A. Conditional Densities used in BCDF

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

\[
\begin{align*}
  x_{p nk} & \sim \text{Multi}(x_{pn}; \zeta_{p n 1}, \ldots, \zeta_{p n K}), \\
  \phi_k & \sim \text{Dir}(a_0 + x_1, \ldots, a_0 + x_P), \\
  \theta_{kn} & \sim \text{Gamma}(r_k h_{kn}^{(1)} + x_{nk}; p_n), \\
  \gamma_0 & \sim \text{Gamma}(c_0 + \sum_{k=1}^K f_0 - \sum_{k=1}^K \ln(1 - p_k)), \\
  h_{kn}^{(1)} & \sim \delta(x_{nk} = 0) \text{Ber}\left(\frac{\pi_{kn}(1-p_n)^{\gamma_n}}{\pi_{kn}(1-p_n)\gamma_n + (1-\pi_{kn})}\right) + \delta(x_{nk} > 0), \\
\end{align*}
\]

where

\[
\begin{align*}
  l_{kn} & \sim \text{CRT} \left( x_{nk}, r_k h_{kn}^{(1)} \right), \\
  l'_k & \sim \text{CRT} \left( \sum_{n=1}^N l_{kn}, \gamma_0 \right), \\
  \zeta_{p nk} & = \frac{\phi_{pk}\theta_{kn}}{\sum_{k=1}^K \phi_{pk}\theta_{kn}}, \\
  p'_k & = -\sum_{n=1}^N h_{kn}^{(1)} \ln(1 - p_n) / c_0 - \sum_{n=1}^N h_{kn}^{(1)} \ln(1 - p_n), \\
  \pi_{kn} & = \sigma \left( \left( w_k^{(1)} \right)^T h_k^{(2)} + c_k^{(1)} \right). \\
\end{align*}
\]

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

\[
d\Gamma = Q(\Gamma)dt + \sqrt{2D(\Gamma)}dW, \quad (4)
\]

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

\[
\Gamma = \begin{pmatrix} \theta \\ v \\ \xi \end{pmatrix}, \quad Q(\Gamma) = \begin{pmatrix} f - \xi v \\ \frac{1}{M} v^T v - 1 \end{pmatrix},
\]

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

\[
p(\Gamma) = \frac{1}{Z} \exp \{ -H(\Gamma) \},
\]

where \( H(\Gamma) \) is usually called the Hamiltonian function of a system. In the following we decompose \( \Gamma \) as \( \Gamma = (\theta, x) \) and \( H(\Gamma) = U(\theta) + E(\theta, x) \) where \( U(\theta) \) 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 \( \theta \) generated by the stochastic differential equation (4) has the target distribution \( p_0(\theta) = \frac{1}{Z} \exp \{ -U(\theta) \} \) as its stationary distribution, if \( p(\Gamma) \) satisfies the following marginalization condition:

\[
\exp \{ -U(\theta) \} \propto \int \exp \{ -U(\theta) - E(\theta, x) \} dx, \quad (5)
\]

\[1\text{DUKE].\]
and if the following condition is also satisfied:
\[ \nabla \cdot (pQ) = \nabla \nabla^T : (pD), \]
where \( \nabla \) represents the vector inner product operator; ";" represents a matrix double dot product, i.e., \( X : Y \triangleq \text{tr}(X^TY) \).

Proof of Theorem 1. For conciseness, we replace the model parameters \( \Psi^0 \) in the main text with notation \( \theta \) in the following. We first reformulate our proposed SGNHT model parameters \( \Psi \) in the main text with notation \( \theta \).

\[ \Gamma = (\theta^T, v^T, \xi_1, \cdots, \xi_M)^T, \]
\[ Q(\Gamma) = (v^T, (f - \Xi v)^T, v_i^2 - 1, \cdots, v_M^2 - 1)^T, \]
\[ D(\Gamma) = \text{diag} \left( 0, \cdots, 0, D, \cdots, D, 0, \cdots, 0 \right). \]

From Theorem 1 in the main text we know
\[ p(\Gamma) = \frac{1}{Z} \exp \left( -\frac{1}{2} v^T v - U(\theta) \right) - \frac{1}{2} \text{tr} \left( \left( \Xi - D \right)^T \left( \Xi - D \right) \right) \]

with \( H(\Gamma) = \frac{1}{2} v^T v + U(\theta) \). 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
\[ \text{LHS} = \sum_i \frac{\partial}{\partial \theta_i} (pQ_i) \\
= \sum_i \frac{\partial p}{\partial \theta_i} Q_i + \frac{\partial Q_i}{\partial \theta_i} p \\
= \sum_i \left( \frac{\partial Q_i}{\partial \theta_i} - \frac{\partial H}{\partial \theta_i} \right) p \\
= - \sum_i \xi_i - v^T (f - \Xi v) + f^T v - \sum_i (\xi_i - D) (v_i^2 - 1) \\
= \sum_i D_{ii} (v_i^2 - 1) p. \]

