Hierarchical Nested Infinite Hidden Markov Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a nonparametric Bayesian prior that models hierarchical state clustering in hidden Markov models so that states in the same cluster have similar transition distributions. The prior is a combination of the hierarchical Dirichlet process (HDP) and the nested Dirichlet process (nDP), where the HDP models the similarity of a distribution over destination states and the nDP models the similarity of distribution over source states. The resulting model, which we named the hierarchical nested infinite hidden Markov model (HN-iHMM), is presented with a sampling-based inference algorithm. The flexibility of the model is demonstrated through evaluation on an artificial sequence and a natural language text.

1 Introduction

The infinite hidden Markov model (iHMM) [2] (also known as hierarchical Dirichlet process hidden Markov model [12]) is one of the most successful nonparametric Bayesian models for sequences. In iHMM, the number N of hidden states is effectively integrated out, and the plausible range of N is automatically inferred from the posterior distribution of the iHMM. The hierarchical Dirichlet process (HDP) [12] is used to represent the set of transition distributions; the upper level represents the distribution of states in the given HMM, and each DP in the lower level corresponds to a transition distribution from a previous state.

However, the general assumption made in the iHMM is sometimes too strong to model realistic data. The model assumes the same prior for every possible state; in other words, every state is considered to be equally different. It is often the case that a state is similar to several other states; for example, application of an HMM to part-of-speech (POS) tagging [8], associating each state to a POS tag of a word, involves such similarities (e.g., intransitive and transitive verbs). In the context of machine learning, neglecting such similarity causes the data sparseness problem, i.e., states with few occurrences, which are likely to exist due to the property of the Dirichlet process, cannot be inferred reliably from a vague prior. Thus, it has to be important to give a better prior to the iHMM by introducing *clusters* of states.

In this paper, we consider a Bayesian prior, in which states are clustered by transition probability, i.e., states in a cluster have similar transition distributions. Then, a state with few occurrences can be helped by the prior transition probability on the basis of other states in the same cluster. As a result, such a prior is expected to fit better to data with local similarities between hidden states. Moreover, such a model can be applied to unsupervised learning for the hierarchical cluster structure from a given sequence.

We propose a nonparametric Bayesian model that represents the hierarchical structure of hidden states in hidden Markov models. We call the model as the hierarchical nested infinite hidden Markov model (HN-iHMM), an extension of the iHMM with the combination of two nonparametric Bayesian techniques: HDPs (with more than two levels) for representing the prior depending on cluster structure, and nested Dirichlet processes (nDP) [9] for representing the clustered distribution of states.

We tested the model with an inference algorithm based on Restricted Collapsed Draws (RCD) sam-

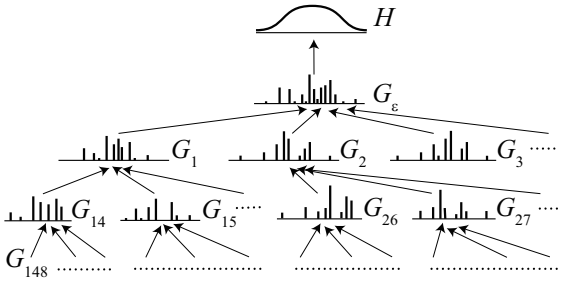

Figure 1: An example structure of a hierarchical DP. Elements of indices are denoted as digits. Each arrow indicates the relationship between a DP and its base measure.

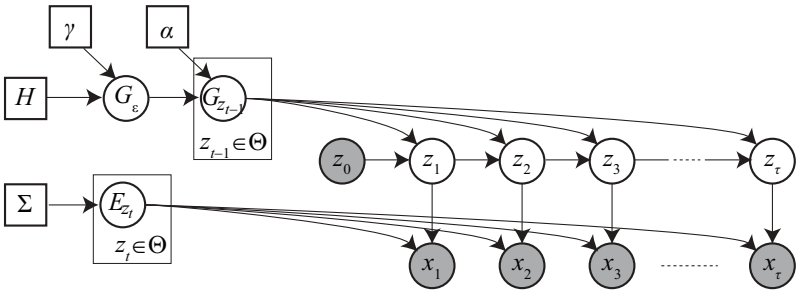

Figure 2: Graphical representation of iHMM.

pler [1], a Markov chain Monte Carlo sampling method that uses the Metropolis-Hastings step [6] to correct bias caused by restricted coupled draws from HCRP. We made empirical evaluations on a small artificial sequence and a natural language text to demonstrate the flexible inference by HN-iHMM.

