Scalable Deep Poisson Factor Analysis for Topic Modeling: Supplementary Material

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

A. Conditional Densities used in BCDF

Using *dot notation* to represent marginal sums, *e.g.*, $x_{\cdot nk} \triangleq \sum_p x_{pnk}$, we can write the conditional densities for DPFA as (Zhou & Carin, 2015)

$$x_{pnk}|- \sim \text{Multi}(x_{pn}; \zeta_{pn1}, \dots, \zeta_{pnK}) \,, \tag{1}$$

$$\boldsymbol{\phi}_k|- \sim \text{Dir}(a_\phi + x_{1\cdot k}, \dots, a_\phi + x_{P\cdot k}) \,,$$

$$\theta_{kn}|- \sim \text{Gamma}(r_k h_{kn}^{(1)} + x_{\cdot nk}, p_n) \,,$$

$$r_k|- \sim \text{Gamma}\left(\gamma_0 + \sum_{n=1}^N l_{kn}, \frac{1}{c_0 - \sum_{n=1}^N h_{kn}^{(1)} \ln(1-p_n)}\right) \,, \tag{2}$$

$$\gamma_0|- \sim \text{Gamma}\left(e_0 + \sum_{k=1}^K l_k^{'}, \frac{1}{f_0 - \sum_{k=1}^K \ln(1-p_k^{'})}\right) \,, \tag{3}$$

$$h_{kn}^{(1)}|- \sim \delta(x_{\cdot nk} = 0)\text{Ber}\left(\frac{\pi_{kn}(1-p_n)^{r_k}}{\pi_{kn}(1-p_n)^{r_k} + (1-\pi_{kn})}\right) + \delta(x_{\cdot nk} > 0) \,,$$

where

$$l_{kn}|- \sim \text{CRT}\left(x_{\cdot nk}, r_k h_{kn}^{(1)}\right) \,, \quad l_k^{\prime}|- \sim \text{ CRT}\left(\sum_{n=1}^N l_{kn}, \gamma_0\right) \,,$$

$$\zeta_{pnk} = \frac{\phi_{pk}\theta_{kn}}{\sum_{k=1}^K \phi_{pk}\theta_{kn}} \,, \quad p_k^{\prime} = \frac{-\sum_{n=1}^N h_{kn}^{(1)} \ln(1-p_n)}{c_0 - \sum_{n=1}^N h_{kn}^{(1)} \ln(1-p_n)} \,,$$

$$\pi_{kn} = \sigma\left((\boldsymbol{w}_k^{(1)})^\top \boldsymbol{h}_n^{(2)} + c_k^{(1)}\right) \,.$$

CRT represents the Chinese Restaurant Table distribution. A CRT random variable $l \sim \text{CRT}(m, r)$ can be generated with the summation of independent Bernoulli random variables as (Zhou & Carin, 2015)

$$l = \sum_{n=1}^m b_n, \quad b_n \sim \text{Ber}\left(\frac{r}{n-1+r}\right) \,.$$

B. Proof of Theorem 1

Consider a general stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{\Gamma} = Q(\boldsymbol{\Gamma})\mathrm{d}t + \sqrt{2D(\boldsymbol{\Gamma})}\mathrm{d}\mathcal{W} \,, \tag{4}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^N$, $Q : \mathbb{R}^N \to \mathbb{R}^N$, $D : \mathbb{R}^M \to \mathbb{R}^{N\times P}$ are measurable functions with P unnecessarily equal to N, and $\mathcal{W}$ is the standard P-dimensional Brownian motion. Taking

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{v} \\ \xi \end{pmatrix} \,, \; Q(\boldsymbol{\Gamma}) = \begin{pmatrix} \boldsymbol{v} \\ f - \xi\boldsymbol{v} \\ \frac{1}{M}\boldsymbol{v}^T\boldsymbol{v} - 1 \end{pmatrix} \,,$$

and $D(\boldsymbol{\Gamma})$ to be constant, *i.e.*, D, recovers the setting in the original *stochastic gradient Nóse-Hoover thermostats* algorithm (Ding et al., 2014) (with the notation of this paper). Furthermore, write the joint distribution of $\boldsymbol{\Gamma}$ as

$$p(\boldsymbol{\Gamma}) = \frac{1}{Z}\exp\left\{-H(\boldsymbol{\Gamma})\right\} \,,$$

where $H(\boldsymbol{\Gamma})$ is usually called the Hamiltonian function of a system. In the following we decompose $\boldsymbol{\Gamma}$ as $\boldsymbol{\Gamma} = (\boldsymbol{\theta}, \boldsymbol{x})$ and $H(\boldsymbol{\Gamma})$ as $H(\boldsymbol{\Gamma}) = U(\boldsymbol{\theta}) + E(\boldsymbol{\theta}, \boldsymbol{x})$ where $U(\boldsymbol{\theta})$ may be the negative log-posterior of a Bayesian model.

