Stochastic Variational Inference for the HDP-HMM

Aonan Zhang San Gultekin John Paisley
Department of Electrical Engineering & Data Science Institute
Columbia University, New York, NY

Abstract

We derive a variational inference algorithm for the HDP-HMM based on the two-level stick breaking construction. This construction has previously been applied to the hierarchical Dirichlet processes (HDP) for mixed membership models, allowing for efficient handling of the coupled weight parameters. However, the same algorithm is not directly applicable to HDP-based infinite hidden Markov models (HDP-HMM) because of extra sequential dependencies in the Markov chain. In this paper we provide a solution to this problem by deriving a variational inference algorithm for the HDP-HMM, as well as its stochastic extension, for which all parameter updates are in closed form. We apply our algorithm to sequential text analysis and audio signal analysis, comparing our results with the beam-sampled iHMM, the parametric HMM, and other variational inference approximations.

1 Introduction

The hierarchical Dirichlet process (HDP) (Teh et al., 2006) is a Bayesian nonparametric prior for generating multiple random measures on the same countably infinite collection of atoms. This property of the HDP makes it a natural tool for modeling groups of data that share hidden components with different mixing proportions. The most well-known application of the HDP is for handling exchangeable data through mixed membership modeling (Airoldi et al., 2014), as well as nonexchangeable data with hidden Markov models (HDP-HMM) (Fox et al., 2008; Teh et al., 2006).

The hierarchical structure of the HDP makes inference a significant problem. For example, various sampling strategies have been developed for these models to improve efficiency: For the mixed membership model, a Chinese restaurant franchise (Teh et al., 2006) sampling method was proposed, while for the HDP-HMM a beam sampling strategy was introduced to enable forward-backward sampling via slice sampling (Neal, 2003) over a dynamic, truncated stick-breaking construction (Van-Gael et al., 2008).

Variational inference provides another promising strategy for inference in Bayesian hierarchical models by restricting the posterior to a simpler form that is able to be deterministically optimized (Jordan et al., 1999). Moreover, stochastic variational inference (SVI) allows for efficient inference over large datasets, and has been applied successfully on locally exchangeable data (Hoffman et al., 2010, 2013; Wang et al., 2011) and nonexchangeable data (Foti et al., 2014; Johnson and Willsky, 2014). For the HDP mixed membership model, batch and stochastic variational inference algorithms have been derived using a two-level stick breaking construction (Wang et al., 2011). However, these algorithms are not immediately transferable to the HDP-HMM because of sequential dependencies in the local variables. Alternative, fully conjugate nonparametric priors for the HMM have also been proposed (Paisley and Carin, 2009).

Previous work has focused on SVI for hidden Markov models (Foti et al., 2014), but is not directly applicable to the HDP-HMM. A recent SVI approach to the HDP-HMM is based on a point estimate strategy to avoid non-conjugacy (Liang et al., 2007; Johnson and Willsky, 2014). In this paper we address posterior inference for the HDP-HMM over all variables by deriving batch and stochastic variational algorithms using the fully conjugate representation of Wang et al. (2011) rather than the representation by Johnson and Willsky (2014), with which we compare.

In Section 2, we present the construction of the HDP-HMM we use for inference, and derive batch and stochastic variational inference algorithms in Section 3. We then apply our model to both artificial data and real data in Section 4, including a sequential text dataset and a large-scale audio dataset. Empirical results demonstrate the effectiveness of our method when compared with the beam-sampled iHMM (Van-Gael et al., 2008), the HDP-HMM with simpler direct assignment variational approximations (Liang et al., 2007; Johnson and Willsky, 2014) and its split-merge variation (Bryant and Sudderth, 2012), as well as the parametric batch and stochastic HMM (Beal, 2003; Foti et al., 2014).

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

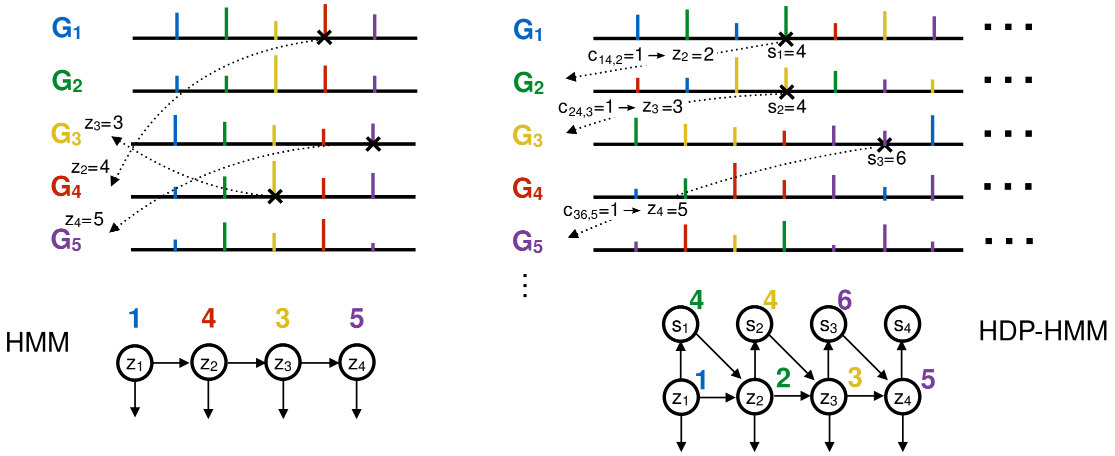

Figure 1: Comparison between the transition structure of the HMM and the stick-breaking construction of HDP-HMM. Each G_i represents a transition distribution from state i, and the color of a stick corresponds to one state. Left: In the HMM the state and the column index are one-to-one. Right: For the stick-breaking construction of the HDP-HMM, multiple sticks in each row may point to the same state (same color) and there is no one-to-one mapping between column and row. For example, a transition from state 3 to 5 takes place if $s = 5$ or $s = 6$. By holding the sequence $\mathbf{s}_d$ fixed and changing a DP indicator c_{km} for a selected stick (i.e., changing the color of a stick), the state transitions for *all* subsequent states may change. This presents an inference challenge for the HDP-HMM not faced by Wang et al. (2011).