2 Background

2.1 Hierarchical Dirichlet Process

Consider a tree structure with a node indexed by a sequence $z = z^{(1)}z^{(2)}\cdots z^{(\ell)} = z^{(1:\ell)}$ of $z \in Z$, where Z is a countably infinite set of index elements (say, integers). Let ε denote the null sequence, and zw is the concatenation of element w to sequence z. We can associate each node in a tree with a Dirichlet process (DP) to compose a hierarchical DP (HDP) [12], in which a DP associated to a node provides a base measure for the DPs associated to child nodes. Let $(\Theta, \mathscr{B})$ be a measurable space, with H a probability measure on the space, and consider the following hierarchy of DPs:

$$G_\varepsilon \sim \mathrm{DP}(\alpha_0, H) \qquad\qquad G_{zw} \sim \mathrm{DP}(\alpha_{|zw|}, G_z) \quad , \tag{1}$$

where $\alpha_\ell \in \mathbb{R}^+$ is the hyperparameter for DPs at the ℓ-th level in the tree (the root node is level 0). Figure 1 represents this relationship.[1]

We can use a CRP representation for this hierarchical DP to integrate out G_z, constituting hierarchical Chinese restaurant process (HCRP), which is also called as Chinese restaurant franchise [12]. A CRP that corresponds to probability measure G_z provides a predictive distribution of ith draw $\theta_{z,i}$ from G_z, conditioned by previous draws $\theta_{z,1}, \ldots, \theta_{z,i-1}$ from G_z by the Pólya urn scheme:

$$\theta_{zw,i}|\theta_{zw,1}, \ldots, \theta_{zw,i-1}, \alpha_{|zw|}, G_z \sim \frac{1}{\alpha_{|zw|}+i}\sum_{j=1}^{i}\delta_{\theta_{zw,j}} + \frac{\alpha_{|zw|}}{\alpha_{|zw|}+i}G_z \tag{2}$$

$$\theta_{\varepsilon,i+1}|\theta_{\varepsilon,1}, \ldots, \theta_{\varepsilon,i}, \alpha_0, H \sim \frac{1}{\alpha_0+i}\sum_{j=1}^{i}\delta_{\theta_{\varepsilon,j}} + \frac{\alpha_0}{\alpha_0+i}H \tag{3}$$

In order to apply Eqs. 2 and 3 recursively to integrate out all probability measures, we need an additional set of variables to track relationship between G_{zw} and its base measure G_z; specifically,

[1] If we assume disjoint space between different levels of elements, the tail element $z^{(\ell)}$ can represent the the whole sequence $z = z^{(1)}z^{(2)}\cdots z^{(\ell)}$, as we used in the index of the nCRP. However, we choose sequenced notation for HDPs so that parent-child relationship is naturally represented.

observing $\theta_{zw,i} = \theta$ from G_{zw} implies either (1) θ has been previously observed from G_{zw}, or (2) θ is observed from its base measure G_z. To represent this, we introduce latent variables $a_{zw,i}$ that represents many-to-one association between random variables such that $\theta_{zw,i} = \theta_{z,a_{zw,i}}$ and that $a_{zw,i} \neq a_{zw',j}$ for all i, j, and $w \neq w'$.[2] Hereafter we denote *seating arrangement* S_z as the set of past observations $\{\theta_{z,1}, \theta_{z,2}, \ldots\}$ and corresponding associations $\{a_{zw,1}, a_{zw,2}, \ldots\}$, $\boldsymbol{S}$ as the collection of all seating arrangements, and $\boldsymbol{\alpha}$ as the vector of hyperparameters $\langle \alpha_0, \alpha_1, \ldots \rangle$. Using these notations, we write $\mathrm{CRP}_z(\theta \,|\boldsymbol{S}, \boldsymbol{\alpha})$ as the predictive distribution of draw θ from $G_{|z|}$ conditioned by $\boldsymbol{S}$.