To prove Theorem 1 in the main text, we make use of Lemma 1 below, which is essentially the main theorem in (Ding et al., 2014), which again is a consequence of the celebrated Fokker-Planck Equation (Risken, 1989).

Lemma 1. *The stochastic process of $\boldsymbol{\theta}$ generated by the stochastic differential equation* (4) *has the target distribution $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{Z}\exp\{-U(\boldsymbol{\theta})\}$ as its stationary distribution, if $p(\boldsymbol{\Gamma})$ satisfies the following marginalization condition:*

$$\exp\{-U(\boldsymbol{\theta})\} \propto \int \exp\{-U(\boldsymbol{\theta}) - E(\boldsymbol{\theta}, \boldsymbol{x})\}\mathrm{d}\boldsymbol{x} \,, \tag{5}$$

and if the following condition is also satisfied:

$$\nabla \cdot (pQ) = \nabla\nabla^\top : (pD)\,, \tag{6}$$

where $\nabla \triangleq (\partial/\partial\boldsymbol{\theta}, \partial/\partial\boldsymbol{x})$, *"$\cdot$" represents the vector inner product operator, ":" represents a matrix double dot product, i.e.,* $X : Y \triangleq$ *tr*$(X^\top Y)$.

Proof of Theorem 1. For conciseness, we replace the model parameters $\boldsymbol{\Psi}_g$ in the main text with notation $\boldsymbol{\theta}$ in the following. We first reformulate our proposed SGNHT as a special case of the general SDE in (4), *i.e.*,

$$\begin{aligned}\boldsymbol{\Gamma} &= (\boldsymbol{\theta}^\top, \boldsymbol{v}^\top, \xi_1, \cdots, \xi_M)^\top\,,\\ Q(\boldsymbol{\Gamma}) &= (\boldsymbol{v}^\top, (f - \Xi\boldsymbol{v})^\top, v_1^2 - 1, \cdots, v_M^2 - 1)^\top\,,\\ D(\boldsymbol{\Gamma}) &= \mathrm{diag}\left(\underbrace{0, \cdots, 0}_{M}, \underbrace{D, \cdots, D}_{M}, \underbrace{0, \cdots, 0}_{M}\right)\,.\end{aligned}$$

From Theorem 1 in the main text we know

$$\begin{aligned}p(\boldsymbol{\Gamma}) = &\frac{1}{Z}\exp\left(-\frac{1}{2}\boldsymbol{v}^\top\boldsymbol{v} - U(\boldsymbol{\theta})\right.\\ &\left.- \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{\Xi} - D)^\top(\boldsymbol{\Xi} - D)\right\}\right)\,,\end{aligned} \tag{7}$$

with $H(\boldsymbol{\Gamma}) = \frac{1}{2}\boldsymbol{v}^\top\boldsymbol{v} + U(\boldsymbol{\theta}) + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{\Xi} - D)^\top(\boldsymbol{\Xi} - D)\right\}$. The marginalization condition (5) is trivially satisfied, we are left to verify condition (6). Substituting p and Q into (6), we have the left-hand side

$$\begin{aligned}\mathrm{LHS} &= \sum_i \frac{\partial}{\partial\Gamma_i}(pQ_i)\\ &= \sum_i \frac{\partial p}{\partial\Gamma_i}Q_i + \frac{\partial Q_i}{\partial\Gamma_i}p\\ &= \sum_i\left(\frac{\partial Q_i}{\partial\Gamma_i} - \frac{\partial H}{\partial\Gamma_i}\right)p\\ &= -\sum_i \xi_i - \boldsymbol{v}^T(f - \Xi\boldsymbol{v}) + f^T\boldsymbol{v} - \sum_i(\xi_i - D)(v_i^2 - 1)\\ &= \sum_i D_{ii}\left(v_i^2 - 1\right)p\,.\end{aligned}$$