2 The HDP-HMM

The HDP-HMM uses the hierarchical Dirichlet process for Bayesian nonparametric inference over the number of states in a hidden Markov model. To review, let G be a top-level Dirichlet process, written $G \sim \text{DP}(a_0\mu)$, with $a_0 > 0$ and μ a non-atomic probability measure. Since G is a.s. discrete, we can write $G = \sum_k \eta_{0k}\delta_{\theta_k}$. The HDP uses G as the base distribution of a possibly infinite number of second-level Dirichlet processes, written $G_k \sim_{iid} \text{DP}(\tau_0 G)$.

In the context of the HDP-HMM, G_k is the transition distribution for state k. To generate a sequence of data, one generates a sequence of parameters $(\theta'_1, \ldots, \theta'_n)$ by first sampling θ'_1 from an initial-state distribution G_0 followed by the rest of the sequence. To this end, we introduce the state index z_i, which equals k if $\theta'_i = \theta_k$. The next parameter in the Markov chain is then generated from the DP indexed by z_i, $\theta'_{i+1} \sim G_{z_i}$. The observed sequence $(x_1, \ldots, x_n)$ is generated using these parameters, where $x_i \sim p(x|\theta'_i)$. In this paper, we will focus on the discrete HMM.

2.1 Stick-breaking construction

Our inference method is based on the stick-breaking construction for the HDP (Sethuraman, 1994), which we briefly review. To generate the top-level DP, we let

$$G = \textstyle\sum_{k=1}^{\infty} \underbrace{\zeta_k \prod_{j=1}^{k-1}(1-\zeta_j)}_{\equiv\, \eta_{0k}} \delta_{\theta_k},$$

$$\zeta_k \overset{iid}{\sim} \text{Beta}(1, a_0), \quad \theta_k \overset{iid}{\sim} \text{Dir}(b_0\mathbf{1}). \tag{1}$$

The infinite number of second-level DP's are then drawn

$$G_k = \textstyle\sum_{m=1}^{\infty} \underbrace{\varepsilon_{km} \prod_{j=1}^{m-1}(1-\varepsilon_{kj})}_{\equiv\, \eta_{km}} \delta_{\phi_{km}},$$

$$\varepsilon_{km} \overset{iid}{\sim} \text{Beta}(1, \tau_0), \quad \phi_{km} \overset{iid}{\sim} G. \tag{2}$$

Since G is discrete almost surely, there is a mapping from ϕ_{km} to θ_i, and many ϕ_{km} will map to the same θ_i. This introduces additional complexity during inference that makes learning parameters more complicated than for the parametric HMM (Beal, 2003).

For inference we introduce the indicator vector c_{km}, which indexes the top-level atom picked for ϕ_{km}. Therefore $c_{km,k'} = 1$ if $\phi_{km} = \theta_{k'}$ and $c_{km} \sim \text{Mult}(\eta_0)$.[1] It turns out that for the HDP mixed membership model these indicator variables are especially important for closed-form updates in variational inference (Wang et al., 2011). Wang et al. (2011) draw from G_k by first drawing a stick indicator s and then mapping to the top-level atom associated with the chosen stick as indicated by c_{ks}.

We will use this auxiliary variable for the HDP-HMM as well. First, for the dth observed sequence, we sample $s_{d,i}|\{z_{d,i-1} = k\} \sim \text{Disc}(\eta_k)$ and then set $z_{d,i} = k'$ if $c_{ks_{d,i},k'} = 1$ to index the next state (see Figure 1). This two-step process of first selecting the stick and then mapping to the top level atom works easily for the HDP in the mixed membership setting, but this algorithm is not directly applicable to the HDP-HMM because of the non-exchangeability of the sequence. That is, given a sequence

[1] We will work only with c_{km} and ignore ϕ_{km} from now on.

Algorithm 1 An outline of VI for the HDP-HMM

```
Iterate the below updates to the variational distributions
    $q(\mathbf{z}_d)$: Forward-backward (appendix) with Eq. (6) approx.
    $q(\theta_k)$: See appendix for discrete HMM case.
    $q(\mathbf{s}_d|\mathbf{z}_d)$: See Eq. (9). This is used in following updates.
    $q(c_{km})$: See Eq. (10) and appendix for expectations.
    $q(\varepsilon_{km})$ and $q(\zeta_i)$: See appendix.
```

of stick indicators $\mathbf{s}_d$, if we change the value of c_{km} we may affect the path of the entire Markov chain $\mathbf{z}_d$. This problem of indicators pointing to indicators constitutes the challenge of variational inference for the HDP-HMM, which we illustrate in Figure 1.

In the next section we derive a variational inference algorithm that works with the marginal distribution (integrating out $\mathbf{s}_d$) to perform forward-backward on $\mathbf{z}_d$, and then reintroduces $\mathbf{s}_d$ as an auxiliary variable to perform local variational inference (Bishop, 2006).

3 Variational inference for the HDP-HMM

The variational objective function is formed by integrating over the model variables in the log joint likelihood using an approximation to the posterior distribution. Since we model each sequence as independent, the joint likelihood of the HDP-HMM can be factorized as

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\varepsilon}, \mathbf{c}, \mathbf{z}) &= p(\boldsymbol{\theta})p(\boldsymbol{\zeta})p(\boldsymbol{\varepsilon})p(\mathbf{c}|\boldsymbol{\zeta}) \\ &\quad \times \textstyle\prod_d p(\mathbf{z}_d|\boldsymbol{\varepsilon}, \mathbf{c}) \prod_i p(x_{di}|\boldsymbol{\theta}, z_{di}), \end{aligned} \tag{3}$$

with additional factorizations on all variables in $p(\boldsymbol{\zeta})$, $p(\boldsymbol{\varepsilon})$, $p(\mathbf{c}|\boldsymbol{\zeta})$ and $p(\boldsymbol{\theta})$. We approximate the posterior of these random variables with the distribution

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\varepsilon}, \mathbf{c}, \mathbf{z}) &= q(\boldsymbol{\theta})q(\boldsymbol{\zeta})q(\boldsymbol{\varepsilon})q(\mathbf{c})q(\mathbf{z}) \qquad (4) \\ &= \prod_k q(\theta_k)q(\zeta_k) \prod_{k,m} q(\varepsilon_{km})q(c_{km}) \prod_d q(\mathbf{z}_d). \end{aligned}$$

We set each variational distribution to be in the same family as the prior (see below for the explicit form). The goal is then to maximize the objective function

$$\mathcal{L} = \mathbb{E}_q[\ln p(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\varepsilon}, \mathbf{c}, \mathbf{z})] - \mathbb{E}_q[\ln q]$$

over the parameters of q to minimize the KL-divergence between q and the posterior (Jordan et al., 1999).