To make a stochastic update of $\boldsymbol{S}$ for a new draw $\theta_{z,i} = \theta$, a procedure addCustomer$(\theta, \mathrm{CRP}_z, \boldsymbol{S})$ is used; combined with corresponding removeCustomer procedure, we obtain a Gibbs sampler for the seating arrangements (see [13] for detail).

2.2 Infinite Hidden Markov Model

Consider an infinite discrete state space Θ, and a collection of random transition kernels $\{G_\theta : \theta \in \Theta\}$ drawn from the following HDP:

$$G_\varepsilon | \gamma, H \sim \mathrm{DP}(\gamma, H) \qquad G_\theta | \alpha, G_\varepsilon \sim \mathrm{DP}(\alpha, G_\varepsilon) \qquad \text{for } \theta \in \Theta, \tag{4}$$

where H is a base measure on probability space $(\Theta, \mathscr{B})$. This is the same as Eq. 1 except the tree structure is restricted to two levels and $Z = \Theta$ is assumed. Then, we can give a conditional distribution of the sequence of hidden states $z_1, \ldots, z_T$ and observations $x_1, \ldots, x_T$ given initial state $z_0 = \theta_0$ and the emission distribution $F_z(x)$ from state $z \in \Theta$ as follows:

$$z_t | z_{t-1}, G_{z_{t-1}} \sim G_{z_{t-1}} \qquad x_t | z_t, F_{z_t} \sim F_{z_t} \qquad \text{for } t = 1, \ldots, T \quad . \tag{5}$$

Figure 2 shows a graphical representation of iHMM.

To understand relation between the iHMM and finite HMM, it is useful to consider stick-breaking representation [10] of the HDP; readers should refer to literature [11] for detail. In this paper, we instead adopt hierarchical Chinese restaurant representation, which gives a predictive distribution of (z_t, x_t) conditioned by hyperparameters $\boldsymbol{\alpha} = \langle \alpha_0, \alpha_1 \rangle$ and seating arrangements $\boldsymbol{S}$ that contains information of past transitions $z_0, \ldots, z_{t-1}$ and latent association variables $a_{z_{t-1},t}$. Then the predictive distribution of z_t given $z_{t-1}, \boldsymbol{\alpha}, \boldsymbol{S}$ can be derived as: $p(z_t | z_{t-1}; \boldsymbol{S}, \boldsymbol{\alpha}) = \mathrm{CRP}_{z_{t-1}}(z_t | \boldsymbol{S}, \boldsymbol{\alpha})$.

2.3 Nested Dirichlet Process

The nested Dirichlet process (nDP) [9] is a nonparametric Bayesian prior for descending paths in an infinitely-branched tree structure. Note the difference between the nDP and the HDP. The HDP corresponds to a prior of a set of distributions that are organized as a tree structure, while the nDP corresponds to a prior of distributions over paths on a tree structure.

More specifically, an L-level nDP represents a distribution of L-length sequence $\theta^{(1)} \theta^{(2)} \cdots \theta^{(\ell)}$:

$$G^\theta \sim \mathrm{DP}(\gamma, H) \qquad \theta^{(\ell)} | \theta^{(\ell-1)}, G^{\theta^{(\ell-1)}} \sim G^{\theta^{(\ell-1)}} \quad (\ell = 1, \ldots, L) \quad . \tag{6}$$

This is similar to Eq. 5, except that distributions G^θ do not have a common base measure with non-zero probability mass. As a result, any two transition distributions G^θ and $G^{\theta'}$ are a.s. disjoint; in other words, there is zero probability of sharing a node between a sequence after θ and another sequence after θ'. Thus, given the index of the root node $\theta^{(0)}$, possible L-length sequence from this nDP forms an infinitely branched tree with L levels.

