2008

# An HDP-HMM for Systems With State Persistence

Emily B. Fox
*University of Pennsylvania*

Erik B. Sudderth
*University of California - Berkeley*

Michael I. Jordan
*University of California - Berkeley*

Alan S. Willsky
*Massachusetts Institute of Technology*

# An HDP-HMM for Systems With State Persistence


## Abstract

The hierarchical Dirichlet process hidden Markov model (HDP-HMM) is a flexible, nonparametric model which allows state spaces of unknown size to be learned from data. We demonstrate some limitations of the original HDP-HMM formulation (Teh et al., 2006), and propose a *sticky* extension which allows more robust learning of smoothly varying dynamics. Using DP mixtures, this formulation also allows learning of more complex, multimodal emission distributions. We further develop a sampling algorithm that employs a truncated approximation of the DP to jointly resample the full state sequence, greatly improving mixing rates. Via extensive experiments with synthetic data and the NIST speaker diarization database, we demonstrate the advantages of our sticky extension, and the utility of the HDP-HMM in real-world applications.



## Disciplines

Statistics and Probability

# An HDP-HMM for Systems with State Persistence

**Emily B. Fox** EBFOX@MIT.EDU
Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139

**Erik B. Sudderth** SUDDERTH@EECS.BERKELEY.EDU
Department of EECS, University of California, Berkeley, CA 94720

**Michael I. Jordan** JORDAN@EECS.BERKELEY.EDU
Department of EECS and Department of Statistics, University of California, Berkeley, CA 94720

**Alan S. Willsky** WILLSKY@MIT.EDU
Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139

## Abstract

The hierarchical Dirichlet process hidden Markov model (HDP-HMM) is a flexible, nonparametric model which allows state spaces of unknown size to be learned from data. We demonstrate some limitations of the original HDP-HMM formulation (Teh et al., 2006), and propose a *sticky* extension which allows more robust learning of smoothly varying dynamics. Using DP mixtures, this formulation also allows learning of more complex, multimodal emission distributions. We further develop a sampling algorithm that employs a truncated approximation of the DP to jointly resample the full state sequence, greatly improving mixing rates. Via extensive experiments with synthetic data and the NIST speaker diarization database, we demonstrate the advantages of our sticky extension, and the utility of the HDP-HMM in real-world applications.

## 1. Introduction

Hidden Markov models (HMMs) have been a major success story in many applied fields; they provide core statistical inference procedures in areas as diverse as speech recognition, genomics, structural biology, machine translation, cryptanalysis and finance. Even after four decades of work on HMMs, however, significant problems remain. One lingering issue is the choice of the hidden state space's cardinality. While standard parametric model selection methods can be adapted to the HMM, there is little understanding of the strengths and weaknesses of such methods in this setting.

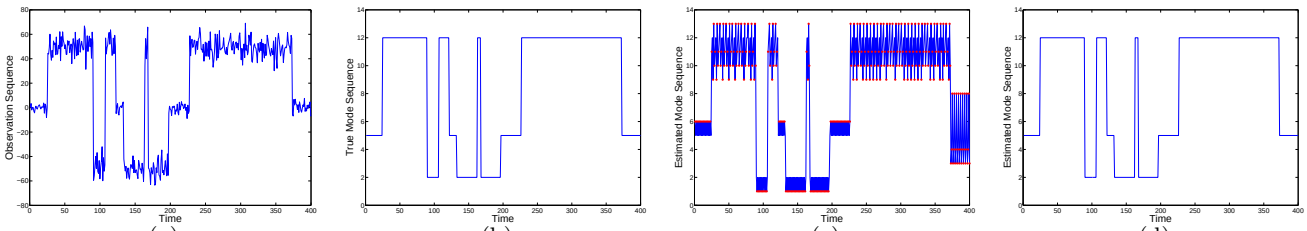
Appearing in *Proceedings of the $25^{th}$ International Conference on Machine Learning*, Helsinki, Finland, 2008. 

Recently, Teh et al. (2006) presented a nonparametric Bayesian approach to HMMs in which a stochastic process, the *hierarchical Dirichlet process* (HDP), defines a prior distribution on transition matrices over countably infinite state spaces. The resulting *HDP-HMM* leads to data–driven learning algorithms which infer posterior distributions over the number of states. This posterior uncertainty can be integrated out when making predictions, effectively averaging over models of varying complexity. The HDP-HMM has shown promise in a variety of applications, including visual scene recognition (Kivinen et al., 2007) and the modeling of genetic recombination (Xing & Sohn, 2007).

One serious limitation of the standard HDP-HMM is that it inadequately models the temporal persistence of states. This problem arises in classical finite HMMs as well, where semi-Markovian models are often proposed as solutions. However, the problem is exacerbated in the nonparametric setting, where the Bayesian bias towards simpler models is insufficient to prevent the HDP-HMM from learning models with unrealistically rapid dynamics, as demonstrated in Fig. 1.

To illustrate the seriousness of this issue, let us consider a challenging application that we revisit in Sec. 5. The problem of *speaker diarization* involves segmenting an audio recording into time intervals associated with individual speakers. This application seems like a natural fit for the HDP-HMM, as the number of true speakers is typically unknown, and may grow as more data is observed. However, this is not a setting in which model averaging is the goal; rather, it is critical to infer the number of speakers as well as the transitions among speakers. As we show in Sec. 5, the HDP-HMM's tendency to rapidly switch among redundant states leads to poor speaker diarization performance.