In the following batch inference algorithm, we work directly with $\mathcal{L}$ to update $q(\mathbf{z}_d)$ and $q(\theta_k)$. Using $\mathcal{L}$ for the remaining q distributions is more difficult, and so we introducing the latent variables $\mathbf{s}_d$ and variational distributions $q(\mathbf{s}_d|\mathbf{z}_d)$ to locally lower bound $\mathcal{L}$. This allows for closed-form updates of $q(\boldsymbol{\zeta})$, $q(\boldsymbol{\varepsilon})$ and $q(\mathbf{c})$. We focus on the novel aspects of our inference algorithm in the following subsections. The parts of our algorithm that overlap with other HMM inference algorithms are given in the appendix. We sketch one batch iteration in Algorithm 1.

3.1 The state transition matrix

The first issue we address is the state transition matrix, which we use to update $q(\mathbf{z}_d)$ and $q(\theta_k)$. Let $A_{kk'}$ be the probability of transitioning from state k to k'.

The challenge here is in the term

$$\mathbb{E}_q \ln A_{kk'} = \mathbb{E}_q \ln \sum_m c_{km,k'}\eta_{km} \tag{5}$$

where $\eta_{km} = \varepsilon_{km} \prod_j (1 - \varepsilon_{kj})$. We recall that this is the sum over all sticks that have been assigned to atom $\theta_{k'}$ for the stick-breaking construction of atom θ_k (i.e., state k). We must account for the possible assignment of each stick to $\theta_{k'}$ since the distribution on c_{km} is discrete almost surely. This expectation is not tractable, and so we form a lower bound and an approximation as follows,

$$\begin{aligned} \mathbb{E}_q \ln \textstyle\sum_m c_{km,k'}\eta_{km} &\geq \mathbb{E}_q \ln \displaystyle\sum_m c_{km,k'} e^{\mathbb{E}_q \ln \eta_{km}} \qquad (6) \\ &\approx \ln \sum_m \mathbb{E}_q[c_{km,k'}] e^{\mathbb{E}_q \ln \eta_{km}}, \end{aligned}$$

We observe that, since we only need the expectation $\mathbb{E}_q \ln \sum_m c_{km,k'}\eta_{km}$ for forward-backward, we could also have sampled c_{km} and η_{km} from their variational q distributions and formed an unbiased approximation. This resulted in a somewhat slower algorithm and we did not empirically observe any difference in performance. We report results with the approximation in Eq. (6) in this paper. We empirically observed that $q(c_{km})$ was nearly deterministic in general, so the approximation to the bound was good.

Making this approximation, we run forward-backward to find $q(\mathbf{z}_d)$ and then use this to update $q(\theta_k)$. These are found as in the parametric HMM (see the appendix).

3.2 A local lower bound using $q(\mathbf{s}_d|\mathbf{z}_d)$

Updating the remaining q distributions on $\boldsymbol{\varepsilon}$, $\mathbf{c}$ and $\boldsymbol{\zeta}$ is difficult because our approximation of the expected log state transition probabilities in Eq. (6) does not yield tractable variational parameter updates for these variables. We address this with a local lower bounding using the sequence $\mathbf{s}_d$. We recall that this latent sequence interacts with $\mathbf{z}_d$ and c_{km} as follows: The pair $(z_{d,i-1} = k, s_{di} = m)$ indicates that the next state z_{di} can be found by choosing the mth stick of the kth DP, and setting $z_{di} = k'$ if $c_{km,k'} = 1$.

In the variational HMM, the state transition from $z_{d,i-1}$ to z_{di} is captured by the marginal $q(z_{d,i-1}, z_{di})$, which is calculated using the output of the forward-backward algorithm. For the variational HDP-HMM, we instead model this transition via the triple $(z_{di}, s_{d,i+1}, c_{z_{di},s_{d,i+1}})$. Introducing these variables creates the local lower bound

$$\mathbb{E}_q \mathbb{1}(z_{d,i-1} = k, z_{di} = k') \ln \sum_m c_{km,k'} \eta_{km} \tag{7}$$

$$\geq \mathbb{E}_q \sum_m c_{km,k'} \mathbb{1}(z_{d,i-1} = k, s_{di} = m) \ln \eta_{km}.$$

We then define the joint variational distribution

$$q(\mathbf{s}_d, \mathbf{z}_d) = q(\mathbf{s}_d|\mathbf{z}_d)q(\mathbf{z}_d) = q(\mathbf{z}_d)\textstyle\prod_i q(s_{di}|\mathbf{z}_d),$$

where $q(\mathbf{z}_d)$ is already calculated using forward-backward.