CRP representation of nDP, which is called nested CRP (nCRP) [3], is straightforward:

$$\theta^{(\ell)} | \theta^{(\ell-1)} \sim \mathrm{CRP}^{\theta^{(\ell-1)}} \qquad \mathrm{CRP}^{\theta^{(\ell-1)}} = \mathrm{CRP}(\gamma, H) \quad . \tag{7}$$

In the following, we introduce a collective notation $\boldsymbol{G} \sim \mathrm{nDP}(\gamma, H^L)$ to denote a set $\{G^\theta \,|\, \theta \in \Theta\}$ of distributions that composes the L-level nDP, and denote a draw of L-length sequence from an nDP as $\theta^{(1:L)} \sim \boldsymbol{G}$. Note that we use superscripts to identify a distribution in the nDP, and subscripts to identify a distribution in HDP.

[2] On the analogy of Chinese restaurants, we can regard $a_{zw,i}$ as the table index of ith customer of restaurant zw. The latter condition means that the sibling restaurants do not share the same table indices.

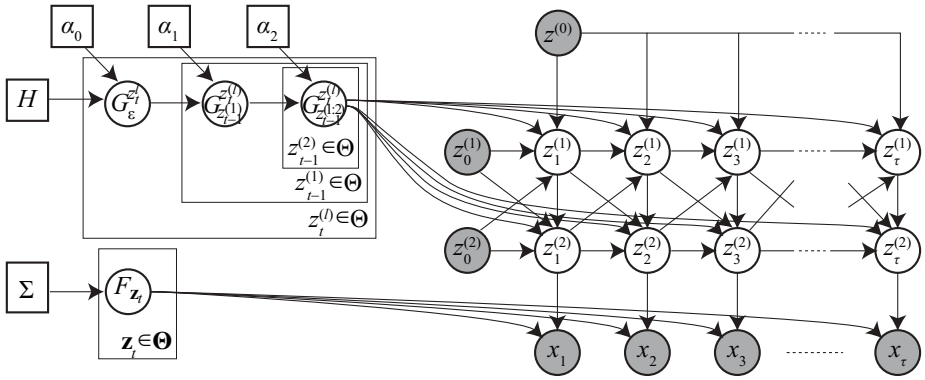

Figure 4: Graphical representation of the HN-iHMM

3 Hierarchical Nested Infinite Hidden Markov Models

In this section, we extend iHMM with hierarchical state clustering. More specifically, our goal is to obtain clustering of states with respect to the similarity on the transition probabilities. We break it down to two requirements to the prior: (a) Correlation between transition probabilities $p(z^*|z)$, $p(z^*|z')$ from states z, z' in the same cluster to another state z^*, and (b) Correlation between transition probabilities $p(z|z^*)$, $p(z'|z^*)$ from a state z^* to states z, z' in the same cluster.

To meet these requirements, state z in HMM is extended to L-dimensional vector $\boldsymbol{z} = z^{(1)} \cdots z^{(L)} = z^{(1:L)}$, providing L-level hierarchical structure of state space. The former requirement is achieved by introducing deeper levels into HDP, and the latter by introducing the nDP; the resulting combination is a new Bayesian prior, which we call the hierarchical nested iHMM (HN-iHMM).

Formally, an L-level HN-iHMM can be represented as follows:

$$G^{\theta}_{\varepsilon} | \sim \mathrm{DP}(\alpha_0, H) \qquad\qquad G^{\theta}_{zw} | \boldsymbol{z}, w, \theta, G^{\theta}_{z} \sim \mathrm{DP}(\alpha_{|zw|}, G^{\theta}_{z}) \quad , \tag{8}$$

where $w, \theta \in \Theta$ are index elements and $\boldsymbol{z}$ is a sequence of index elements $z^{(1:\ell)}$ with length $\ell < L$.