In contrast, the methods that we develop in this paper yield a state-of-the-art speaker diarization method, as




(a) (b) (c) (d)

*Figure 1.* Sensitivity of the HDP-HMM to within-state variations in the observations. (a) Observation sequence; (b) true state sequence; estimated state sequence after 100 Gibbs iterations for the (c) original and (d) sticky HDP-HMM, with errors indicated in red. Without an extra self–transition bias, the HDP-HMM rapidly transitions among redundant states.

well as a general solution to the problem of state persistence in HDP-HMMs. The approach is easily stated—we simply augment the HDP-HMM to include a parameter for self-transition bias, and place a separate prior on this parameter. The challenge is to consistently execute this idea in a nonparametric Bayesian framework. Earlier papers have also proposed self-transition parameters for HMMs with infinite state spaces (Beal et al., 2002; Xing & Sohn, 2007), but did not formulate general solutions that integrate fully with nonparametric Bayesian inference.

While the HDP-HMM treats the state transition distribution nonparametrically, it is also desirable to allow more flexible, nonparametric emission distributions. In classical applications of HMMs, finite Gaussian mixtures are often used to model multimodal observations. Dirichlet process (DP) mixtures provide an appealing alternative which avoids fixing the number of observation modes. Such emission distributions are not identifiable for the standard HDP-HMM, due to the tendency to rapidly switch between redundant states. With an additional self-transition bias, however, we show that a fully nonparametric HMM leads to effective learning algorithms. In particular, we develop a blocked Gibbs sampler which leverages forward–backward recursions to jointly resample the state and emission assignments for all observations.

In Sec. 2, we begin by presenting background material on the HDP. Sec. 3 then links these nonparametric methods with HMMs, and extends them to account for state persistence. We further augment the model with multimodal emission distributions in Sec. 4, and present results using synthetic data and the NIST speaker diarization database in Sec. 5.

## 2. Background: Dirichlet Processes

A Dirichlet process (DP), denoted by $\mathrm{DP}(\gamma, H)$, is a distribution over countably infinite random measures

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta - \theta_k) \qquad \theta_k \sim H \tag{1}$$

on a parameter space $\Theta$. The weights are sampled via a *stick-breaking construction* (Sethuraman, 1994):

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell) \qquad \beta'_k \sim \mathrm{Beta}(1, \gamma) \tag{2}$$

We denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$.

The DP is commonly used as a prior on the parameters of a mixture model of unknown complexity, resulting in a *DPMM* (see Fig. 2(a)). To generate observations, we choose $\bar{\theta}_i \sim G_0$ and $y_i \sim F(\bar{\theta}_i)$. This sampling process is often described via a discrete variable $z_i \sim \beta$ indicating which component generates $y_i \sim F(\theta_{z_i})$.

The *hierarchical Dirichlet process* (HDP) (Teh et al., 2006) extends the DP to cases in which groups of data are produced by related, yet unique, generative processes. Taking a hierarchical Bayesian approach, the HDP places a global Dirichlet process prior $\mathrm{DP}(\alpha, G_0)$ on $\Theta$, and then draws group specific distributions $G_j \sim \mathrm{DP}(\alpha, G_0)$. Here, the base measure $G_0$ acts as an "average" distribution ($E[G_j] = G_0$) encoding the frequency of each shared, global parameter:

$$G_j(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta(\theta - \tilde{\theta}_{jt}) \qquad \tilde{\pi}_j \sim \mathrm{GEM}(\alpha) \tag{3}$$

$$= \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta - \theta_k) \qquad \pi_j \sim \mathrm{DP}(\alpha, \beta) \tag{4}$$

Because $G_0$ is discrete, multiple $\tilde{\theta}_{jt} \sim G_0$ may take identical values $\theta_k$. Eq. (4) aggregates these probabilities, allowing an observation $y_{ji}$ to be directly associated with the unique global parameters via an indicator random variable $z_{ji} \sim \pi_j$. See Fig. 2(b).

We can alternatively represent this generative process via indicator variables $t_{ji} \sim \tilde{\pi}_j$ and $k_{jt} \sim \beta$, as in Fig. 2(c). The stick-breaking priors on these mixture weights can be analytically marginalized, yielding simple forms for the predictive distributions of assignments. The resulting distribution on partitions is sometimes described using the metaphor of a *Chinese restaurant franchise* (CRF). There are $J$ restaurants (groups), each with infinitely many tables (clusters) at

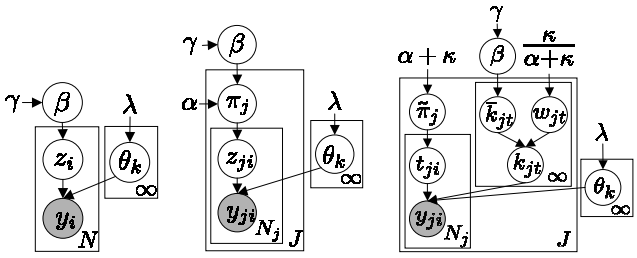


*Figure 2.* (a) DPMM in which $\beta \sim \text{GEM}(\gamma)$, $\theta_k \sim H(\lambda)$, $z_i \sim \beta$, and $y_i \sim f(y \mid \theta_{z_i})$. (b) HDP mixture model with $\beta \sim \text{GEM}(\gamma)$, $\pi_j \sim \text{DP}(\alpha, \beta)$, $\theta_k \sim H(\lambda)$, $z_{ji} \sim \pi_j$, and $y_{ji} \sim f(y \mid \theta_{z_{ji}})$. (c) CRF with loyal customers. Customers $y_{ji}$ sit at table $t_{ji} \sim \tilde{\pi}_j$ which *considers* dish $\bar{k}_{jt} \sim \beta$, but override variables $w_{jt} \sim \text{Ber}(\kappa/\alpha + \kappa)$ can force the *served* dish $k_{jt}$ to be $j$. The original CRF, as described in Sec. 2, has $\kappa = 0$ so that $k_{jt} = \bar{k}_{jt}$.

which customers (observations) sit. Upon entering the $j^{th}$ restaurant, customer $y_{ji}$ sits at currently occupied tables $t_{ji}$ with probability proportional to the number of currently seated customers, or starts a new table $\tilde{t}$ with probability proportional to $\alpha$. Each table chooses a dish (parameter) $\tilde{\theta}_{jt} = \theta_{k_{jt}}$ with probability proportional to the number of other tables in the franchise that ordered that dish, or orders a new dish $\theta_{\tilde{k}}$ with probability proportional to $\gamma$. Observation $y_{ji}$ is then generated by global parameter $\theta_{z_{ji}} = \tilde{\theta}_{jt_{ji}} = \theta_{k_{jt_{ji}}}$.