Since D is independent of $\boldsymbol{\Gamma}$, we have the right-hand side

$$\begin{aligned}\mathrm{RHS} &= \sum_i\sum_j D_{ij}\frac{\partial^2}{\partial\Gamma_i\partial\Gamma_j}p\\ &= \sum_i\sum_j D_{ij}\frac{\partial}{\partial v_j}\left(-\frac{\partial H}{\partial v_i}p\right)\\ &= \sum_i D_{ii}\left(v_i^2 - 1\right)p\\ &\equiv \mathrm{LHS}\,.\end{aligned}$$

Now we have verified both conditions of Lemma 1, this makes the distribution (7) the equilibrium distribution of our proposed SGNHT algorithm. □

C. Sampling Topic-Word Distributions using the Stochastic Gradient Riemannian Langevin Dynamics

The Langevin dynamic (diffusion) is defined via a stochastic differential equation of the following form:

$$\mathrm{d}\boldsymbol{\theta}_t = \frac{1}{2}\nabla_{\boldsymbol{\theta}_t}\log U(\boldsymbol{\theta}_t)\mathrm{d}t + \mathrm{d}\mathcal{W}_t\,, \tag{8}$$

where t is the time index, $\boldsymbol{\theta}_t \in \mathbb{R}^M$ is the model parameter, $U(\boldsymbol{\theta}_t) \triangleq \left(\prod_{i=1}^N p(x_i|\boldsymbol{\theta}_t)\right)p(\boldsymbol{\theta}_t)$ is the model posterior, and $\mathcal{W}_t$ is the standard M-dimensional Brownian motion. The law of the Langevin dynamic is described by the Fokker-Planck equation (Risken, 1989), and the equilibrium distribution $p(\boldsymbol{\theta})$ of $\boldsymbol{\theta}_t$ can be shown to be the model posterior $U(\boldsymbol{\theta})$ (Øksendal, 2003). In Bayesian learning with big data, the stochastic gradient Langevin dynamic (SGLD) (Welling & Teh, 2011) generalizes the Langevin dynamic (8) by replacing $U(\boldsymbol{\theta}_t)$ with a stochastic version $\tilde{U}(\boldsymbol{\theta}_t)$ evaluated on a subset of data, *e.g.*, $\tilde{U}(\boldsymbol{\theta}_t) \triangleq \left(\prod_{i\in\mathcal{D}} p(x_i|\boldsymbol{\theta}_t)\right)p(\boldsymbol{\theta}_t)$, where $\mathcal{D}$ is a random subset of $\{1, \cdots, N\}$. Furthermore, the corresponding SDE of SGLD is then solved via the Euler-Maruyama scheme (Tuckerman, 2010) using a decreasing step sizes sequence (Welling & Teh, 2011), *i.e.*, samples of the SGLD are generated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \frac{\delta_t}{2}\nabla_{\boldsymbol{\theta}}\log\tilde{U}(\boldsymbol{\theta}_t) + \zeta_t, \quad \zeta_t \sim N(0, \delta_t\mathbf{I})\,,$$

where $\mathbf{I}$ is the identity matrix and $\{\delta_t\}$ is a decreasing sequence such that $\lim_{t\to\infty}\delta_t = 0$ and $\sum_{t=1}^\infty \delta_t = \infty$. Under certain assumptions, it is shown that the SGLD is consistent with the Langevin dynamics of (8), *i.e.*, it generates correct samples from the posterior $U(\boldsymbol{\theta})$ (Teh et al., 2014; Vollmer et al., 2015).

The stochastic gradient Riemannian Langevin dynamics (SGRLD) (Patterson & Teh, 2013) extends the SGLD by defining it on Riemannian manifolds (Girolami & Calderhead, 2011; Byrne & Girolami, 2013). Specifically, given a Riemannian metric $G(\theta)$ (Jürgen, 2008), the SGRLD generalizes the Langevin dynamic (8) on a Riemannian manifold and generates samples using the following proposal

