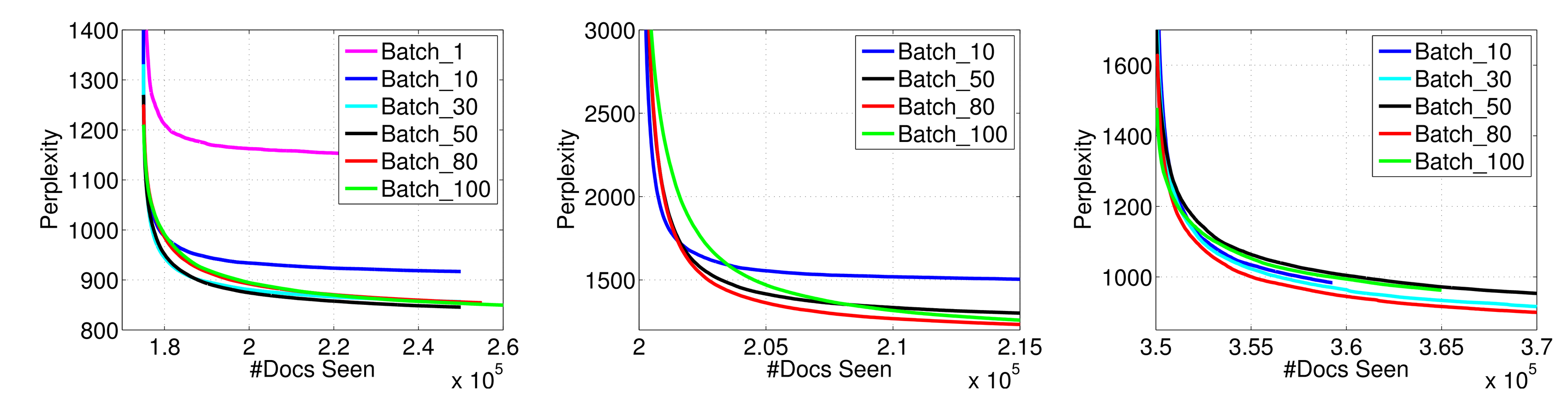

Figure: (Left) *20 Newsgroups*. (Middle) *RCV1-v2*. (Right) *Wikipedia*.

Topics learned from the *20 Newsgroup* :

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

Visualization: Sports, Computers, and Poltics/Law.

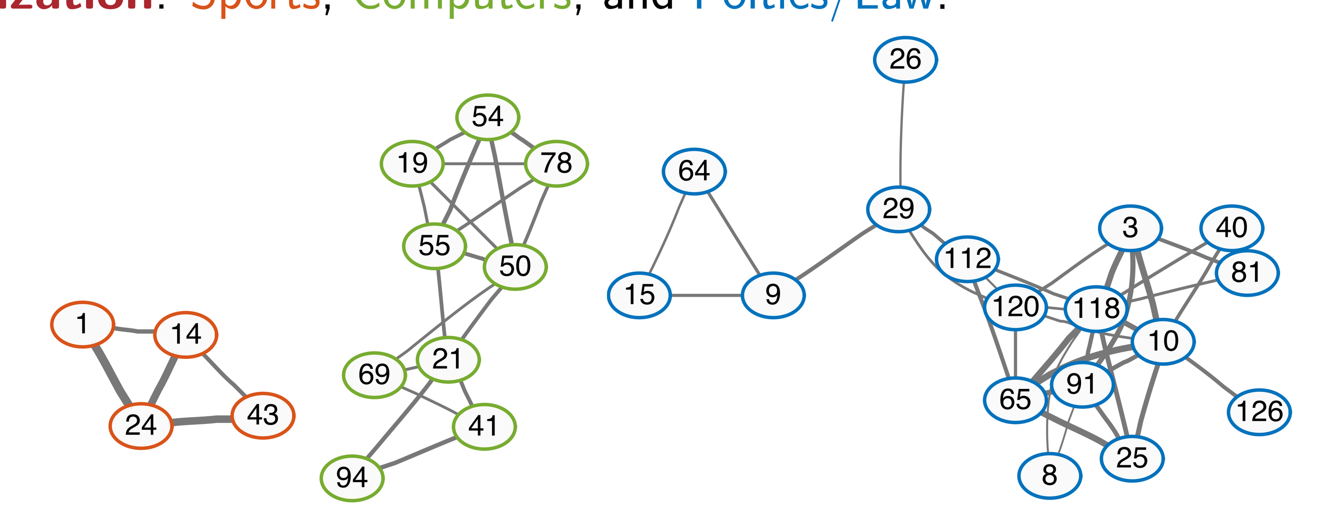

Figure: Graphs induced by the correlation structure learned by DPFA for the *20 Newsgroups*.

ACKNOWLEDGEMENTS

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