Supplemental Information: Streaming Variational Inference for Bayesian Nonparametric Mixture Models

Alex Tank
Department of Statistics
University of Washington

Nicholas J. Foti
Department of Statistics
University of Washington

Emily B. Fox
Department of Statistics
University of Washington

1 Derivation of Global Update

This section motivates the use of the mean field update for the global variables, given in Eq. (19) of the main text, as an approximation to the optimal update for ADF. The presentation adapts that of [1] for ADF-NRM.

Let $\hat{p}(\theta, z_{1:n}|x_{1:n}) \propto p(x_n|\theta, z_n)p(z_n|z_{1:n})\hat{q}(\theta, z_{1:n-1})$ denote the approximate posterior under the past variational updates after adding in the nth observation. The optimal $q(\theta_k)$ under ADF is given by the marginal distribution of $\hat{p}$:

$$\hat{q}_n(\theta_k) \propto \int \sum_{z_{1:n}} \hat{p}(\theta_k|\theta_{\setminus k}, z_{1:n}, x_{1:n})\hat{p}(\theta_{\setminus k}|z_{1:n}, x_{1:n})\hat{p}(z_{1:n}|x_{1:n})d\theta_{\setminus k}. \tag{1}$$

Both the sums and integrals are intractable so we use the approximations: $\hat{p}(\theta_{\setminus k}|z_{1:n}, x_{1:n}) \approx \hat{q}_n(\theta_{\setminus k})$ and $\hat{p}(z_{1:n}|x_{1:n}) \approx \hat{q}_n(z_{1:n}) = \prod_{i=1}^n \hat{q}_n(z_i)$ which yields:

$$\begin{aligned}
\hat{q}_n(\theta_k) &\overset{\approx}{\propto} \int \sum_{z_{1:n}} \hat{p}(\theta_k|\theta_{\setminus k}, z_{1:n}, x_{1:n})\hat{q}_n(\theta_{\setminus k})\hat{q}_n(z_{1:n})d\theta_{\setminus k} && (2)\\
&= E_{\hat{q}(z_{1:n}),\hat{q}(\theta_{\setminus k})}[\hat{p}(\theta_k|\theta_{\setminus k}, z_{1:n}, x_{1:n})] && (3)\\
&= \exp\{\log E_{\hat{q}(z_{1:n}),\hat{q}(\theta_{-k})}[\hat{p}(\theta_k|\theta_{\setminus k}, z_{1:n}, x_{1:n})]\} && (4)\\
&\leq \exp\{E_{\hat{q}(z_{1:n}),\hat{q}(\theta_{\setminus k})}[\log \hat{p}(\theta_k|\theta_{\setminus k}, z_{1:n}, x_{1:n})]\} && (5)\\
&\propto \exp\{E_{\hat{q}(z_{1:n}),\hat{q}(\theta_{\setminus k})}[\log \hat{p}(\theta, z_{1:n}|x_{1:n})]\} && (6)
\end{aligned}$$

where the inequality follows by Jensen's inequality [1]. The approximation is tight when $\hat{q}(z_{1:n})$ and $\hat{q}(\theta_{\setminus k})$ approach Dirac measures. Eq. (6) is that of the standard mean field update for $\hat{q}(\theta_k)$ [2]. Since the $q(\theta_k)$ distributions are unknown for all k, we could perform coordinate ascent and cycle through these updates for each of the θ_k given the other $\theta_{\setminus k}$ and $\hat{q}(z_{1:n})$. Conveniently, since the $\hat{q}(z_{1:n})$ is already optimized and the θ_ks are conditionally independent given the assignments in the mixture model we can perform a single mean field update for each θ_k,

$$\hat{q}_n(\theta_k) \propto p(x_n|z_{nk}, \theta_k)^{\hat{q}(z_{nk})}\hat{q}_{n-1}(\theta_k). \tag{7}$$

2 Derivation of approximate NRM predictive rule

In this section we provide the derivation of $q^{\mathrm{pr}}(z_n)$ for NRMs given in Eq. (25) of the main text. We start by presenting the derivation for general NRMs and then demonstrate how to apply ideas to NGGPs. The presentation in this section is adapted from [3, 4].