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \frac{\delta_t}{2}\mu(\boldsymbol{\theta}_t) + G(\boldsymbol{\theta}_t)^{-1/2}\zeta_t, \quad \zeta_t \sim N(0, \delta_t\mathbf{I})\,, \tag{9}$$

where the j-th component of $\mu(\boldsymbol{\theta}_t)$ is given by $\mu(\boldsymbol{\theta}_t)_j =$

$$\begin{aligned}&\left(G^{-1}(\boldsymbol{\theta}_t)\ \nabla_{\boldsymbol{\theta}_t}\tilde{U}(\boldsymbol{\theta}_t)\right)_j - 2\sum_{k=1}^M\left(G^{-1}(\boldsymbol{\theta}_t)\frac{\partial G(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_{tk}}G^{-1}(\boldsymbol{\theta}_t)\right)_{jk}\\ &\quad + \sum_{k=1}^M\left(G^{-1}(\boldsymbol{\theta}_t)\right)_{jk}\mathrm{tr}\left(G^{-1}(\boldsymbol{\theta}_t)\frac{\partial G(\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_{tk}}\right)\,.\end{aligned}$$

The SGRLD makes moves along a Riemannian manifold defined by the metric $G(\boldsymbol{\theta})$, thus having the advantage of converging faster towards the optimal solution compared to the naive SGLD. In the Bayesian setting, the Fisher Information matrix (Rao, 1945; Amari, 1990; 1997) is usually chosen as the metric $G(\boldsymbol{\theta})$, which has convenient forms for some models.

SGRLD for Deep Poisson Factor Analysis We use the SGRLD algorithm to sample the topic-word distributions $\{\boldsymbol{\phi}_k\}$'s. From Section A, we see that based on the augmentation x_{pnk}, the posterior of $\boldsymbol{\phi}_k$ is a Dirichlet distribution, *i.e.*, the joint conditional likelihood of $(\{x_{pnk}\}, \boldsymbol{\phi}_k)$ is $p(\{x_{pnk}\}, \boldsymbol{\phi}_k|\cdot) \propto \prod_p (\phi_{pk})^{x_{p \cdot k}}$.

Now we reparameterize $\boldsymbol{\phi}_k$ using the *expanded-mean* schema (Patterson & Teh, 2013) as $\phi_{pk} = \frac{\tilde{\phi}_{pk}}{\sum_{p'} \tilde{\phi}_{pk'}}$. Together with the prior on $\boldsymbol{\phi}_k$ as $\boldsymbol{\phi}_k \sim \text{Dir}(a_\phi)$, this gives us a stochastic gradient on a subset of data $\mathcal{D}$ as

$$\frac{\partial \log p(\tilde{\boldsymbol{\phi}}_k|\cdot)}{\partial \tilde{\phi}_{pk}} \approx \frac{a_\phi - 1}{\tilde{\phi}_{pk}} - 1 + \frac{N}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} \mathbb{E}_{\{x_{pnk}\}} \left[\frac{x_{pnk}}{\tilde{\phi}_{pk}} - \frac{x_{\cdot nk}}{\tilde{\phi}_{k\cdot}} \right] ,$$

where the expectation is taken over the posterior of $\{x_{pnk}\}$, which is infeasible. As a result, we use samples from the posterior to approximate the expectation. To this end, we use the conditional posterior of $\{x_{pnk}\}$ in (1) to collect samples for the current mini-batch $\mathcal{D}$ after a few burn-in steps.

To apply the SGRLD algorithm, we need to choose a metric $G(\boldsymbol{\theta})$ for the Riemannian manifold. Because the posterior of $\boldsymbol{\phi}_k$ is Dirichlet, we thus use the same Riemannian metric as in LDA (Patterson & Teh, 2013), *e.g.*, $G(\phi) = \text{diag}(\phi_{11}, \cdots, \phi_{P1}, \cdots, \phi_{1K}, \cdots, \phi_{PK})^{-1}$, and use the same *mirror idea* (Patterson & Teh, 2013) to avoid parameters going out of boundary. After plugging $G(\theta)$ into (9) and simplifying, we get the update for $\tilde{\phi}_{pk}$ as:

$$\tilde{\phi}_{pk} = \Bigg| \tilde{\phi}_{pk} + \frac{\delta}{2} \Bigg(a_\phi - \tilde{\phi}_{pk} + \frac{N}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbb{E}_{\{x_{pnk}\}} \left[x_{pnk} - x_{\cdot nk} \phi_{pk} \right] \Bigg) + (\tilde{\phi}_{pk})^{1/2} \zeta \Bigg| ,$$

where we have omitted the *time index* t in the parameters for simplicity.