The RHS of Eq. (7) involves three separate expectations because of the factorization of q. The expectation of η is discussed in the appendix. One novelty introduced by our construction of the HDP-HMM is the term

$$\begin{aligned} q(z_{d,i-1} = k, s_{di} = m) &= \mathbb{E}_q \mathbb{1}(z_{d,i-1} = k, s_{di} = m) \\ &\equiv \xi_{di}(k, m), \end{aligned} \tag{8}$$

which is the variational marginal probability of picking the mth stick from the kth DP in step i of sequence d. This value serves a similar purpose as the marginal state transition probability in the parametric HMM, only it is not a distribution between states, but between a state and a stick that must be mapped to a top-level DP state. We find ξ_{di} by calculating

$$\xi_{di}(k, m) \propto \exp\{\mathbb{E}_q \ln p(\mathbf{x}_d, z_{d,i-1} = k, s_{di} = m, -)\},$$

and normalizing ("$-$" indicates all other variables). Similar to the variational HMM, this requires the forward α_d and backward β_d calculations found when updating $q(\mathbf{z}_d)$ (see the appendix). As a result, we update the marginal

$$\xi_{di}(k, m) \propto \alpha_{d,i-1}(k) \exp\{\mathbb{E}_q \ln \eta_{km}\} \times \tag{9}$$

$$\textstyle\prod_{k'} \left[\exp\{\mathbb{E}_q[\ln \theta_{k',x_{di}}]\}\beta_{di}(k')\right]^{\varphi_{km,k'}}.$$

φ_{km} is the variational multinomial parameter for c_{km} derived later. The difference between this term and the corresponding term for the HMM is the product over the assignment of stick m (which is known *a priori* in that model). We obtain $\xi_{di}(k, m)$ by normalizing this matrix.

3.3 Mapping atoms between DP levels

For c_{km}, being the indicator of the atom associated with the mth stick in the kth DP, we let $q(c_{km,k'} = 1) \equiv \varphi_{km,k'}$ where

$$\varphi_{km,k'} \propto \exp\{\mathbb{E}_q \ln \eta_{0,k'} + \textstyle\sum_{d,i} \xi_{di}(k, m)\mathbb{E}_q \ln \theta_{k',x_{di}}\}. \tag{10}$$

We give the expectations in the appendix. A similar calculation appears in the HDP mixed membership model (Wang et al., 2011).

3.4 Stochastic variational inference

For computationally intensive scenarios in which we have a large collection of sequences over which to learn q, the proposed inference algorithm can be scaled with stochastic variational inference (Hoffman et al., 2013). SVI works in this context by subsampling a set of sequences $\mathbf{x}_d$, where $d \in B_t \subset \{1, \dots, D\}$ at iteration t. It then optimizes the q distributions for these sequences and takes a weighted step in the direction of the natural gradient of the global variational parameters.[2]

Since stochastic inference for $q(\theta_k)$ and $q(\varepsilon_{km})$ are common calculations, we discuss them in the appendix. The stochastic update for $q(c_{km})$ requires the following new SVI derivation. First, we restrict the scaled variational objective function to B_t and terms involving φ_{km},

$$\begin{aligned} \mathcal{L}^{(t)}_{c_{km}} = \sum_{k'} & \varphi_{km,k'}\mathbb{E}_q \ln \eta_{0,k'} - \varphi_{km,k'} \ln \varphi_{km,k'} + \\ & \varphi_{km,k'} \tfrac{D}{|B_t|} \sum_{d \in B_t, i} \xi_{di}(k, m)\mathbb{E}_q \ln \theta_{k',x_{di}}. \end{aligned} \tag{11}$$

The natural parameter of $q(c_{km})$ is $\ln \varphi_{km}$ (taking elementwise logarithm), and so the natural gradient update is

$$\ln \varphi_{km} \leftarrow \ln \varphi_{km} + \rho_t M_{km}^{-1} \nabla_{\ln \varphi_{km}} \mathcal{L}^{(t)}_{c_{km}}, \tag{12}$$

$$\begin{aligned} M_{km} &= \mathbb{E}_q \left[\frac{d \ln q(c_{km})}{d \ln \varphi_{km}} \frac{d \ln q(c_{km})}{d \ln \varphi^T_{km}} \right] \\ &= \text{diag}(\varphi_{km}). \end{aligned} \tag{13}$$

Next, we observe from Eq. (10) that we can write the update of this multinomial distribution in the form

$$q(c_{km,k'} = 1) = \varphi_{km,k'} \propto \exp\{\lambda_{km,k'}\}. \tag{14}$$

Swapping in this representation, we find that the natural gradient step over the scale term λ followed by the restriction to the simplex gives the update

$$\varphi_{km,k'} \propto \exp\{\lambda^{(t)}_{km,k'}\}, \tag{15}$$

where $\lambda^{(t)}_{km,k'}$ is the typical weighted average

$$\lambda^{(t)}_{km,k'} = (1 - \rho_t)\, \lambda^{(t-1)}_{km,k'} + \rho_t\, \lambda'_{km,k'} \tag{16}$$

$$\lambda'_{km,k'} = \mathbb{E}_q \ln \eta_{0,k'} + \frac{D}{|B_t|} \sum_{d \in B_t, i} \xi_{di}(k, m)\mathbb{E}_q \ln \theta_{k',x_{di}}$$

4 Experiments

We perform experiments on artificial data, batch inference experiments using the "Alice" dataset and large-scale experiments using discretized audio sequences. We list the methods we compare with in Table 1.

[2] We note that our algorithm will scale with the number of sequences, not the length of the sequence as in Foti et al. (2014). We compare with SVI for the parametric HMM in this many short sequences setting, but still reference Foti et al. (2014) in this case.

Table 1: Methods compared with in our experiments. Top half are batch methods and bottom half are stochastic methods.

Method	**Notation**	**Reference**
Batch variational HMM	HMM	Beal (2003)
Beam sampling HDP-HMM	Beam	Van-Gael et al. (2008)
Direct assignment variational HDP-HMM	HDPHMM-p	Liang et al. (2007)
Direct assignment variational HDP-HMM + mean-field assumption	HDPHMM-p-mf	Liang et al. (2007)
Two-level stick breaking for variational HDP-HMM	HDPHMM-sb	Our method
Stochastic variational HMM	oHMM	Foti et al. (2014)
Direct assignment stochastic variational HDP-HMM	oHDPHMM-p	Johnson and Willsky (2014)
Direct assignment stochastic variational HDP-HMM + split-merge	oHDPHMM-sm	Bryant and Sudderth (2012)
Two-level stick breaking for stochastic variational HDP-HMM	oHDPHMM-sb	Our method