2.1 General NRMs

We assume the mixture model specification in Eq. (3) from the main text. In particular we note that the unnormalized mixture weights $\pi = (\pi_1, \pi_2, \ldots)$ are drawn from a completely random measure with Lévy measure $\lambda(d\pi)$. We also introduce the *exponentially tilted Lévy measure* as $e^{-U\pi}\lambda(d\pi)$ which will appear below.

First, we expand the sum in the approximate predictive distribution, $q^{\mathrm{pr}}(z_n)$, to include the unnormalized masses, π, and the auxiliary variable U_{n-1}:

$$q^{\mathrm{pr}}(z_n) = \sum_{z_{1:n-1}} p(z_n|z_{1:n-1}) \prod_{i=1}^{n-1} \hat{q}_{n-1}(z_i) \tag{8}$$

$$= \sum_{z_{1:n-1}} \iint p(z_n|\pi)p(\pi|U_{n-1}, z_{1:n-1})p(U_{n-1}|z_{1:n-1})dU_{n-1}d\pi \prod_{i=1}^{n-1} \hat{q}_{n-1}(z_i) \tag{9}$$

where the conditional distribution of the auxiliary variables U_{n-1} given the past assignments is given by:

$$p(U_{n-1}|z_{1:n-1}) = U_{n-1}^{n-1} e^{-\phi(U_{n-1})} \prod_{k=1}^{K_{n-1}} \kappa_{n_k}(U_{n-1}) \tag{10}$$

where $\phi(U)$ is the Laplace exponent of the underlying CRM, $\phi(U) = \int(1 - e^{-Us})\lambda(ds)$, and $\kappa_m(U)$ denotes the mth moment of the exponentially tilted Lévy measure, $\kappa_m = \int s^m e^{-Us}\lambda(ds)$.

Let K_{n-1} denote the number of components considered for the observations $z_{1:n-1}$. The conditional distribution of the random measure, $\pi = (\pi^*, \pi_1, \ldots, \pi_{K_{n-1}})$, given U_{n-1} and the assignments, $z_{1:n-1}$, is:

$$p(\pi|U_{n-1}, z_{1:n-1}) = p(\pi^*|U_{n-1}) \prod_{k=1}^{K_{n-1}} p(\pi_k|z_{1:n-1}, U_{n-1}). \tag{11}$$

where $\pi_{1:K_{n-1}}$ are the masses of all the instantiated components and π^* denotes the mass assigned to the uninstantiated components. The distribution of π_k is given by

$$p(\pi_k|z_{1:n-1}, U_{n-1}) \propto \pi_k^{n_k} e^{-U_{n-1}\pi_k}\lambda(d\pi_k), \tag{12}$$

where n_k is the number of observations assigned to component k in $z_{1:n-1}$ and π^* follows a Poisson process (PP) with exponentially tilted Lévy measure, $\pi^* \sim \mathrm{PP}(e^{-U_{n-1}\pi^*}\lambda(d\pi^*))$, where again $\lambda(ds)$ is the Lévy measure of the unnormalized masses. Since the integral in Eq (9) is intractable, we introduce a variational approximation for π and U_{n-1}. In particular, we use a partially factorized approximation

$$p(\pi|U_{n-1}, z_{1:n-1})p(U_{n-1}|z_{1:n-1})\hat{q}(z_{1:n-1}) \approx q(\pi|U_{n-1})q(U_{n-1})\hat{q}(z_{1:n-1}), \tag{13}$$

where $\hat{q}(z_{1:n-1}) = \prod_{i=1}^{n-1} \hat{q}_{n-1}(z_i)$ is fixed and given from previous iterations. We perform a mean field step to minimize the KL divergence between the left and right hand sides of Eq (13). Specifically, we compute the optimal $q(U_{n-1})$ and then given that we compute the optimal $q(\pi|U_{n-1})$. Because of the factorization

given in the left hand of Eq. (13) this procedure gives the optimal distributions. According to standard mean field updates the optimal distribution for $q(U_{n-1})$ is given by:

$$\log q(U_{n-1}) = E_{\hat{q}(z_{1:n-1})} \log p(U_{n-1}|z_{1:n-1}) + C \tag{14}$$

where $p(U_{n-1}|z_{1:n-1})$ is given in Eq. (10). The tractability of this variational approximation for U_{n-1} will depend on the NRM under consideration. For the NGGP it is conveniently given in closed form, as detailed below in Section 2.2. However, efficient numerical algorithms can be used to compute the necessary integrals for general NRMs.