We can interpret the HN-iHMM in two ways. In the first way, the HN-iHMM can be seen as an extension of the nDP, in which each DP in Eq. 7 is replaced with an $L+1$-level HDP. Given a fixed superscript θ, Eq. 8 are equivalent to the HDP model shown in Eq. 1. In other words, HN-iHMM consists of an infinite number of independent HDPs indexed by θ, which is the same structure as an nDP that consists of an infinite number of independent DPs. Starting from constant $\theta = \theta^{(0)}$, recursive draw of $\theta^{(\ell)}$ from an HDP indexed by $\theta^{(\ell-1)}$ constitutes the stochastic process for generating L-length sequence $\theta^{(1)} \cdots \theta^{(L)}$.

In the second way, HN-iHMM can be seen as an extension of HDP, in which each DP in Eq. 1 is replaced with an nDP. Using a collective notation $\boldsymbol{G}$ we introduced at Sec. 2.3, Eq. 8 can be rewritten to be similar to Eq. 1:

$$\boldsymbol{G}_{\varepsilon} \sim \mathrm{nDP}(\alpha_0, H^L) \qquad\qquad \boldsymbol{G}_{zw} | \boldsymbol{z}, w, \boldsymbol{G}_{z} \sim \mathrm{nDP}(\alpha_{|zw|}, \boldsymbol{G}_{z}) \quad , \tag{9}$$

where the notation $\boldsymbol{G} \sim \mathrm{nDP}(\alpha, \boldsymbol{G}')$ in Eq. 9 denotes $G^{\theta} \sim \mathrm{DP}(\alpha, G'^{\theta})$ for all $\theta \in \Theta$. In other words, the HN-iHMM is a hierarchically structured nDP, in which an nDP indexed by subscript zw has a "base measure," i.e., a set of distributions associated to another nDP indexed by subscript z.

The transition distribution of destination states $\boldsymbol{z}_t = z_t{}^{(1:L)}$, given source state $\boldsymbol{z}_{t-1}$, is as follows.

$$z_t^{(0)} = z^{(0)} \quad \text{(constant)} \qquad\qquad z_t^{(\ell)} | z_t^{(\ell-1)}, \boldsymbol{z}_{t-1}, G^{z_t^{(\ell-1)}}_{\boldsymbol{z}_{t-1}} \sim G^{z_t^{(\ell-1)}}_{\boldsymbol{z}_{t-1}} \quad . \tag{10}$$

This means that, given the source state $\boldsymbol{z}_{t-1}$, elements $z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(L)}$ are sampled according to the nDP represented as $\boldsymbol{G}_{\boldsymbol{z}_{t-1}}$, and the vector $\boldsymbol{z}_t$ becomes the destination state .

The state transition probability (in HMM sense) can be recovered by regarding the vector of elements $\boldsymbol{z}_t$ as a state. In particular, the predictive distribution of transition probability $p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1})$ from state $\boldsymbol{z}_{t-1}$ to state $\boldsymbol{z}_t$ can be obtained by introducing CRP representation to integrate out all probability

measure G as follows (omitting $\boldsymbol{S}$ from the conditioning variables):

$$p(z_t^{(\ell)}|z_t^{(\ell-1)},\boldsymbol{z}_{t-1}) = \mathrm{CRP}^{z_t^{(\ell-1)}}_{z_{t-1}}(z_t^{(\ell)}) \tag{11}$$

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \prod_{\ell=1}^{L} p(z_t^{(\ell)}|z_t^{(\ell-1)},\boldsymbol{z}_{t-1}) = \mathrm{CRP}^{z^{(0)}}_{z_{t-1}}(z_t^{(1)}) \cdot \mathrm{CRP}^{z_t^{(1)}}_{z_{t-1}}(z_t^{(2)}) \cdots \mathrm{CRP}^{z_t^{(L-1)}}_{z_{t-1}}(z_t^{(L)}) \tag{12}$$