Since \( D \) is independent of \( \Gamma \), we have the right-hand side
\[ \text{RHS} = \sum_i \sum_j D_{ij} \frac{\partial^2}{\partial \theta_i \partial \theta_j} p \\
= \sum_i \sum_j D_{ij} \frac{\partial}{\partial \theta_j} \left( - \frac{\partial H}{\partial \theta_i} p \right) \\
= \sum_i D_{ii} (v_i^2 - 1) p \\
= \text{LHS}. \]

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:
\[ d\theta_t = \frac{1}{2} \nabla \theta_t \log U(\theta_t) dt + dW_t \]
where \( t \) is the time index, \( \theta_t \in \mathbb{R}^M \) is the model parameter, \( U(\theta_t) \triangleq \left( \prod_{i=1}^N p(x_i | \theta_t) \right) p(\theta_t) \) is the model posterior, and \( W_t \) 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 \( p(\theta) \) of \( \theta_t \) can be shown to be the model posterior \( U(\theta) \) (Oksendal, 2003).

In Bayesian learning with big data, the stochastic gradient Langevin dynamic (SGLD) (Welling & Teh, 2011) generalizes the Langevin dynamic (8) by replacing \( U(\theta_t) \) with a stochastic version \( \hat{U}(\theta_t) \) evaluated on a subset of data, e.g., \( \hat{U}(\theta_t) \triangleq \left( \prod_{i \in D(t)} p(x_i | \theta_t) \right) p(\theta_t) \), where \( D \) 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
\[ \theta_{t+1} = \theta_t + \frac{\delta_t}{2} \nabla \theta_t \log U(\theta_t) + \zeta_t, \quad \zeta_t \sim N(0, \delta_t I), \]
where \( I \) 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 \( U(\theta) \) (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
\[ \theta_{t+1} = \theta_t + \frac{\delta_t}{2} \mu(\theta_t) + G(\theta_t)^{-1/2} \zeta_t, \quad \zeta_t \sim N(0, \delta_t I), \]
where the \( j \)-th component of \( \mu(\theta_t) \) is given by \( \mu(\theta_t)_j = (G^{-1}(\theta_t) \nabla_{\theta_j} \hat{U}(\theta_t))_j - 2 \sum_{k=1}^M \left( G^{-1}(\theta_t) \frac{\partial G(\theta_t)}{\partial \theta_{jk}} (G^{-1}(\theta_t))_{jk} \right)_{jk} + \sum_{k=1}^M (G^{-1}(\theta_t))_{jk} \text{tr} \left( G^{-1}(\theta_t) \frac{\partial G(\theta_t)}{\partial \theta_{jk}} \right)_{jk}. \]
The SGRLD makes moves along a Riemannian manifold defined by the metric $G(\theta)$, 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 $G(\theta)$, which has convenient forms for some models.

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

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

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

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 $D$ after a few burn-in steps.

To apply the SGRLD algorithm, we need to choose a metric $G(\theta)$ for the Riemannian manifold. Because the posterior of $\phi_k$ is Dirichlet, we thus use the same Riemannian metric as in LDA (Patterson & Teh, 2013), e.g., $G(\phi) = \text{diag}(\phi_{11}, \ldots, \phi_{ PP}, \ldots, \phi_{kK}, \ldots, \phi_{ P K})^{-1}$, 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 $\phi_{pk}$ as:

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

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

For the other global model parameters such as $r_k$ and $\gamma_0$, we make use of their posterior distributions in (2) and (3), respectively. We only briefly describe the sampling for $r_k$, sampling $\gamma_0$ is similar. To simplified, rewrite the posterior $r_k$ as

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

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 $L_k$. Now we use the reparameterization for $r_k$ as $r_k = e^{\tilde{r}_k}$ with $\tilde{r}_k \sim \text{Log-Gamma}(\gamma_0, 1/c_0)$, 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

$$\frac{\partial \log p(\tilde{r}_k \mid \cdot)}{\partial \tilde{r}_k} = \frac{\partial}{\partial \tilde{r}_k} \mathbb{E}_{L_k} \left[ (a_k + \gamma_0 - 1)\tilde{r}_k - e^{-(c_0 + b_k)\tilde{r}_k} \right] = \mathbb{E}_{L_k} \left[ a_k + \gamma_0 - 1 - (c_0 + b_k)r_k^{c_0 + b_k} \right].$$

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 $\mathbb{Y}$. For the PFA-based models, the test perplexity is calculated as (Zhou et al., 2012)

$$\text{exp} \left( -\frac{1}{y} \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{s=1}^{S} \sum_{k=1}^{K} \phi_{pk}^s \phi_{kn}^s + \sum_{s=1}^{S} \sum_{p=1}^{P} \sum_{k=1}^{K} \phi_{pk}^s \phi_{kn}^s \right),$$

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

The conditional distribution of $y_{pn}$ given $h_n$, in the Replicated Softmax model (RSM) is specified as

$$y_{pn} \sim \text{Multi} (D_n ; \beta_n),$$

$$\beta_{pn} = \frac{\exp(w_{p}^\top h_n + c_p)}{\sum_{p'=1}^{P} \exp(w_{p'}^\top h_n + c_{p'})},$$

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

$$\text{exp} \left( -\frac{1}{y} \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{s=1}^{S} \sum_{k=1}^{K} \phi_{pk}^s \phi_{kn}^s \right).$$

**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.

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.

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


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.