4.1 Artificial data

In this subsection we demonstrate the effectiveness of our variational HDP-HMM on artificial data. We generate discrete training data of length 1,000 from two four-state HMMs with transition matrices A_{pos}, A_{neg} and emission matrix B set to,

$$A_{pos} = \begin{bmatrix} .99 & .01 & 0 & 0 \\ 0 & .99 & .01 & 0 \\ 0 & 0 & .99 & .01 \\ .01 & 0 & 0 & .99 \end{bmatrix}, \quad A_{neg} = \begin{bmatrix} .01 & .99 & 0 & 0 \\ 0 & .01 & .99 & 0 \\ 0 & 0 & .01 & .99 \\ .99 & 0 & 0 & .01 \end{bmatrix},$$

$$B = \frac{1}{3}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

For each scenario we ran 20 experiments with a newly generated sequence and average the results. For the variational HDP-HMM, we truncate the posterior to 10 states and set $a_0 = 1$, $\tau_0 = 1$, $b_0 = 1$. For beam sampling we use the same setting, and we randomly assign each observation to an initial state between 1 and 10. We initialize the variational parameter (see appendix) to $\hat{\theta}_k/100 \sim \mathrm{Dir}(10 \times \mathbf{1})$. For larger state truncations the results were the same, but converged over a longer timescale.

We analyze the convergence of both methods using their respective approximations of the log marginal likelihood on the top of Figure 2. For this small-scale problem, the variational method converges in less than 2.5 seconds (∼0.03s per iteration) while beam sampling converges in a longer time (∼0.02s per iteration). Since beam sampling requires multiple samples after the burn-in phase, it requires significantly more computation time. However, the constrained posterior q distribution is restrictive for variational inference, while beam sampling learns a slightly better model in terms of the log marginal likelihood.

We also compare the accuracy of the number of posterior states recovered by both methods. For variational inference we count the minimum number of occupied states k to cover 99.5% of the data. For beam sampling we record the number of states used per iteration. On the bottom of Figure 2, we can see that it takes more than 2,000 iterations (∼40 seconds) for beam sampling to find the true number of states. For variational inference, this requires less than 100 iterations (∼3 seconds).

4.2 Alice's Adventures in Wonderland

We also compare our HDP-HMM algorithm with the parametric HMM and the direct assignment approach to the HDP-HMM. We recall that the direct assignment learns a point estimate of the top-level truncated DP and represents each second level with a finite Dirichlet distribution.

We consider a sequential text analysis problem in which we collect 12 chapters in "Alice's Adventures in Wonderland" and filter out all the symbols other than the 26 characters and whitespace. We use these symbols as codewords giving a codebook of size of 27. The entire text sequence was encoded as a sequence of these codewords. For each chapter,

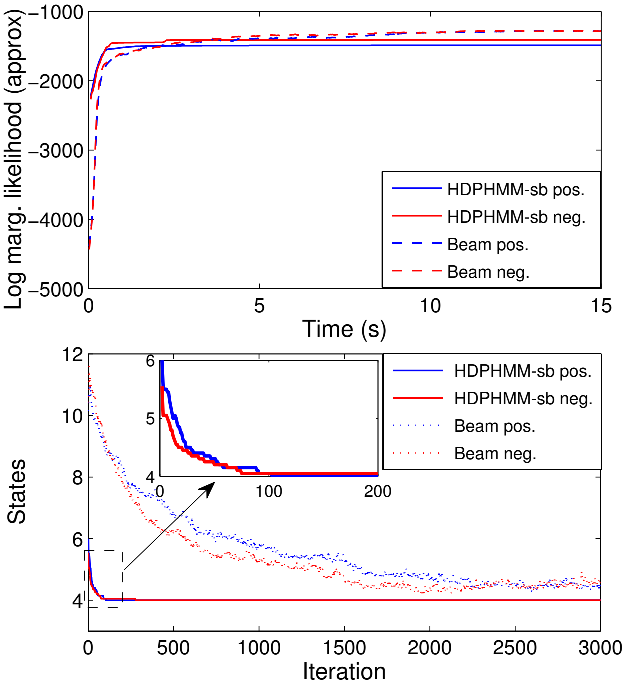

Figure 2: Top: Log likelihood for our variational HDP-HMM and beam sampling. Bottom: Number of posterior states inferred by variational HDP-HMM and beam sampling. Results are averaged over 20 random experiments. "pos" indicates A_{pos} and "neg" indicates A_{neg} was used.

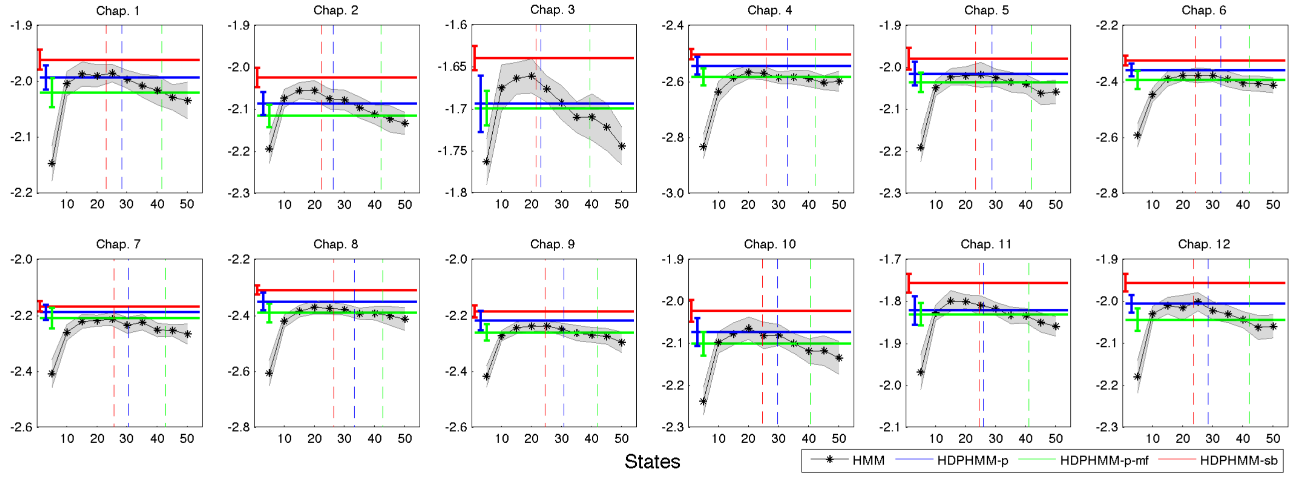

Figure 3: Posterior variational bound for each chapter $(\times 10^4)$. The black plots represents results for HMM with various number of states. The red, blue, green bars separately show the average result for the HDP-HMM, HDP-HMM with direct assignment, and HDP-HMM with direct assignment and fully-factorized mean-field assumption in variational posterior, all with a truncation level of 50. The standard deviation is shown on the left hand side, for each HDP-HMM model. The dashed lines are the number of states used in the posterior of according models averaged over multiple runs.