An alternative, non–constructive characterization of samples $G_0 \sim \text{DP}(\gamma, H)$ from a Dirichlet process states that for every finite partition $\{A_1, \ldots, A_K\}$ of $\Theta$,

$$(G_0(A_1), \ldots, G_0(A_K)) \sim \text{Dir}(\gamma H(A_1), \ldots, \gamma H(A_K)). \quad (5)$$

Using this expression, it can be shown that the following finite, hierarchical mixture model converges in distribution to the HDP as $L \to \infty$ (Ishwaran & Zarepour, 2002; Teh et al., 2006):

$$\begin{aligned} \beta &\sim \text{Dir}(\gamma/L, \ldots, \gamma/L) \\ \pi_j &\sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_L). \end{aligned} \quad (6)$$

Later sections use this *weak limit* approximation to develop efficient, blocked sampling algorithms.

## 3. The Sticky HDP-HMM

The HDP can be used to develop an HMM with an unknown, potentially infinite state space (Teh et al., 2006). For this HDP-HMM, each HDP group-specific distribution, $\pi_j$, is a state-specific transition distribution and, due to the infinite state space, there are infinitely many groups. Let $z_t$ denote the state of the Markov chain at time $t$. For Markov chains $z_t \sim \pi_{z_{t-1}}$, so that $z_{t-1}$ indexes the group to which $y_t$ is assigned. The current HMM state $z_t$ then indexes the parameter $\theta_{z_t}$ used to generate observation $y_t$ (see Fig. 3).

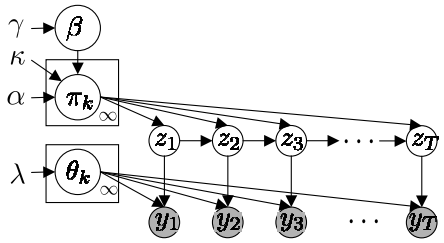


*Figure 3.* Graph of the sticky HDP-HMM. The state evolves as $z_{t+1} \sim \pi_{z_t}$, where $\pi_k \sim \text{DP}(\alpha + \kappa, (\alpha\beta + \kappa\delta_k)/(\alpha + \kappa))$ and $\beta \sim \text{GEM}(\gamma)$, and observations are generated as $y_t \sim F(\theta_{z_t})$. The original HDP-HMM has $\kappa = 0$.

By sampling $\pi_j \sim \text{DP}(\alpha, \beta)$, the HDP prior encourages states to have similar transition distributions ($E[\pi_{jk}] = \beta_k$). However, it does not differentiate self–transitions from moves between states. When modeling systems with state persistence, the flexible nature of the HDP-HMM prior allows for state sequences with unrealistically fast dynamics to have large posterior probability. For example, with Gaussian emissions, as in Fig. 1, a good explanation of the data is to divide an observation block into two small–variance states with slightly different means, and then rapidly switch between them (see Fig. 1). In such cases, many models with redundant states may have large posterior probability, thus impeding our ability to identify a single dynamical model which best explains the observations. The problem is compounded by the fact that once this alternating pattern has been instantiated by the sampler, its persistence is then reinforced by the properties of the Chinese restaurant franchise, thus slowing mixing rates. Furthermore, when observations are high-dimensional, this fragmentation of data into redundant states may reduce predictive performance. In many applications, one would thus like to be able to incorporate prior knowledge that slow, smoothly varying dynamics are more likely.

To address these issues, we propose to instead sample transition distributions $\pi_j$ as follows:

$$\pi_j \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}\right). \quad (7)$$

Here, $(\alpha\beta + \kappa\delta_j)$ indicates that an amount $\kappa > 0$ is added to the $j^{th}$ component of $\alpha\beta$. The measure of $\pi_j$ over a finite partition $(Z_1, \ldots, Z_K)$ of the positive integers $\mathbb{Z}_+$, as described by Eq. (5), adds an amount $\kappa$ only to the arbitrarily small partition containing $j$, corresponding to a self-transition. When $\kappa = 0$ the original HDP-HMM is recovered. Because positive $\kappa$ values increase the prior probability $E[\pi_{jj}]$ of self–transitions, we refer to this extension as the *sticky* HDP-HMM.

In some ways, this $\kappa$ parameter is reminiscent of the infinite HMM's self-transition bias (Beal et al., 2002).

However, that paper relied on a heuristic, approximate Gibbs sampler. The full connection between the infinite HMM and an underlying nonparametric Bayesian prior, as well as the development of a globally consistent inference algorithm, was made in Teh et al. (2006), but without a treatment of a self-transition parameter.