Given the optimal $q(U_{n-1})$, the optimal variational approximations to the masses, $q(\pi|U_{n-1}) = q(\pi^*|U_{n-1}) \prod_{i=1}^{K_{n-1}} q(\pi_j|U_{n-1})$, are given by

$$q(\pi_k|U_{n-1}) \propto \pi_k^{\mathbb{E}_{\hat{q}}[n_k]} e^{-U\pi_k} \lambda(d\pi_k) \text{ for } k = 1 \dots K_{n-1}, \tag{15}$$

where $\mathbb{E}_{\hat{q}}[n_k]$ is the expected number of assignments to component k and is given by:

$$\mathbb{E}_{\hat{q}}[n_k] = \sum_{i=1}^{n-1} \hat{q}(z_{ik}). \tag{16}$$

Under $q(\pi^*|U_{n-1})$, π^* is still drawn from $\mathrm{PP}(e^{-U_{n-1}w}\lambda(dw))$. Using these approximations Eq. (9) becomes

$$q^{\mathrm{pr}}(z_n) = \sum_{z_{1:n-1}} \iint p(z_n|\pi) p(\pi|U, z_{1:n-1}) p(U_{n-1}|z_{1:n-1}) dU_{n-1} d\pi \prod_{i=1}^{n-1} \hat{q}_{n-1}(z_i) \tag{17}$$

$$\approx \iint p(z_n|\pi) q(\pi|U_{n-1}) q(U_{n-1}) d\pi dU_{n-1} \tag{18}$$

$$= \int q(z_n|U_{n-1}) q(U_{n-1}) dU_{n-1} \tag{19}$$

where

$$q(z_{nk}|U_{n-1}) \propto \begin{cases} \max\left(\frac{\kappa_{E_{\hat{q}}[n_k]+1}(U_{n-1})}{\kappa_{\mathbb{E}_{\hat{q}}[n_k]}(U_{n-1})}, 0 \right), & \text{if } k \leq K_{n-1} \\ \kappa_1(U_{n-1}), & \text{if } k = K_{n-1}+1. \end{cases} \tag{20}$$

Eq. (19) arises from (18) by an application of Prop. 2.1 in [3]. In Eq. (20), the maximum with zero is necessary since if the expected number of clusters assigned to a cluster k, $\mathbb{E}_{\hat{q}}[n_k]$, is small then the variational distribution for π_k given in Eq. (15) might be degenerate at zero and so there will be zero probability of a new observation being assigned to that cluster. More details for the NGGP case are given in Section 2.2.

2.2 Predictive Rule for the NGGP

For NGGPs, the general equations for NRMs described above reduce to simple, analytically tractable forms. In particular, the variational approximation $q(U_{n-1})$ is given by

$$q(U_{n-1}) \propto \frac{U_{n-1}^{n-1}}{(U_{n-1}+\tau)^{n-1-a\mathbb{E}_{\hat{q}(z_{1:n-1})}[K'_{n-1}]}} e^{-\frac{a}{\sigma}(U_{n-1}+\tau)^{\sigma}} \tag{21}$$

where $\mathbb{E}_{\hat{q}(z_{1:n-1})}[K'_{n-1}]$ is the expected number of clusters instantiated thus far. This expectation is given by:

$$\mathbb{E}_{\hat{q}(z_{1:n-1})}[K'_{n-1}] = K_{n-1} - \sum_{j=1}^{K_{n-1}} \left(\prod_{i=1}^{n-1} (1 - \hat{q}(z_{ij})) \right) \tag{22}$$

$$\overset{n\to\infty}{\to} K_{n-1}. \tag{23}$$

Note that Eq. (22) does not require all past soft assignments to be saved; instead, only $\prod_{i=1}^{n-1}(1-\hat{q}(z_{ij}))$ must be stored for each component and updated after each observation. In practice we find that using $\mathbb{E}_{\hat{q}(z_{1:n-1})}[K'_{n-1}] \approx K_{n-1}$ leads to comparable performance to evaluating the complete expectation. This occurs because, given our thresholding scheme for mixture components, each component has a few $\hat{q}(z_{ik})$ that are close to one, making the product close to zero.