we truncate the entire sequence into small chunks of size 200. This gave 663 sequences in total and 55 sequences per chapter on average. We pick 537 sequences for training and hold out the remaining 126 for testing, modeling each chapter separately. For the HMM we use variational inference (Beal, 2003) and tried various number of states K ranging from 5 to 50 with the transition Dirichlet parameter set to $5/K$. For the HDP-HMM we truncate the posterior to 50 states and set the first-level scale parameter $a_0 = 5$ and second-level scale parameter $\tau_0 = 3$. For all models we set the emission Dirichlet scale parameter $b_0 = 1/27$. We ran 20 trials for each experiment. For the performance criterion, we use the predictive variational bound on the test data (using sampling in Eq. (5) rather than lower bound).

In Figure 3 we show the predictive variational bounds for the HMM (black lines) as a function of state number, and for our HDP-HMM (red lines). For all experiments we show both the mean and standard deviation. From the figure we can see that when the model grows too large, the HMM may overfit the training data, resulting in a drop in predictive performance. For different chapters, the best performance in the HMM varies, which makes model selection more time consuming. For every chapter, our HDP-HMM out performs HMM in predictive performance and learns roughly the ideal number of states according to the HMM. We mark the average number of occupied states in our HDP-HMM posterior with a red vertical line. The number of states varies from 21.4 (Chap. 3) to 26.4 (Chap. 8), which shows flexibility when the data complexity varies.

We also show the result for the HDP-HMM with direct assignment (blue lines). In general, our algorithm converges to a better solution that uses slightly fewer states, indicating the benefit of our representation. In addition to the direct assignment model, we also made a mean-field assumption in which $q(\mathbf{z}_d) = \prod_i q(z_{di})$ (green lines). This factorization is required in the split-merge stochastic model we compare with (Bryant and Sudderth, 2012) in the next section. We see that a mean-field assumption on $q(\mathbf{z}_d)$ significantly overestimates the number of states.

4.3 Million Song dataset

We also conduct large-scale experiments on discretized audio sequences extracted from the Million Song dataset. We first extract audio features from 371K songs and learn a codebook of size 256 using K-means. We split all the sequences into small chunks of length 50 and learn a single HMM on all sequences.

We compare with the beam-sampled iHMM (Van-Gael et al., 2008), for which we initialized all experiments by randomly assigning observations to one of 500 hidden states, and trained with the same parameter setting as the variational HDP-HMM. For the stochastically-learned parametric HMM models (Foti et al., 2014) we set the transition Dirichlet parameter to $20/K$ and the emission Dirichlet parameter to $b_0 = 0.1$. For the HDP-HMM models, including ours and Johnson and Willsky (2014), we truncated to 500 states and set $a_0 = 20, \tau_0 = 3, b_0 = 0.1$.

In addition, we compare with the stochastic HDP-HMM using a direct assignment and a split-merge strategy during online learning (Bryant and Sudderth, 2012). The split-merge method, originally introduced for stochastic variational HDP, can adaptively create (split) new states or merge old states during each learning iteration in a data-driven manner. In practice we can start with a few states and let the algorithm gradually learn more states. We

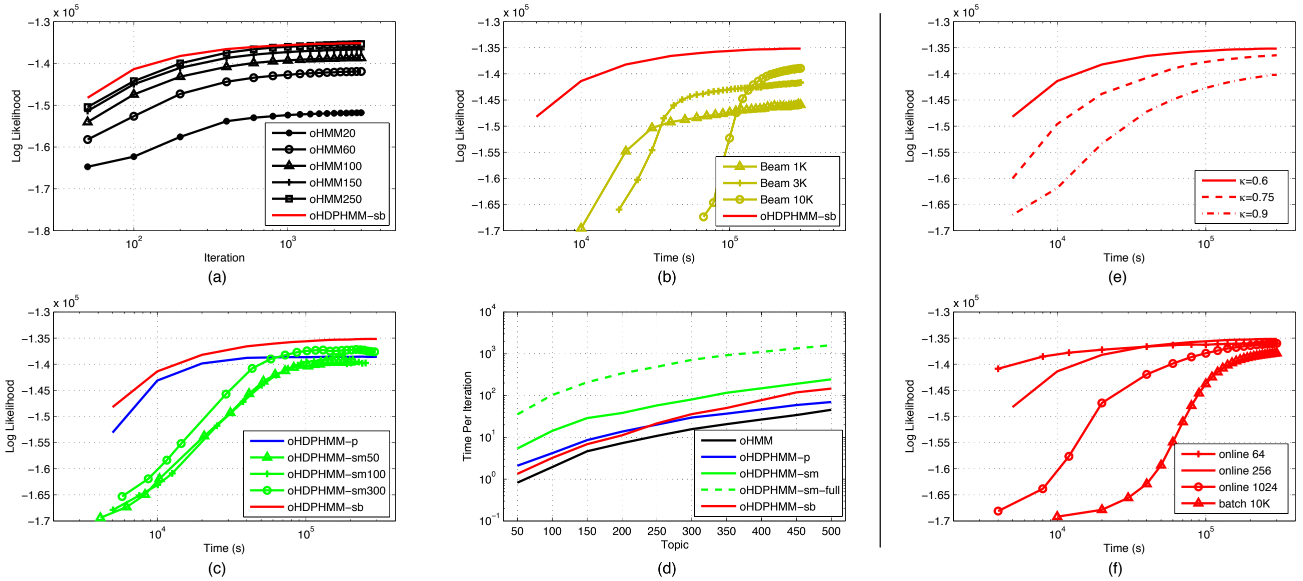

Figure 4: (a). Comparison between stochastic HMM and stochastic HDP-HMM. (b). Comparison between stochastic HDP-HMM and beam sampling. (c). Comparison between the two-level stick-breaking stochastic HDP-HMM with the direct assignment stochastic HDP-HMM with and without split-merge updates. (d). Comparison of time cost per iteration among various methods. (e). Comparison among our stochastic HDP-HMM with various learning rate. (f). Comparison among our stochastic HDP-HMMs with various mini-batch sizes and batch HDP-HMMs trained with 10k sequences.