### 3.1. A CRF with Loyal Customers

We further abuse the Chinese restaurant metaphor by extending it to the sticky HDP-HMM, where our franchise now has restaurants with loyal customers. Each restaurant has a specialty dish with the same index as that of the restaurant. Although this dish is served elsewhere, it is more popular in the dish's namesake restaurant. We see this increased popularity from the fact that a table's dish is now drawn as

$$k_{jt} \sim \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}. \tag{8}$$

We will refer to $z_t$ as the parent and $z_{t+1}$ as the child. The parent enters a restaurant $j$ determined by its parent (the grandparent), $z_{t-1} = j$. We assume there is a bijective mapping of indices $f : t \to ji$. The parent then chooses a table $t_{ji} \sim \tilde{\pi}_j$ and that table is served a dish indexed by $k_{jt_{ji}}$. Noting that $z_t = z_{ji} = k_{jt_{ji}}$, the increased popularity of the house specialty dish implies that children are more likely to eat in the same restaurant as their parent and, in turn, more likely to eat the restaurant's specialty dish. This develops family loyalty to a given restaurant in the franchise. However, if the parent chooses a dish other than the house specialty, the child will then go to the restaurant where this dish is the specialty and will in turn be more likely to eat this dish, too. One might say that for the sticky HDP-HMM, children have similar tastebuds to their parents and will always go the restaurant that prepares their parent's dish best. Often, this keeps many generations eating in the same restaurant.

The inference algorithm is simplified if we introduce a set of auxiliary random variables $\bar{k}_{jt}$ and $w_{jt}$ as follows:

$$\bar{k}_{jt} \sim \beta, \qquad w_{jt} \sim \text{Ber}\left(\frac{\kappa}{\alpha+\kappa}\right), \qquad k_{jt} = \begin{cases} \bar{k}_{jt}, & w_{jt} = 0; \\ j, & w_{jt} = 1, \end{cases} \tag{9}$$

where $\text{Ber}(p)$ represents the Bernoulli distribution. The table first chooses a dish $\bar{k}_{jt}$ without taking the restaurant's specialty into consideration (i.e., the original CRF.) With some probability, this *considered* dish is overridden (perhaps by a waiter's suggestion) and the table is served the specialty dish $j$. Thus, $k_{jt}$ represents the *served* dish. We refer to $w_{jt}$ as the *override* variable. For the original HDP-HMM, when $\kappa = 0$, the considered dish is always the served dish since $w_{jt} = 0$ for all tables. See Fig. 2(c).

### 3.2. Sampling via Direct Assignments

In this section we describe a modified version of the direct assignment Rao-Blackwellized Gibbs sampler of Teh et al. (2006) which circumvents the complicated bookkeeping of the CRF by sampling indicator random variables directly. Throughout this section, we refer to the variables in the graph of Fig. 3. For this sampler, a set of auxiliary variables $m_{jk}$, $\bar{m}_{jk}$, and $w_{jt}$ must be added (as illustrated in Fig. 2(c)).

**Sampling** $z_t$ The posterior distribution factors as:

$$p(z_t = k \mid z_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa, \lambda) \propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda). \tag{10}$$

The properties of the Dirichlet process dictate that on the finite partition $\{1, \ldots, K, \tilde{k}\}$ we have the following form for the group-specific transition distributions:

$$\pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_K, \alpha\beta_{\tilde{k}}). \tag{11}$$

We use the above definition of $\pi_j$ and the Dirichlet distribution's conjugacy to the multinomial observations $z_t$ to marginalize $\pi_j$ and derive the following conditional distribution over the states assignments:

$$p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \propto (\alpha\beta_k + n^{-t}_{z_{t-1}k} + \kappa\delta(z_{t-1}, k)) \left( \frac{\alpha\beta_{z_{t+1}} + n^{-t}_{kz_{t+1}} + \kappa\delta(k, z_{t+1}) + \delta(z_{t-1}, k)\delta(k, z_{t+1})}{\alpha + n^{-t}_{k\cdot} + \kappa + \delta(z_{t-1}, k)} \right). \tag{12}$$

This formula is more complex than that of the standard HDP sampler due to potential dependencies in the marginalization of $\pi_{z_{t-1}}$ and $\pi_{z_t}$. For a detailed derivation, see Fox et al. (2007). The notation $n_{jk}$ represents the number of Markov chain transitions from state $j$ to $k$, $n_{j\cdot} = \sum_k n_{jk}$, and $n^{-t}_{jk}$ the number of transitions from state $j$ to $k$ not counting the transition $z_{t-1}$ to $z_t$ or $z_t$ to $z_{t+1}$. Intuitively, this expression chooses a state $k$ with probability depending on how many times we have seen other $z_{t-1}$ to $k$ and $k$ to $z_{t+1}$ transitions. Note that there is a dependency on whether either or both of these transitions correspond to a self-transition, which is strongest when $\kappa > 0$.

As in Teh et al. (2006), by placing a conjugate prior on the parameter space, there is a closed analytic form for the likelihood component $p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda)$.

**Sampling** $\beta$ Assume there are currently $\bar{K}$ unique dishes being *considered* and take a finite partition $\{\theta_1, \theta_2, \ldots, \theta_{\bar{K}}, \theta_{\tilde{k}}\}$ of $\Theta$, where $\theta_{\tilde{k}} = \Theta \setminus \bigcup_{k=1}^{\bar{K}} \{\theta_k\}$. Since $\tilde{\theta}_{jt} \sim G_0$ and $\bar{m}_{\cdot k}$ tables are considering dish $\theta_k$, the properties of the Dirichlet distribution dictate:

$$p((\beta_1, \ldots, \beta_{\bar{K}}, \beta_{\tilde{k}}) \mid \bar{\boldsymbol{k}}, \gamma) \propto \text{Dir}(\bar{m}_{\cdot 1}, \ldots, \bar{m}_{\cdot \bar{K}}, \gamma). \tag{13}$$

From the above, we see that $\{\bar{m}_{\cdot k}\}_{k=1}^{\bar{K}}$ is a set of sufficient statistics for resampling $\beta$ on this partition.