Additionally, in the case of the NGGP the $\kappa_m(U)$ functions needed in Eq. (19) are given by

$$\kappa_m(U) = \frac{a}{(U+\tau)^{m-\sigma}} \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}, \tag{24}$$

which when plugged into Eq. (20) yields

$$q(z_{nk}|U_{n-1}) \propto \begin{cases} \max\left(\sum_{i=1}^{n-1} \hat{q}(z_{ik}) - \sigma, 0\right), & \text{if } k \leq K_{n-1} \\ a(U_{n-1}+\tau)^{\sigma}, & \text{if } k = K_{n-1}+1. \end{cases} \tag{25}$$

When we approximate the integral in Eq. (19) with a delta function about the maximum, $\hat{U}_{n-1} = \arg\max q(U_{n-1})$ we see that $q^{\text{pr}}(z_{nk}) \approx q(z_{nk}|\hat{U}_{n-1})$, which is exactly Eq. (25) of the main text. Alternatively, one could compute the integral in Eq. (19) numerically by first performing a change of variables, $V_{n-1} = \log U_{n-1}$, to obtain a log-convex density over V_{n-1} and then use adaptive rejection sampling to sample from V_{n-1}, as proposed in [3]. The efficiency of this method depends on the sampling process and we leave such investigations to future work. Intuitively, $q(z_{nk}|U_{n-1}) = 0$ for some k when $\sum_{i=1}^{n-1} \hat{q}_i(z_{ik}) < \sigma$ since $q(\pi_k|U_{n-1})$ will be degenerate in Eq. (15). This means that σ acts as a natural threshold for the instantiated clusters as clusters with mass (under the variational distribution) smaller than σ will have zero probability of having observations assigned to it.

3 EP-NRM derivation

In this section we modify the EP derivation in [5] for our EP-NRM algorithm for batch inference. The resulting algorithm is conceptually similar to ADF-NRM, except now we also save a local contribution to the variational approximation for each data point. The algorithm cycles through the observations repeatedly, refining the variational approximations for $z_{1:N}$ and θ. Due to the fact that local contributions must be saved, the algorithm is applicable to moderately sized data sets. The full EP-NRM algorithm is shown in Alg. 1.

Assume we have an approximation to the batch posterior

$$p(\theta, z_{1:N}|x_{1:N}) \approx \hat{q}(\theta, z_{1:N}) = \prod_{k=1}^{\infty} \hat{q}(\theta_k) \prod_{i=1}^{N} \hat{q}(z_i) \propto \prod_{i=1}^{N} \bar{q}_i(\theta, z_{1:n}), \tag{26}$$

where $\bar{q}_i(\theta, z_{1:n})$ are the *local contributions* as described in the main text. Furthermore, assume that

$$\bar{q}_i(\theta, z_{1:n}) = \bar{q}_i(z_i) \prod_{k=1}^{\infty} \bar{q}_i(\theta_k) \tag{27}$$

and that $\bar{q}_i(z_i) = \hat{q}(z_i)$. This holds initially since since $\bar{q}_i(\theta, z_{1:n})$ is initialized during ADF to $\bar{q}_i(\theta, z_{1:n}) \propto \frac{\hat{q}_i(\theta, z_{1:i})}{\hat{q}_{i-1}(\theta, z_{1:i-1})}$. Since $\bar{q}_i(\theta, z_{1:N})$ only depends on z_i we henceforth refer to this quantity as $\bar{q}_i(\theta, z_i)$. Under

these assumptions we can rewrite the approximation to the full posterior excluding data point i as

$$\hat{q}_{\backslash i}(\theta, z_{\backslash i}) \propto \frac{\hat{q}(\theta, z_{1:N})}{\bar{q}_i(\theta, z_i)} \tag{28}$$

$$= \frac{\prod_{k=1}^{\infty} \hat{q}(\theta_k) \prod_{j=1}^{n} \hat{q}(z_j)}{\hat{q}(z_i) \prod_{k=1}^{\infty} \bar{q}_i(\theta_k)} \tag{29}$$

$$= \prod_{k=1}^{\infty} \frac{\hat{q}(\theta_k)}{\bar{q}_i(\theta_k)} \prod_{j \neq i} \hat{q}(z_j). \tag{30}$$

$$= \prod_{k=1}^{\infty} \hat{q}_{\backslash i}(\theta_k) \prod_{j \neq i} \hat{q}(z_j). \tag{31}$$