For the other global model parameters such as r_k and γ_0, we make use of their posterior distributions in (2) and (3), respectively. We only briefly describe the sampling for r_k, sampling γ_0 is similar. To simplified, rewrite the posterior r_k as

$$p(r_k|\cdot) \sim \text{Gamma}\left(a_k + c_0, 1/(c_0 + b_k)\right) ,$$

where a_k and b_k are parameters for the Gamma distribution containing local parameters (or augmented parameters) from the current mini-batch, *e.g.*, a_k contains $\{l_{kn}\}$ for the current mini-batch. Denote these local parameters as $\boldsymbol{L}_{\boldsymbol{x}}$. Now we use the reparameterization for r_k as $r_k = e^{\tilde{r}_k}$ with $\tilde{r}_k \sim$ Log-Gamma$(\gamma_0, 1/c_0)$, this is equivalent as putting a Gamma$(\gamma_0, 1/c_0)$ prior on r_k. Now the stochastic gradient for $\tilde{r}_k$ can be easily seen as

$$\begin{aligned}\frac{\partial \log p(\tilde{r}_k|\cdot)}{\partial \tilde{r}_k} &= \frac{\partial}{\partial \tilde{r}_k} \mathbb{E}_{\boldsymbol{L}_{\boldsymbol{x}}} \left[(a_k + \gamma_0 - 1)\tilde{r}_k - e^{-(c_0 + b_k)\tilde{r}_k} \right] \\ &= \mathbb{E}_{\boldsymbol{L}_{\boldsymbol{x}}} \left[a_k + \gamma_0 - 1 - (c_0 + b_k) r_k^{c_0 + b_k} \right] .\end{aligned}$$

Again, the expectation can be approximated using Monte Carlo integration by drawing samples from the posterior, then the stochastic gradient Langevin dynamic can be straightforwardly applied.

D. Evaluation Details on Perplexities

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 remaining 20% subset, denoted as $\mathbf{Y}$. For the PFA-based models, the test perplexity is calculated as (Zhou et al., 2012)

$$\exp\left(-\frac{1}{y_{\cdot\cdot}} \sum_{p=1}^{P} \sum_{n=1}^{N} y_{pn} \log \frac{\sum_{s=1}^{S} \sum_{k=1}^{K} \phi_{pk}^s \theta_{kn}^s}{\sum_{s=1}^{S} \sum_{p=1}^{P} \sum_{k=1}^{K} \phi_{pk}^s \theta_{kn}^s} \right) ,$$

where S is the total number of collected samples, $y_{\cdot\cdot} = \sum_{p=1}^{P} \sum_{n=1}^{N} y_{pn}$ and y_{pn} is an element of matrix $\mathbf{Y}$.

The conditional distribution of $\boldsymbol{y}_n$ given $\boldsymbol{h}_n$, in the Replicated Softmax model (RSM) is specified as

$$\begin{aligned}\boldsymbol{y}_n &\sim \text{Multi}(D_n; \boldsymbol{\beta}_n) , \\ \beta_{pn} &= \frac{\exp(\boldsymbol{w}_p^\top \boldsymbol{h}_n + c_p)}{\sum_{p'=1}^{P} \exp(\boldsymbol{w}_{p'}^\top \boldsymbol{h}_n + c_{p'})} ,\end{aligned}$$

where $\boldsymbol{y}_n$ is the nth column of $\mathbf{Y}$, and $D_n = \sum_{p=1}^{P} y_{pn}$. $\mathbf{W} = [\boldsymbol{w}_1, \ldots . \boldsymbol{w}_P]^\top \in \mathbb{R}^{P \times K}$ is the mapping from $\boldsymbol{h}_n$ to $\boldsymbol{y}_n$, and $\boldsymbol{c} = [c_1, \ldots . c_P]^\top \in \mathbb{R}^{P \times 1}$ is the bias term. Based on this, the predictive test perplexity for RSM can be calculated as

$$\exp\left(-\frac{1}{y_{\cdot\cdot}} \sum_{p=1}^{P} \sum_{n=1}^{N} y_{pn} \log \beta_{pn} \right) .$$