However, this requires sampling two additional variables, $m_{jk}$ and $w_{jt}$, corresponding to the number of tables in restaurant $j$ *served* dish $k$ and the corresponding overwrite variables. We jointly sample from

$$p(\boldsymbol{m}, \boldsymbol{w}, \bar{\boldsymbol{m}} \mid z_{1:T}, \beta, \alpha, \kappa) = p(\bar{\boldsymbol{m}} \mid \boldsymbol{m}, \boldsymbol{w}, z_{1:T}, \beta, \alpha, \kappa) \\ p(\boldsymbol{w} \mid \boldsymbol{m}, z_{1:T}, \beta, \alpha, \kappa) p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa). \quad (14)$$

We start by examining $p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa)$. Having the state index assignments $z_{1:T}$ effectively partitions the data (customers) into both restaurants and dishes, though the table assignments are unknown since multiple tables can be served the same dish. Thus, sampling $m_{jk}$ is in effect equivalent to sampling table assignments for each customer *after* knowing the dish assignment. This conditional distribution is given by:

$$p(t_{ji} = t \mid k_{jt} = k, \boldsymbol{t}^{-ji}, \boldsymbol{k}^{-jt}, y_{1:T}, \beta, \alpha, \kappa) \\ \propto \begin{cases} \tilde{n}_{jt}^{-ji}, & t \in \{1, \ldots, T_j\}; \\ \alpha\beta_k + \kappa\delta(k, j), & t = \tilde{t}_j, \end{cases} \quad (15)$$

where $\tilde{n}_{jt}^{-ji}$ is the number of customers at table $t$ in restaurant $j$, not counting $y_{ji}$. The form of Eq. (15) implies that a customer's table assignment conditioned on a dish assignment $k$ follows a DP with concentration parameter $\alpha\beta_k + \kappa\delta(k, j)$ and may be sampled by simulating the associated Chinese restaurant process.

We now derive the conditional distribution for the override variables $w_{jt}$. The table counts provide that $m_{jk}$ tables are serving dish $k$ in restaurant $j$. If $k \neq j$, we automatically have $m_{jk}$ tables with $w_{jt} = 0$ since the served dish is not the house specialty. Otherwise,

$$p(w_{jt} \mid k_{jt} = j, \beta, \rho) \propto \begin{cases} \beta_j(1 - \rho), & w_{jt} = 0; \\ \rho, & w_{jt} = 1, \end{cases} \quad (16)$$

where $\rho = \frac{\kappa}{\alpha + \kappa}$ is the prior probability that $w_{jt} = 1$. Observing served dish $k_{jt} = j$ makes it more likely that the considered dish $\bar{k}_{jt}$ was overridden than the prior suggests. We draw $m_{jj}$ samples of $w_{jt}$ from Eq. (16).

Given $m_{jk}$ for all $j$ and $k$ and $w_{jt}$ for each of these instantiated tables, we can now deterministically compute $\bar{m}_{jk}$. Any table that was overridden is an uninformative observation for the posterior of $\bar{m}_{jk}$ so that

$$\bar{m}_{jk} = \begin{cases} m_{jk}, & j \neq k; \\ m_{jj} - w_{j\cdot}, & j = k. \end{cases} \quad (17)$$

**Sampling Hyperparameters** Rather than fixing the sticky HDP-HMM's hyperparameters, we place vague gamma priors on $\gamma$ and $(\alpha + \kappa)$, and a beta prior on $\kappa/(\alpha+\kappa)$. As detailed in Fox et al. (2007), the auxiliary variables introduced in the preceding section then allow tractable resampling of these hyperparameters. This allows the number of occupied states, and the degree of self–transition bias, to be strongly influenced by the statistics of observed data, as desired.

### 3.3. Blocked Sampling of State Sequences

The HDP-HMM direct assignment sampler can exhibit slow mixing rates since global state sequence changes are forced to occur coordinate by coordinate. This is explored in Scott (2002) for the finite HMM. Although the sticky HDP-HMM reduces the posterior uncertainty caused by fast state-switching explanations of the data, the self-transition bias can cause two continuous and temporally separated sets of observations of a given state to be grouped into two states. If this occurs, the high probability of self-transition makes it challenging for the sequential sampler to group those two examples into a single state.

A variant of the HMM forward-backward procedure (Rabiner, 1989) allows us to harness the Markov structure and jointly sample the state sequence $z_{1:T}$ given the observations $y_{1:T}$, transitions probabilities $\pi_j$, and model parameters $\theta_k$. To take advantage of this procedure, we now must sample the previously marginalized transition distributions and model parameters. In practice, this requires approximating the theoretically countably infinite transition distributions. One approach is the degree $L$ *weak limit approximation* to the DP (Ishwaran & Zarepour, 2002),

$$\mathrm{GEM}_L(\alpha) \triangleq \mathrm{Dir}(\alpha/L, \ldots, \alpha/L), \quad (18)$$

where $L$ is a number that exceeds the total number of expected HMM states. This approximation encourages the learning of models with fewer than $L$ components while allowing the generation of new components, upper bounded by $L$, as new data are observed.

The posterior distributions of $\beta$ and $\pi_j$ are given by:

$$\beta \sim \mathrm{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \ldots, \gamma/L + \bar{m}_{\cdot L}) \quad (19)$$
$$\pi_j \sim \mathrm{Dir}(\alpha\beta_1 + n_{j1}, \ldots, \alpha\beta_j + \kappa + n_{jj}, \ldots, \alpha\beta_L + n_{jL}).$$

Depending on the form of the emission distribution and base measure on the parameter space $\Theta$, we sample parameters for each of the currently instantiated states from the updated posterior distribution:

$$\theta_j \sim p(\theta \mid \{y_t \mid z_t = j\}, \lambda). \quad (20)$$

Now that we are sampling $\theta_j$ directly, we can use a non-conjugate base measure.