adapt this method to the HDP-HMM and compare with our method.

For all experiments we ran the algorithms for 3×10^5 seconds (~3.5 days), during which stochastic HDP-HMM can process around 1 million sequences. We also held out 634 sequences for testing and use the predictive log marginal likelihood on this test set as a performance measure. We use $\rho_t = (100 + t)^{-0.6}$ as the learning rate and $|B_t| = 256$ as mini-batch size.

In Figure 4(a) we show the comparison between the parametric stochastic HMM and the nonparametric stochastic HDP-HMM. The predictive log marginal likelihood for stochastic HMM will stop increasing at around 200 to 250 states. The performance for stochastic HDP-HMM is roughly equal to the best of the stochastic HMM.

In Figure 4(b) we also compare the stochastic HDP-HMM with beam sampling. Since there is no stochastic solution for beam sampling, we performed batch inference with various amounts of data and present experiments as a function of time in order to compare the efficiency of the algorithms. As shown, stochastic HDP-HMM outperforms beam sampling because it uses more data. When beam sampling is trained with a limited amount of data (for example, 1K sequences) it will converge more quickly, but the performance will suffer. On the other hand, using too much data for beam sampling will be computationally inefficient. For instance, if we use 10K sequences for training, beam sampling can only draw 70 samples in three days. Beam sampling also did not efficiently infer the number of states. In our experiments, beam sampling will use more than 500 states, while the HDP-HMM occupies around 250 states in its posterior.

In Figure 4(c) we compare our stochastic HDP-HMM with the direct assignment method and the split-merge methods. The predictive likelihood for our method outperforms the direct assignment method. For the split-merge method we do three trials by starting with $\{50, 100, 300\}$ states. All three of these cases converged to around 270 states. However, split-merge is restricted to using the fully factorized mean-field assumption over $q(\mathbf{z}_d)$ as discussed previously. Also, we cannot try all split-merge candidates during each online iteration, otherwise split-merge will be computationally prohibitive.[3] Considering all these factors, the split-merge method performs slightly better than the direct assignment method without split-merge, but still worse than our method.

In Figure 4(d) we show the time per iteration as a function of state number. HMM is the fastest. Direct assignment also performs fast for large numbers of states since the point estimate of the top-level DP significantly reduces the complexity during learning. Our two-level stick-breaking method is slightly slower than the direct assignment method because of the additional of the posterior complexity. The split-merge method is clearly the slow-

[3] The split-merge strategy requires more computation in addition to the direct assignment method by applying a "restricted iteration" for the split part, and a checking over $K(K-1)/2$ potential candidates (K is the number of states) for the merge part. For details, see Bryant and Sudderth (2012).

est, even when we reduce the computation by not checking all possible split-merge moves (green solid). When checking all split-merge options, the algorithm is much slower (green dashed line).

In Figure 4(e) we compare our stochastic HDP-HMM as a function of learning rate. For batch size $|B_t| = 256$, we set $\rho_t = (100+t)^{-\kappa}$ with $\kappa \in \{0.6, 0.75, 0.9\}$. When $\kappa = 0.6$, the learning rate decays more slowly, which gives the best result in our experiments. In Figure 4(f) we compare our stochastic HDP-HMM with various batch sizes $|B_t| \in \{64, 256, 1024\}$ and $\kappa = 0.6$, as well as the batch HDP-HMM trained with 10K sequences. Similar to beam sampling, batch HDP-HMM is inefficient in processing large amounts of data. For stochastic algorithms, choosing small mini-batch sizes will result in fast convergence. On the other hand, choosing larger mini-batch sizes can give better performance, but with slower convergence speed.

5 Conclusion

We have presented a scalable variational inference algorithm for the HDP-HMM. Using a two-level stick-breaking construction, we were able to infer approximate posteriors of all model variables with closed form updates. We compared our algorithm with beam sampling of the HDP-HMM, the parametric HMM, and the direct assignment methods, showing that our inference algorithm is competitive in batch inference settings, and often better in large scale settings using stochastic variational inference. We observe that this algorithm can be applied more generally to extend other latent Markov modeling frameworks to the nonparametric setting (Zhang and Paisley, 2015).

6 Appendix

Updating $q(\mathbf{z}_d)$ and $q(\theta_k)$: For the forward-backward algorithm we define

$$\begin{aligned} \tilde{p}(x_{di}|\theta_k) &= \exp\{\mathbb{E}_q \ln p(x_{di}|\theta_k)\}, & (17) \\ \tilde{A}_{kk'} &= \exp\{\mathbb{E}_q \ln A_{kk'}\}. & (18) \end{aligned}$$