E. Sensitivity Analysis

We examined the sensitivity of the model performance with respect to batch sizes in SGNHT on the three corpora considered. The results are shown in Figure 1. We found that

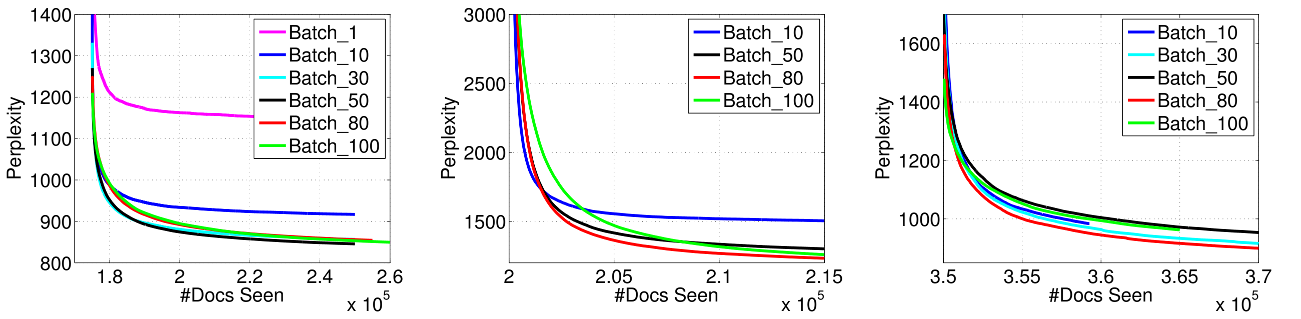

Figure 1. Test perplexities *w.r.t.* mini-batch sizes on the three corpora. The number of hidden units in each layer is 128, 64, 32, respectively. (Left) *20 Newsgroups*. (Middle) *RCV1-v2*. (Right) *Wikipedia*.

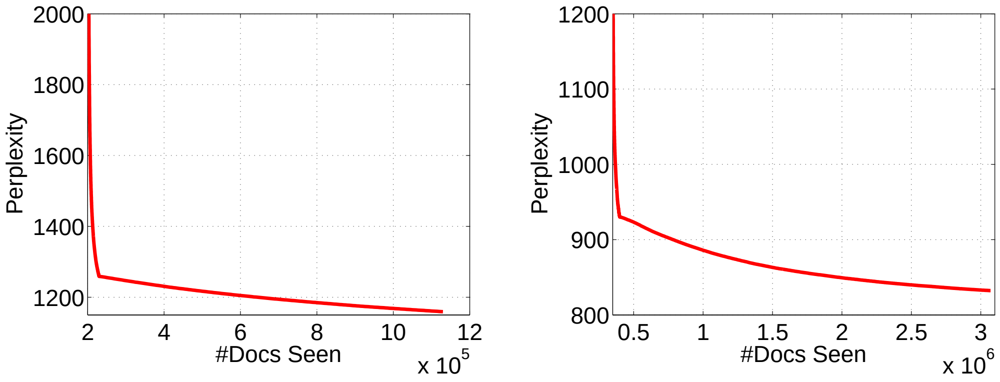

Figure 2. Test perplexities as a function of training documents seen. The number of hidden units in each layer is 128, 64, 32, respectively. (Left) *RCV1-v2*. (Right) *Wikipedia*.

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.

We run the SGNHT algorithm on the *RCV1-v2* and *Wikipedia* datasets long enough so that the whole corpora can be traversed. The results are shown in Figure 2. As can be seen, performance smoothly improves as the amount of data processed increases.

F. Additional Results

We compare our DPFA model with stronger shallow baselines, such as the Marked-Gamma-NB and Marked-Beta-NB model described in Zhou & Carin (2015). Direct Gibbs sampling is only suitable for relatively small datasets, so only results on the *20Newsgroups* are reported, shown in Table 1. As can be seen, these advanced models can achieve perplexity results comparable to those of a two-layer DPFA model.

Table 1. Additional results on *20 Newsgroups*.

MODEL	METHOD	DIM	PERP.
MARKED-BETA-NB	GIBBS	128	853
MARKED-GAMMA-NB	GIBBS	128	854

G. Source Code

The source code, along with the topics we inferred from the model, are available at `https://github.com/zhegan27/dpfa_icml2015`. This package is made publicly available for reproducibility purposes, and it is not optimized for speed, minimally documented but fully functional.