The EP-NRM algorithm consists of the following two steps. First, update the global variational approximations, $\hat{q}(\theta_k)$ and $\hat{q}(z_i)$. Second, use these to refine $\bar{q}_i(\theta_k)$ and $\bar{q}_i(z_i)$ (see Alg. 1). Define $\hat{p}(\theta, z_{1:N}) \overset{\triangle}{\propto} \hat{q}_{\backslash i}(\theta, z_{\backslash i}) p(x_i|z_i, \theta) p(z_i|z_{\backslash i})$. The optimal variational distributions are found by solving

$$\hat{q}(\theta, z_{1:n}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(\hat{p}(\theta, z_{1:N}) || q(\theta, z_{1:N})). \tag{32}$$

As in ADF-NRM we do not need to update $\hat{q}(z_j)$ for $j \neq i$; the optimal update for $\hat{q}(z_i)$ is given by:

$$\hat{q}(z_{ik}) \propto q_{\backslash i}^{\mathrm{pr}}(z_{ik}) \int p(x_i|z_{ik}, \theta_k) \hat{q}(\theta_k) d\theta_k \quad k = 1, \ldots, K+1, \tag{33}$$

where K is the number of instantiated clusters and $\hat{q}(\theta_{K+1}) = p(\theta_{K+1})$. Similar to ADF, the predictive distribution for the NGGP in Eq.(33) is given by:

$$q_{\backslash i}^{\mathrm{pr}}(z_{ik}) \propto \begin{cases} \max\left(\sum_{i \neq j} \hat{q}(z_{jk}) - \sigma, 0\right), & k \leq K \\ a(\hat{U}_{\backslash i} + \tau)^{\sigma}, & k = K+1 \end{cases} \tag{34}$$

where $q(U_{\backslash i})$ is given by:

$$q(U_{\backslash i}) \propto \frac{U_{\backslash i}^{N-1}}{(U_{\backslash i} + \tau)^{N-1-a\mathbb{E}_{\hat{q}(z_{\backslash i})}[K']}} e^{-\frac{a}{\sigma}(U_{\backslash i}+\tau)^{\sigma}}. \tag{35}$$

where K' is the number of unique assignments in $z_{\backslash i}$ and $\hat{U}_{\backslash i} = \arg\max q(U_{\backslash i})$.

Following the ADF discussion in the main text, the optimal variational distributions for the θ_ks are given by:

$$\hat{q}(\theta_k) \propto p(x_i|z_{ik}, \theta_k)^{\hat{q}(z_{ik})} \hat{q}_{\backslash i}(\theta_k). \tag{36}$$

Given updated approximations $\hat{q}(\theta_k)$ and $\hat{q}(z_i)$, the local contribution for observation i is refined as:

$$\bar{q}_i(\theta, z_i) = \frac{\hat{q}(\theta, z_{1:N})}{\hat{q}_{\backslash i}(\theta, z_{\backslash i})} \tag{37}$$

$$= \hat{q}(z_{ik}) \prod_{k=1}^{\infty} p(x_i|z_{ik}, \theta_k)^{\hat{q}(z_{ik})} \tag{38}$$

$$= \bar{q}_i(z_i) \prod_{k=1}^{\infty} \bar{q}_i(\theta_k) \tag{39}$$

which takes the form we assumed in Eq. (27).

When $\hat{q}(\theta_k)$ is in the exponential family with sufficient statistics $\hat{\nu}^k$, then Eqs. (30) and (37) are given adding and subtracting the corresponding sufficient statistics [5].