We block sample $z_{1:T}$ by first computing backward messages $m_{t,t-1}(z_{t-1}) \propto p(y_{t:T} | z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta})$ and then recursively sampling each $z_t$ conditioned on $z_{t-1}$ from

$$p(z_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \\ p(z_t \mid \pi_{z_{t-1}}) p(y_t \mid \theta_{z_t}) m_{t+1,t}(z_t). \quad (21)$$

A similar sampler has been used for learning HDP hidden Markov trees (Kivinen et al., 2007). However, this work did not consider the complications introduced by multimodal emissions, as we explore next.

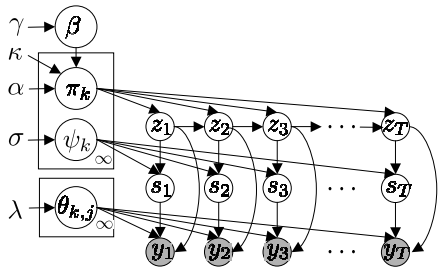


*Figure 4.* Sticky HDP-HMM with DP emissions, where $s_t$ indexes the state-specific mixture component generating observation $y_t$. The DP prior dictates that $s_t \sim \psi_{z_t}$ for $\psi_k \sim \text{GEM}(\sigma)$. The $j^{th}$ Gaussian component of the $k^{th}$ mixture density is parameterized by $\theta_{k,j}$ so $y_t \sim F(\theta_{z_t,s_t})$.

## 4. Multimodal Emission Distributions

For many application domains, the data associated with each hidden state may have a complex, multimodal distribution. We propose to approximate such emission distributions nonparametrically, using an infinite DP mixture of Gaussians. This formulation is related to the nested DP (Rodriguez et al., 2006). The bias towards self-transitions allow us to distinguish between the underlying HDP-HMM states. If the model were free to both rapidly switch between HDP-HMM states and associate multiple Gaussians per state, there would be considerable posterior uncertainty. Thus, it is only with the sticky HDP-HMM that we can effectively learn such models.

We augment the HDP-HMM state $z_t$ with a term $s_t$ indexing the mixture component of the $z_t^{th}$ emission density. For each HDP-HMM state, there is a unique stick-breaking distribution $\psi_k \sim \text{GEM}(\sigma)$ defining the mixture weights of the $k^{th}$ emission density so that $s_t \sim \psi_{z_t}$. The observation $y_t$ is generated by the Gaussian component with parameter $\theta_{z_t,s_t}$. See Fig. 4.

To implement blocked resampling of $(z_{1:T}, s_{1:T})$, we use weak limit approximations to both the HDP-HMM and Dirichlet process emissions, approximated to levels $L$ and $L'$, respectively. The posterior distributions of $\beta$ and $\pi_k$ remain unchanged; that of $\psi_k$ is given by:

$$\psi_k \sim \text{Dir}(\sigma/L' + n'_{k1}, \ldots, \sigma/L' + n'_{kL'}), \qquad (22)$$

where $n'_{kl}$ are the number of observations assigned to the $l^{th}$ mixture component of the $k^{th}$ HMM state. The posterior distribution for each Gaussian's mean and covariance, $\theta_{k,j}$, is determined by the observations assigned to this component, namely,

$$\theta_{k,j} \sim p(\theta \mid \{y_t \mid (z_t = k, s_t = j)\}, \lambda). \qquad (23)$$

The augmented state $(z_t, s_t)$ is sampled from

$$p(z_t, s_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}) \propto$$
$$p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t}) p(y_t \mid \theta_{z_t,s_t}) m_{t+1,t}(z_t). \qquad (24)$$

Since the Markov structure is only on the $z_t$ component of the augmented state, the backward message $m_{t,t-1}(z_{t-1})$ from $(z_t, s_t)$ to $(z_{t-1}, s_{t-1})$ is solely a function of $z_{t-1}$. These messages are given by:

$$m_{t,t-1}(z_{t-1}) \propto \sum_{z_t} \sum_{s_t} p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t})$$
$$p(y_t \mid \theta_{z_t,s_t}) m_{t+1,t}(z_t). \qquad (25)$$

## 5. Results

**Synthetic Data** We generated test data from a three-state Gaussian emission HMM with: 0.97 probability of self-transition; means 50, 0, and -50; and variances 50, 10, and 50 (see Fig. 1(a).) For the blocked sampler, we used a truncation level of $L = 15$.

Fig. 5 shows the clear advantage of considering a sticky HDP-HMM with blocked sampling. The Hamming distance error is calculated by greedily mapping the indices of the estimated state sequence to those maximizing overlap with the true sequence. The apparent slow convergence of the sticky HDP-HMM direct assignment sampler (Fig. 5(b)) can be attributed to the sampler splitting temporally separated segments of a true state into multiple, redundant states. Although not depicted due to space constraints, both sticky HDP-HMM samplers result in estimated models with significantly larger likelihoods of the true state sequence than those of the original HDP-HMM.

To test the model of Sec. 4, we generated data from a two-state HMM, where each state had a two-Gaussian mixture emission distribution with equally weighted components defined by means $(0, 10)$ and $(-7, 7)$, and variances of 10. The probability of self-transition was set to 0.98. The resulting observation and true state sequences are shown in Fig. 6(a) and (b).