Finally, we present emission distribution to complete the HN-iHMM:

$$x_t|\boldsymbol{z}_t, F_{z_t} \sim F_{z_t} \quad . \tag{13}$$

It is straightforward to consider Bayesian prior to emission distribution F_z as well. When the emission distributions have a DP prior, we can also introduce similarity on emission symbols for the states in the same cluster by introducing an HDP indexed by z, as we have done in the experiments. Figure 4 shows a graphical representation of HN-iHMM with $L = 2$ levels. The left part corresponds to the hierarchical nDP (or nested HDP), and the right part shows HMM with hierarchical states. If we set $L = 1$, the model becomes identical to the original iHMM [2].

Let us confirm the two requirements (a),(b) above to see that our goal is achieved. Consider the states z, z' are in the same cluster at the ℓth level, i.e., they share the prefix $\boldsymbol{z}_p = z^{(1)} \cdots z^{(\ell)}$.

- $\boldsymbol{G}_z$ and $\boldsymbol{G}_{z'}$ have similar priors through the common base measure $\boldsymbol{G}_{\boldsymbol{z}_p}$ associated to the cluster $\boldsymbol{z}_p$. Thus, the destination distributions from the two states z and z' are likely to be similar.
- According to Eq. 12, transition probabilities from another state z^* to the two states z and z' share ℓ out of L factors. Since each factor in Eq. 12 is independently drawn from the prior, sharing the factors induces correlation between the two transition probabilities.

Figure 5 shows an example of generated transition matrix from an HN-iHMM. As you can see, the states within a cluster has similar transition destinations, and transition probabilities to the states within a cluster from another state is also correlated.

4 Inference

Here we provide a sampling algorithm for inferring the posterior distribution of an HN-iHMM given a sequence of emission symbols $x_1, \ldots, x_T$. The proposed algorithm is basically a step-wise Gibbs sampling, i.e., for each t in random order, resamples $\boldsymbol{z}_t = (z_t{}^{(1)}, \ldots, z_t{}^{(L)})$ from the posterior distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}, \boldsymbol{z}_{t+1}, \ldots, \boldsymbol{z}_T)$. We did not apply a conventional MCMC sampling technique that uses stick-breaking representation [14, 5] because it is computationally expensive and mixes slowly if the model contains a large number of distributions, such as HN-iHMM does. However, to adopt CRP representation, we need a special care to handle probabilities of coupled draws from CRP. Even in the case of $L = 1$ (ordinary iHMM), resampling z_t involves conditional joint distribution of two transitions, $z_{t-1} \to z_t$ and $z_t \to z_{t+1}$, which are coupled through collapsed distributions $G_{z_{t-1}}$, G_{z_t} and G_ε. It is complicating to make sampling from the coupled draw with restriction that the draws are consistent with the rest of the sequence. In fact, the first inference algorithm for iHMM [2] was a biased sampler that ignores the coupling term.

The RCD sampler introduces Metropolis-Hastings step to correct coupling term, and applicable to HN-iHMM. The key observation of the RCD sampler is that we actually need a proposal of a pair $(z_i^*, a^*_{z_{i-1},i})$, where z_i^* is a draw from HCRP z_t and $a^*_{z_{i-1},i}$ is the corresponding table assignment (as described in Sec. 2.1). The RCD sampler factorizes the proposal distribution $q(z_i^*, a^*_{z_{i-1},i})$ into $q_z(z_i^*)q_a(a^*_{z_{i-1},i}|z_i^*)$, and efficiently calculates the factor $r^* = p(z_i^*, a^*_{z_{i-1},i})/q_a(a^*_{z_{i-1},i}|z_i^*)$ that is needed in calculating acceptance ratio R, along the standard CRP update procedures addCustomer/removeCustomer, while leaving freedom to use any draw proposal $q_z(z_i^*)$.