For discrete HMMs we let $q(\theta_k) = \text{Dir}(\hat{\theta}_k)$. In this case we have the same variational expectation as for the parametric model, $\mathbb{E}_q \ln p(x_{di}|\theta_k) = \psi(\hat{\theta}_{k,x_{di}}) - \psi(\sum_j \hat{\theta}_{k,j})$. For Eq. (18) we use the approximation in Sec. 3.1. We recall that for the variational forward-backward algorithm, $\alpha_{di}(k)$ is the variational joint probability of $z_{di} = k$ and the sequence $\mathbf{x}_d$ up to step i, and $\beta_{di}(k)$ is the variational probability of the sequence $\mathbf{x}_d$ after step i conditioned on $z_{di} = k$ (Beal, 2003). We then iterate forward over α_{di} and backward over β_{di} as follows,

$$\alpha_{di}(k) = \tilde{p}(x_{di}|\theta_k) \textstyle\sum_{j=1}^{\infty} \alpha_{d,i-1}(j)\tilde{A}_{jk}, \qquad (19)$$

$$\beta_{di}(k) = \textstyle\sum_{j=1}^{\infty} \tilde{A}_{kj}\tilde{p}(x_{d,i+1}|\theta_j)\beta_{d,i+1}(j). \qquad (20)$$

Having these values, we can make the following update to the marginal of z_{di} for $q(\mathbf{z}_d)$,

$$\gamma_{di}(k) = \frac{\alpha_{di}(k)\beta_{di}(k)}{\sum_j \alpha_{di}(j)\beta_{di}(j)}. \qquad (21)$$

The variational marginal on the state transition $(z_{di}, z_{d,i+1})$ used for the parametric HMM is not used by our algorithm. Given each γ_{di}, we can update $\hat{\theta}_{k,v}$ in $q(\theta_k) = \text{Dir}(\hat{\theta}_k)$ exactly as in the parametric HMM,

$$\hat{\theta}_{k,v} = b_0 + \textstyle\sum_{d,i} \gamma_{di}(k)\mathbb{1}(x_{di} = v). \qquad (22)$$

Updating $q(c)$, $q(\varepsilon)$ and $q(\zeta)$: The update of $q(c)$ uses $\mathbb{E}_q \ln \eta_{0,k'} = \mathbb{E}_q \ln \zeta_{k'} + \sum_{j<k'} \mathbb{E}_q \ln(1-\zeta_j)$. Using the variational distribution $q(\zeta_k) = \text{Beta}(c_k, d_k)$, these expectations are

$$\begin{aligned} \mathbb{E}_q \ln \zeta_k &= \psi(c_k) - \psi(c_k + d_k), \\ \mathbb{E}_q \ln(1-\zeta_j) &= \psi(d_j) - \psi(c_j + d_j). \qquad (23) \end{aligned}$$

Also, $\mathbb{E}_q \ln \eta_{km} = \mathbb{E}_q \ln \varepsilon_{km} + \sum_j \mathbb{E}_q \ln(1-\varepsilon_{kj})$ used elsewhere is similarly calculated as above, only using $q(\varepsilon_{km}) = \text{Beta}(a_{km}, b_{km})$.

To update $q(\varepsilon_{km}) = \text{Beta}(a_{km}, b_{km})$, we have

$$\begin{aligned} a_{km} &= 1 + \textstyle\sum_{d,i} \xi_{di}(k,m), & (24) \\ b_{km} &= \tau_0 + \textstyle\sum_{d,i} \sum_{m'>m} \xi_{di}(k,m'). & (25) \end{aligned}$$

As is evident, given the allocations ξ_{di} (defined in Eq.(8)), this is simply the expected counts used for updating exchangeable stick-breaking mixture models (Blei and Jordan, 2006).

Finally, we have the top-level stick-breaking construction update $q(\zeta_k) = \text{Beta}(c_k, d_k)$. We have

$$\begin{aligned} c_k &= 1 + \textstyle\sum_{k',m} \varphi_{k'm,k}, & (26) \\ d_k &= a_0 + \textstyle\sum_{k',m} \sum_{j>k} \varphi_{k'm,j}. & (27) \end{aligned}$$

where $\varphi_{k'm,k} = \mathbb{E}_q c_{k'm,k}$.

Stochastic inference: For the HDP-HMM, the global q distributions (whose parameter updates are linked to the data size) are on c_{km}, θ_k and ε_{km}. Though ζ_k is also a global parameter, it is conditionally independent of the data given $\mathbf{c}$, and so stochastic inference isn't necessary for this q distribution.

The stochastic updates for $q(\theta_k)$ and $q(\varepsilon_{km})$ are the same as those used by similar models (Hoffman et al., 2013). First, form the scaled closed form updates restricted to B_t, denoted $\hat{\theta}'_{k,v}$, a'_{km} and b'_{km}, and then average with the previous variational parameters,

$$q(\theta_k): \quad \hat{\theta}^{(t)}_{k,v} = (1-\rho_t)\hat{\theta}^{(t-1)}_{k,v} + \rho_t \hat{\theta}'_{k,v}, \qquad (28)$$

$$\begin{aligned} q(\varepsilon_{km}): \quad a^{(t)}_{km} &= (1-\rho_t)a^{(t-1)}_{km} + \rho_t a'_{km}, \\ b^{(t)}_{km} &= (1-\rho_t)b^{(t-1)}_{km} + \rho_t b'_{km}. \qquad (29) \end{aligned}$$

Acknowledgements. This project was supported in part by Lawrence Livermore National Laboratory.

References

Airoldi, E., Blei, D., Erosheva, E., and Fienberg, S., editors (2014). *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC Handbooks of Modern Statistical Methods.

Beal, M. (2003). Variational algorithms for approximate Bayesian inference. *Ph. D. Thesis, University College London*.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Bryant, M. and Sudderth, E. B. (2012). Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*.

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems*.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2008). An HDP-HMM for systems with state persistence. In *International Conference on Machine Learning*.

Hoffman, M., Blei, D., and Bach, F. (2010). Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1005–1031.

Johnson, M. J. and Willsky, A. (2014). Stochastic variational inference for Bayesian time series models. In *International Conference on Machine Learning*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K., editors (1999). *An introduction to variational methods for graphical models*. MIT Press, Cambridge.

Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31:705–741.

Paisley, J. and Carin, L. (2009). Hidden markov models with stick breaking priors. *IEEE Transactions on Signal Processing*, 57:3905–3917.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Van-Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *International Conference on Machine Learning*.

Wang, C., Paisley, J., and Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*.

Zhang, A. and Paisley, J. (2015). Markov mixed membership models. In *International Conference on Machine Learning*.