References

Amari, S. *Differential geometrical methods in statistics*. Lecture Notes in Statistics 28, Springer-Verlag, 1990.

Amari, S. Natural gradient works efficiently in learning. *Neural computation*, 1997.

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

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and

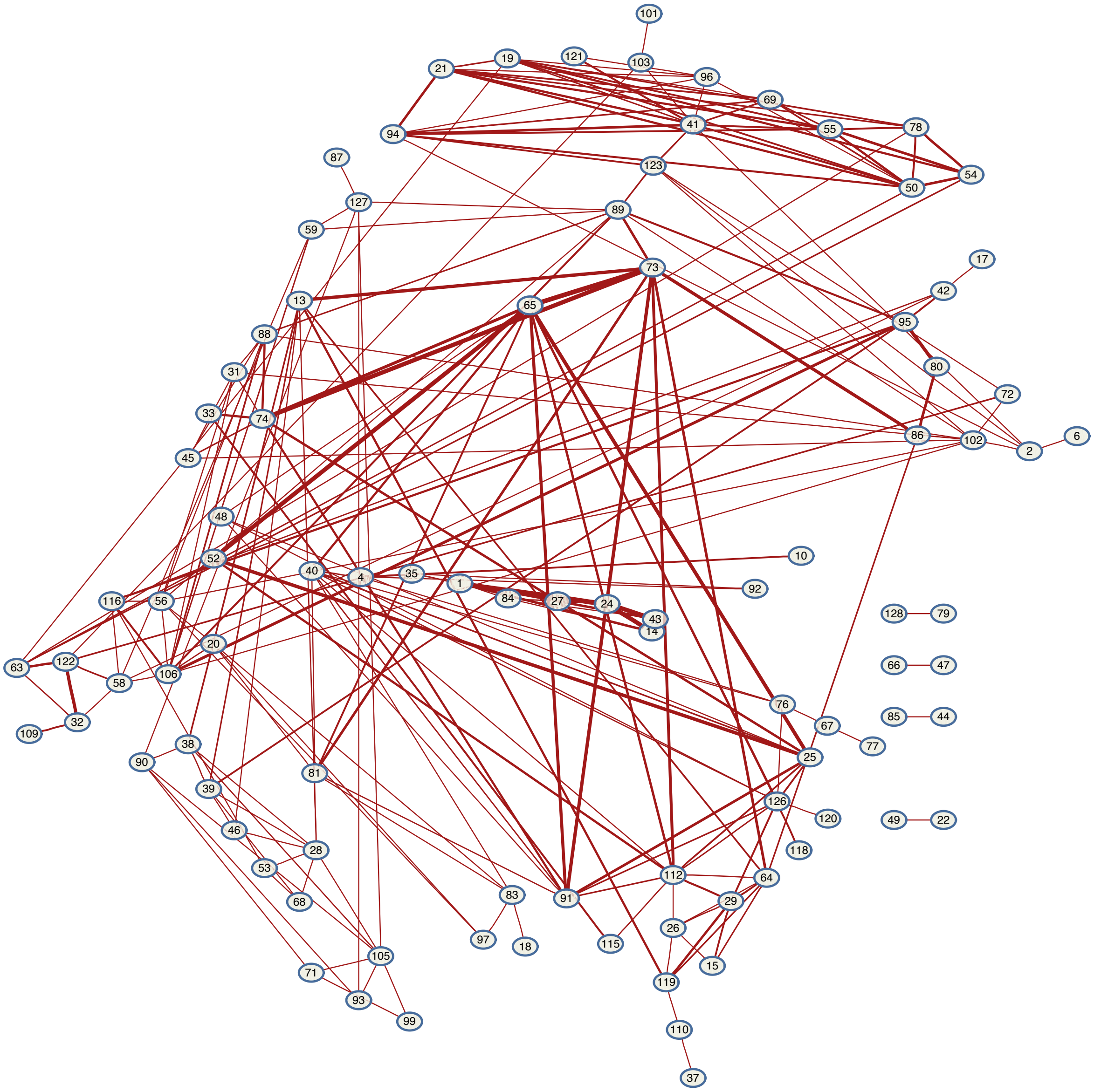

Figure 3. Full graph induced by the correlation structure learned by DPFA-SBN for the *20 Newsgroups* corpus.