Fig. 6(e)-(h) compares the performance of the sticky and original HDP-HMM with single and infinite Gaussian mixture emissions. All results are for the blocked sampler with truncation levels $L = L' = 15$. Intuitively, when constrained to single Gaussian emissions, the best explanation of the data is to associate each true mixture component with a separate state and then quickly switch between these states, resulting in the large Hamming distances of Fig. 6(g)-(h). Although not the desired effect in this scenario, this behavior, as depicted in Fig. 6(c), demonstrates the flexibility of the sticky HDP-HMM: if the best explanation of the data according to the model is fast state-switching, the sticky HDP-HMM still allows for this by learning a small bias towards self-transitions. The sticky HDP-HMM occasionally has more accurate state sequence estimates by grouping a true state's Gaussian mixture components into a single Gaussian with large variance. By far the best performance is

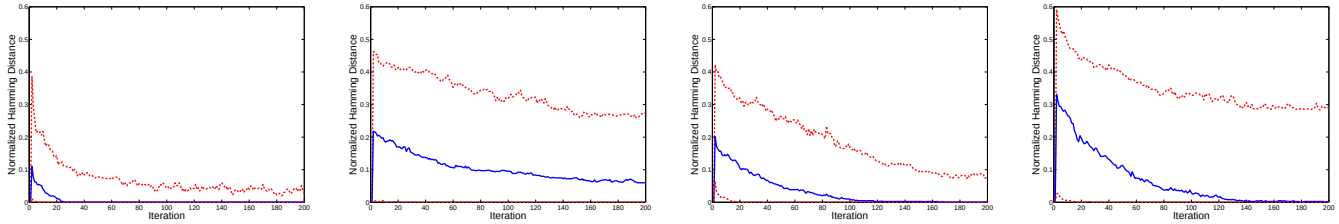



*Figure 5.* Hamming distance between true and estimated state sequences over 100 iterations for the sticky HDP-HMM (a) blocked and (b) direct assignment samplers and the original HDP-HMM (c) blocked and (d) direct assignment samplers. These plots show the median (solid blue) and $10^{th}$ and $90^{th}$ quantiles (dashed red) from 200 initializations.

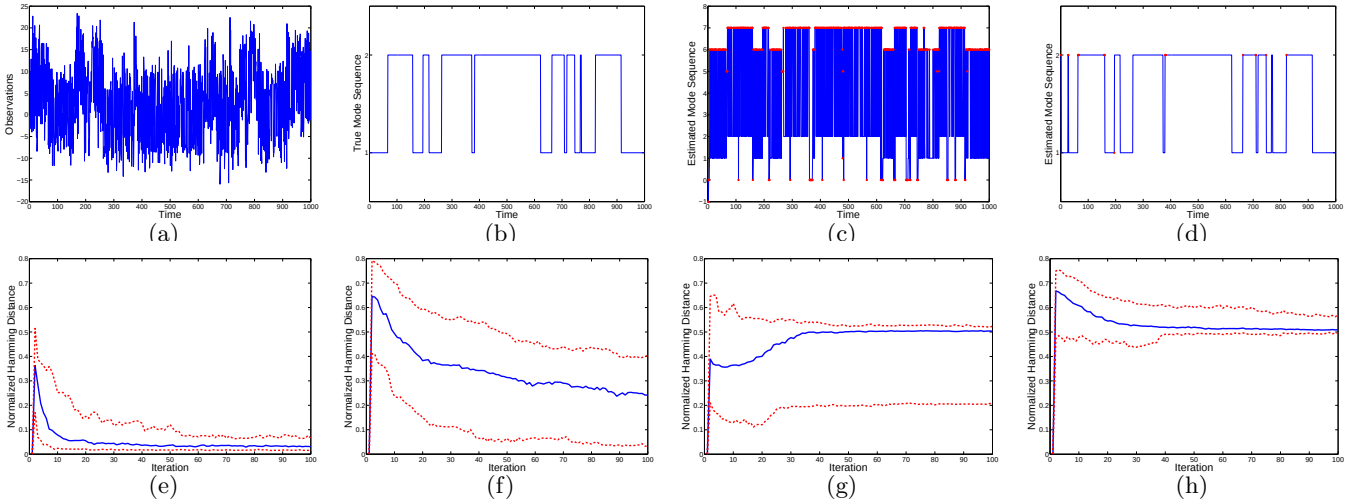
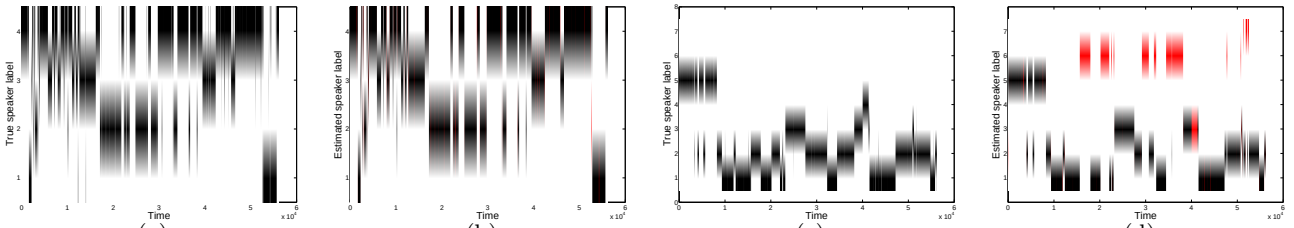



*Figure 6.* Performance of inference on data generated by an HMM with Gaussian mixture emissions. (a) Observation sequence; (b) true HMM state sequence; estimated HMM state sequence using the sticky HDP-HMM model with (c) single and (d) infinite Gaussian mixture emissions. Errors are indicated by red markers. The bottom row contains Hamming distance plots, as in Fig. 5, for infinite Gaussian mixture emissions and the (e) sticky HDP-HMM and (f) original HDP-HMM, and single Gaussian emissions for the (g) sticky HDP-HMM and (h) original HDP-HMM.

achieved by the sticky HDP-HMM with infinite Gaussian mixture emissions (see Fig. 6(e) and (d)); comparing to Fig. 6(f), we see that the gain can be attributed to modeling rather than just improved mixing rates.