Algorithm 1 EP-NRM algorithm

$\hat{q}(\theta_{1:K}), S_{1:K}, \bar{q}(z_{1:N}), \bar{q}_{1:N}(\theta_{1:K}) \leftarrow$ ADF-NRM$(x_{1:N})$ // Initialize via ADF with data contributions.
while $\hat{q}(\theta_{1:K})$ not converged **do**
  for $i = 1$ **to** N **do**
    $\hat{U}_{\setminus i} = \arg\max q(U_{\setminus i})$
    for $k = 1$ to K **do**
      $S_k = S_k - \bar{q}_i(z_{ik})$
      $q^{\mathrm{pr}}(z_{ik}) \propto \max(S_k - \sigma, 0)$
      $\hat{q}_{\setminus i}(\theta_k) \propto \frac{\hat{q}(\theta_k)}{\bar{q}_i(\theta_k)}$
      $\hat{q}(z_{ik}) \propto q^{\mathrm{pr}}(z_{ik}) \int p(x_i|z_{ik}, \theta_k)\hat{q}_{\setminus i}(\theta_k)d\theta_k.$
    end for
    $q^{\mathrm{pr}}(z_{i,K+1}) \propto a(\hat{U}_{\setminus i} + \tau)^{\sigma}$
    $\hat{q}(z_{i,K+1}) \propto q^{\mathrm{pr}}(z_{i,K+1}) \int p(x_i|z_{i,K+1}, \theta)p(\theta_{K+1})d\theta_{K+1}$
    normalize $\hat{q}(z_{i(1:K+1)})$
    if $\hat{q}(z_{i,K+1}) > \epsilon$ **then**
      $K = K + 1$, $S_K = 0$, $\hat{q}(\theta_K) = p(\theta_K)$
    else
      normalize $\hat{q}(z_{i(1:K)})$
    end if
    for $k = 1$ to K **do**
      $\hat{q}(\theta_k) \propto p(x_i|z_{ik}, \theta_k)^{\hat{q}(z_{ik})}\hat{q}_{\setminus i}(\theta_k)$
      $S_k = S_k + \hat{q}(z_{ik})$
      $\bar{q}_i(z_{ik}) = \hat{q}(z_{ik})$
      $\bar{q}_i(\theta_k) \propto p(x_i|z_{ik}, \theta_k)^{\hat{q}(z_{ik})}$
    end for
    Remove all clusters for which $S_k < \epsilon$
  end for
end while

4 Experiments

In this section we provide details on how we select hyperparameter values of the IG and DP for the experiments in the main text. We also present additional experimental results regarding the convergence of EP-NRM and comparisons with the Gibbs sampler.

4.1 Selecting Hyperparameters: a, τ, and α

In order to compare the modeling performance of the IG and DP on the document corpora considered in the main text, we must first select the values of the hyperparameters, a and τ (since σ is known in both cases). It is well known that the hyperparameters of both the IG (a and τ) and the DP (a) strongly affect the posterior distribution over the number of inferred clusters. For all experiments where the IG and DP are compared we adapt a method to determine the hyperparameters for GGP mixture models originally developed for batch inference [6] to the streaming setting of interest. Specifically, for a given corpus, we consider a small

subset of the entire corpus (5% for the NYT corpus and 10% for both the KOS and synthetic data) which we then split into a training and testing sets used to determine the hyperparameters. We run ADF-NRM on the training portion of the subset of documents (95% of the subset for NYT and 80% for both KOS and synthetic data) for a grid of parameters $a \in [1, 10, 100, 1000]$ and $\tau \in [.1, 1, 10, 100, 1000]$. For the DP we only consider a and for the IG we consider both a and τ. For each parameter value we compute the heldout log-likelihood of the test portion of the subset and choose the values of a and τ with the largest heldout log-likelihood to use when running ADF-NRM and EP-NRM on the remainder of the corpora. This setup mimics a streaming scenario in that an initial subset of the data is collected for preliminary analysis and then the algorithm is let loose on the entire data set as it arrives.

For the Pitman-Yor data set, the grid search resulted in $a = 100$ for the DP and $a = 1, \tau = 1000$ for the IG. The resulting parameter values for the KOS corpus were $a = 100$ for the DP and $a = 10, \tau = 100$ for the IG. Last, on the NYT corpus we obtained the parameter values $a = 1000$ for the DP and $a = 100, \tau = 100$ for the IG.