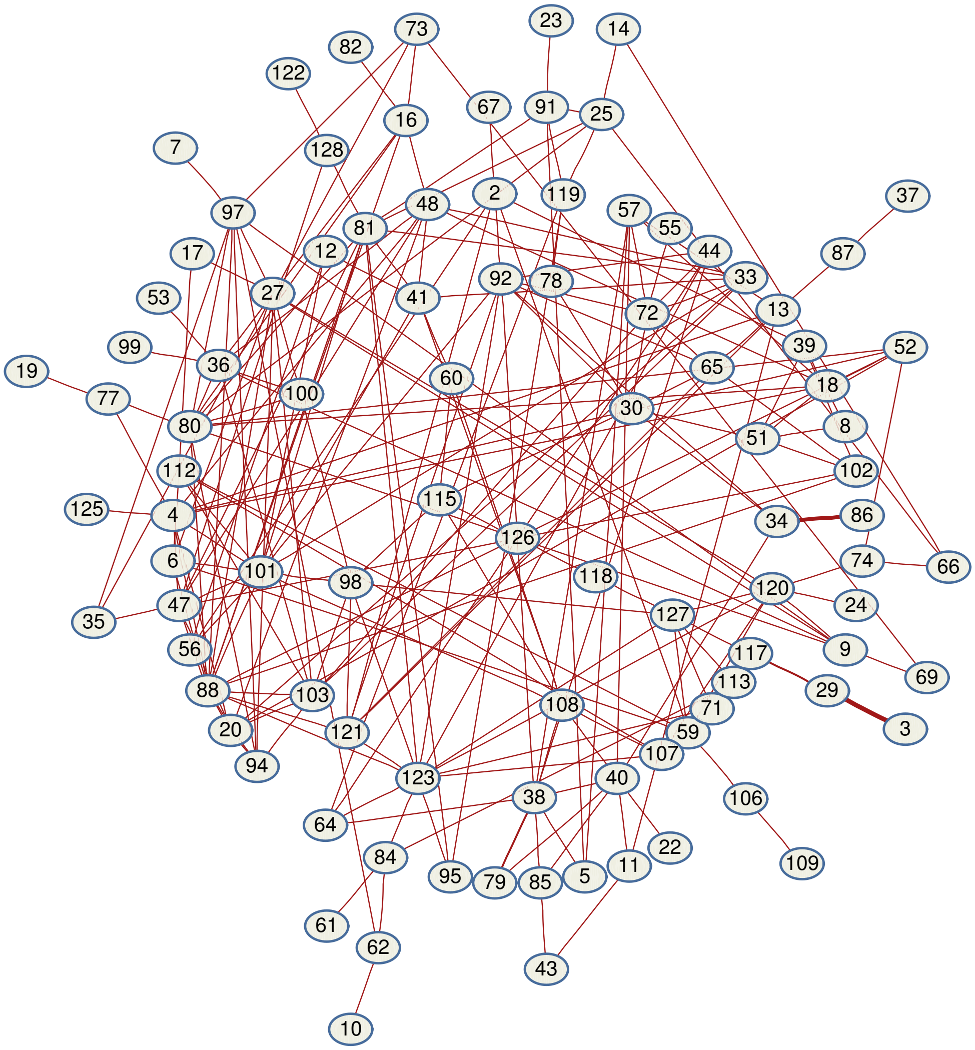

Figure 4. Full graph induced by the correlation structure learned by DPFA-SBN for the *RCV1-v2* corpus.

Neven, H. Bayesian sampling using stochastic gradient thermostats. *NIPS*, 2014.

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

Jürgen, J. *Riemannian Geometry and Geometric Analysis*. Springer-Verlag, 2008.

Øksendal, B. *Stochastic differential equations*. Springer, 2003.

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

Rao, C. R. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta mathematical society*, 1945.

Risken, H. *The Fokker-Planck Equation*. Springer-Verlag, New York, 1989.

Teh, Y. W., Thiery, A. H., and Vollmer, S. J. Consistency and fluctuations for stochastic gradient Langevin dynamics. *arXiv:1409.0578*, 2014.

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

Vollmer, S. J., Zygalakis, K. C., and Teh, Y. W. (Non-) asymptotic properties of stochastic gradient Langevin dynamics. *arXiv:1501.00438*, 2015.

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

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.