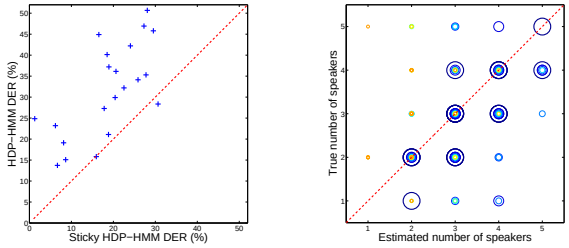
**Speaker Diarization Data** The *speaker diarization* task involves segmenting an audio recording into speaker-homogeneous regions, while simultaneously identifying the number of speakers. We tested the utility of the sticky HDP-HMM for this task on the data distributed by NIST as part of the Rich Transcription 2004-2007 meeting recognition evaluations (NIST, 2007). We use the first 19 Mel Frequency Cepstral Coefficients (MFCCs), computed over a 30ms window every 10ms, as our feature vector. When working with this dataset, we discovered that: (1) the high frequency content of these features contained little discriminative information, and (2) without a minimum speaker duration, the sticky HDP-HMM learned within speaker dynamics in addition to global speaker changes. To jointly address these issues, we instead model feature averages computed over 250ms, non–overlapping blocks. A minimum speaker duration of 500ms is set by associating two average features with each hidden state. We also tie the covariances of within–state mixture components. We found single–Gaussian emission distributions to be less effective.

For each of 21 meetings, we compare 10 initializations of the original and sticky HDP-HMM blocked samplers. In Fig. 8(a), we report the official NIST diarization error rate (DER) of the run with the largest observation sequence likelihood, given parameters estimated at the 1000th Gibbs iteration. The sticky HDP-HMM's temporal smoothing provides substantial performance gains. Fig 8(b) plots the estimated versus true number of speakers who talk for more than 10% of the meeting time, and shows our model's ability to adapt to a varying number of speakers. As a further comparison, the ICSI team's algorithm (Wooters & Huijbregts, 2007), by far the best performer at the 2007 competition, has an overall DER of 18.37%, simi-




(a) (b) (c) (d)

*Figure 7.* True state sequences for meetings (a) AMI_20041210-1052 and (c) VT_20050304-1300, with the corresponding most likely state estimates shown in (b) and (d), respectively, with incorrect labels shown in red.




(a) (b)

*Figure 8.* For the 21 meeting database: (a) plot of sticky vs. original HDP-HMM most likely sequence DER; and (b) plot of true vs. estimated number of speakers for samples drawn from 10 random initializations of each meeting (larger circles have higher likelihood).

lar to our 19.04%. Our best and worst DER are 1.26% and 31.42%, respectively, compared to their 4.39% and 32.23%. We use the same non-speech pre-processing, so that the differences are due to changes in the identified speakers. As depicted in Fig. 7, a significant proportion of our errors can be attributed to splitting or merging speakers. The ICSI team's algorithm uses agglomerative clustering, and requires significant tuning of parameters on representative training data. In contrast, our hyperparameters are automatically set meeting-by-meeting, so that each component's expected mean and covariance are that of the entire feature sequence. Note that the selected runs plotted in Fig. 8 are not necessarily those with the smallest DER. For example, the run depicted in Fig. 7(d) had 24.06% DER, while another run on the same meeting had 4.37% (versus ICSI's 22.00%.) There is inherent posterior uncertainty in this task, and our sampler has the advantage of giving several interpretations. When considering the best per-meeting DER for the five most likely samples, our overall DER drops to 15.14%; we hope to explore automated ways of combining multiple samples in future work. Regardless, our results demonstrate that the sticky HDP-HMM provides an elegant and empirically effective speaker diarization method.

## 6. Discussion

We have demonstrated the considerable benefits of an extended HDP-HMM in which a separate parameter captures state persistence. We have also shown that this sticky HDP-HMM allows a fully nonparametric treatment of multimodal emissions, disambiguated by its bias towards self-transitions, and presented efficient sampling techniques with mixing rates that improve on the state-of-the-art. Results on synthetic data, and a challenging speaker diarization task, clearly demonstrate the practical importance of our extensions.

## Acknowledgments

We thank O. Vinyals, G. Friedland, and N. Morgan for helpful discussions about the NIST dataset. This research was supported in part by DARPA contract NBCHD030010, and MURIs funded through ARO Grant W911NF-06-1-0076 and AFOSR Grant FA9550-06-1-0324. E.B.F. was partially funded by an NDSEG fellowship.

## References

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2002). The infinite hidden Markov model. *NIPS* (pp. 577–584).

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2007). A tempered HDP-HMM for systems with state persistence. *MIT LIDS, TR #2777.*

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum–representations for the Dirichlet process. *Can. J. Stat.*, *30*, 269–283.

Kivinen, J. J., Sudderth, E. B., & Jordan, M. I. (2007). Learning multiscale representations of natural scenes using Dirichlet processes. *ICCV* (pp. 1–8).

NIST (2007). Rich transcriptions database. *http://www.nist.gov/speech/tests/rt/.*

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, *77*, 257–286.

Rodriguez, A., Dunson, D., & Gelfand, A. (2006). The nested Dirichlet process. *Duke ISDS, TR #06-19.*

Scott, S. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Stat. Assoc.*, *97*, 337–351.

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

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Stat. Assoc.*, *101*, 1566–1581.

Wooters, C., & Huijbregts, M. (2007). The ICSI RT07s speaker diarization system. *To appear in LNCS.*

Xing, E., & Sohn, K.-A. (2007). Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayes. Analysis*, *2*, 501–528.