In the synthetic bars experiments α was set to 0.5, however correct recovery of the bars was robust to values within a reasonable range, $\alpha \in [0.1, 0.9]$. For the Pitman-Yor synthetic data, the cluster centers were drawn from a Dirichlet with $\alpha = 0.75$ to ensure overlap between clusters; $\alpha = 0.75$ was used for inference as well. For the KOS corpus $\alpha = 0.1$ was used because it was found to provide the best overall fit under repeated trials. Finally, for the NYT data set $\alpha = 0.5$ was used, as is common for this corpus [1].

4.2 KOS Corpus

While the ADF-NRM algorithm makes a single pass through the corpus, a more accurate posterior approximation can be achieved by revisiting observations as in EP-NRM. Figure 1 shows the predictive performance for EP-NRM applied to the KOS corpus. We see a rapid increase in predictive performance in the first epoch which corresponds to ADF-NRM. Predictive performance continues to rise during subsequent epochs indicating an improved variational posterior.

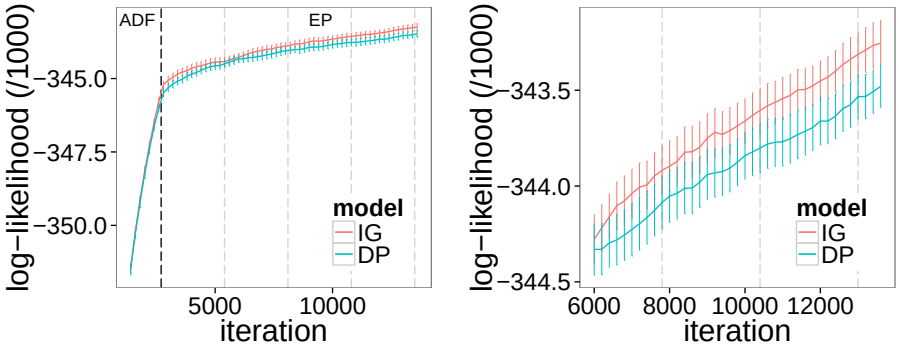

Figure 1: (*left*) EP predictive performance on KOS corpus for both models continues to rise after the pass through the data (equivalent to ADF-NRM). The black vertical line indicates the completion of the first pass through the data. The other grey vertical lines indicate subsequent epochs. (*right*) Zoom in of the plot on the left.

We compare the predictive performance of ADF-NRM, EP-NRM, and the Gibbs sampler for the IG model on the KOS corpus and present the results in Figure 2. In particular, we compare the predictive log-likelihood of held-out data versus the number of complete passes through the data (epochs). Both ADF-NRM and EP-NRM are initialized as in the main text and the Gibbs sampler is initialized so that all data points are

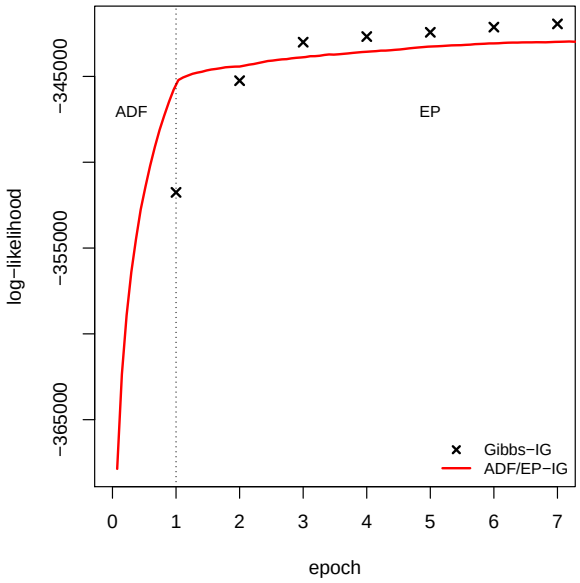

Figure 2: Comparison of the predictive performance of ADF-NRM, EP-NRM, and the Gibbs sampler for both the IG model. The predictive log-likelihood is plotted against the number of epochs through the KOS corpus.

assigned to a single component. We found this Gibbs initialization to outperform random cluster initialization. ADF-NRM performs significantly better than Gibbs after the first epoch and it takes three full epochs for Gibbs to outperform ADF-NRM and EP-NRM method. Both methods are implemented in Python and a per epoch timing comparison shows that ADF-NRM takes an average of 220 seconds per epoch while the Gibbs sampler takes an average of 160 seconds per epoch. The ADF-/EP-NRM methods take longer since the auxiliary variable U must be updated after each data point has been processed, while in the collapsed sampler U is only sampled once per epoch. Furthermore, in the Gibbs sampler, after a cluster assignment has been sampled only the sufficient statistics for the corresponding component must be updated, while in ADF-NRM, all component parameters are updated after every data point. Importantly, our goal is not to beat the Gibbs sampler, neither in performance nor compute time, but only to show that the streaming ADF-NRM reaches competitive performance to Gibbs after only a single pass through the data. Remember, Gibbs is inherently not suited to our streaming data of interest.

4.3 New York Times

As seen in the main paper, the IG both introduces more clusters than the DP and attains superior predictive performance. To further explore the difference in the inferred clusters between the two models we plot the normalized variational cluster weights in decreasing order in Figure 3. In particular, let $S_k = \sum_{i=1}^{N} \hat{q}_i(z_{ik})$ be the total weight assigned to cluster k after a full pass through the data and $\hat{p}_k = \frac{S_k}{\sum_{j=1}^{K_N} S_j}$ be the normalized weight. We can interpret $\hat{p}_k$ as the posterior probability of an observation being assigned to cluster k. We see in Figure 3 that the distribution of weights for the IG has a heavier tail than the DP. The plots are similar for the large and medium sized clusters but diverge for the small clusters, indicating that the IG emphasizes capturing structure at a finer scale.

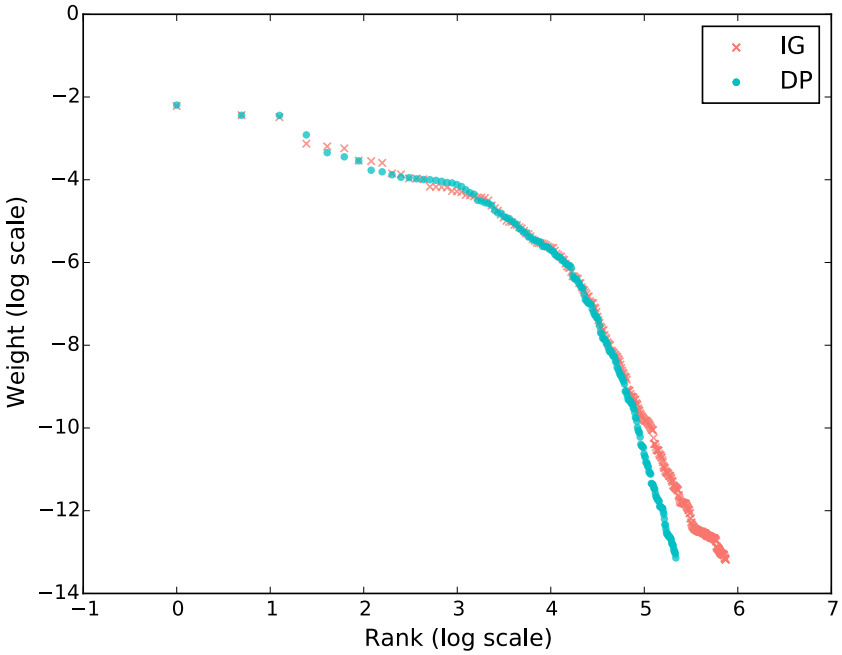

Figure 3: Variational cluster weights in decreasing order. The IG exhibits a heavier tail than the DP.

References

[1] C. Wang and D. M. Blei. Truncation-free online variational inference for Bayesian nonparametric models. In *Advances in Neural Information Processing Systems*. 2012.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[3] S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359, August 2013.

[4] L. F. James, A. Lijoi, and I. Prunster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[5] T. Minka. Expectation propagation for approximate Bayesian inference. In *Advances in Neural Information Processing Systems*, 2001.

[6] E. Barrios, A. Lijoi, L. E. Nieto-Barajas, and I. Prunster. Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334, 08 2013.