In HN-iHMM, we use the approximated sampler that ignores coupling for proposal distribution, with a small adjustment to avoid cases where $p(\boldsymbol{z}_t^* = \boldsymbol{z}_{t+1})$ becomes zero. Algorithm 1 shows the Step-wise Gibbs sampling for HN-iHMM. For detail of the RCD sampler, please refer [1].

5 Experiments

We performed two sets of experiments to evaluate the proposed model. The first is on a small artificial sequence, and the second is on a sequence of natural language words.

Algorithm 1 A Step-wise Gibbs for HN-iHMM

Input: $x_{1:T}$: observed emissions, $z_{1:T}$: previously inferred states, $\boldsymbol{S}$: CRP seating arrangements
for $t = 1, \ldots, T$, in random order **do**
$z_t^{old} := z_t$; $\boldsymbol{S}^{old} := \boldsymbol{S}$; $p_3^{old} := F_{z_t^{old}}(x_t)$
$\boldsymbol{S}_{-1} := \mathsf{removeCustomer}(z_{t+1}, \boldsymbol{CRP}_{z_t^{old}}, \boldsymbol{S}^{old})$; $p_2^{old} := \boldsymbol{CRP}_{z_t^{old}}(z_{t+1}|\boldsymbol{S}_{-1})$
$\boldsymbol{S}_0 := \mathsf{removeCustomer}(z_t^{old}, \boldsymbol{CRP}_{z_{t-1}}, \boldsymbol{S}_{-1})$; $p_1^{old} := \boldsymbol{CRP}_{z_{t-1}}(z_t^{old}|\boldsymbol{S}_0)$
$\boldsymbol{S}^{adj} := \mathsf{addCustomer}(z_{t+1}, \boldsymbol{CRP}_{\varepsilon}, \boldsymbol{S}_0)$ {adjustment for avoiding zero probability}
Sample z_t^* in proportional to $q_z(z_t) := \mathrm{CRP}_{z_{t-1}}(z_t|\boldsymbol{S}^{adj}) \cdot F_{z_t}(x_t) \cdot \mathrm{CRP}_{z_t}(z_{t+1}|\boldsymbol{S}^{adj})$
$p_1^* := \boldsymbol{CRP}_{z_{t-1}}(z_t^*|\boldsymbol{S}_0)$; $\boldsymbol{S}_1 := \mathsf{addCustomer}(z_t^*, \boldsymbol{CRP}_{z_{t-1}}, \boldsymbol{S}_0)$
$p_2^* := \boldsymbol{CRP}_{z_t^*}(z_{t+1}|\boldsymbol{S}_1)$; $\boldsymbol{S}_* := \mathsf{addCustomer}(z_t, \boldsymbol{CRP}_{z_t^*}, \boldsymbol{S}_1)$; $p_3^* := F_{z_t^*}(x_t)$
$r^* := p_1^* \cdot p_2^* \cdot p_3^*$; $r^{old} := p_1^{old} \cdot p_2^{old} \cdot p_3^{old}$; $R := \min\left\{1, \frac{r^*}{r^{old}} \frac{q_z(z_t^{old}|\boldsymbol{S}_0)}{q_z(z_t^*|\boldsymbol{S}_0)}\right\}$ {Acceptance ratio}
$\langle z_t, \boldsymbol{S}\rangle := \begin{cases} \langle z_t^*, \boldsymbol{S}^*\rangle & \text{with probability } R \\ \langle z_t^{old}, \boldsymbol{S}^{old}\rangle & \text{otherwise} \end{cases}$
end for

As for the artificial sequence, we prepare a 500-length artificial sequence with 7 symbols, generated from an artificial HMM with 30 states (Fig. 6). This sequence consists of many similar sequences (noisy A-B-C emission) with variations; we expect that clusters in the HN-iHMM capture the common feature of the sequences, and states within a cluster capture the variation among the sequences. Since we have true states $s_{1:T}$ that generate the symbols, we compared the model in terms of the mutual information between the true states and inferred hidden states $z_{1:T}$: $I(s_{1:T}; z_{1:T}) = \sum_{s,z} p(s,z) \log_2[p(s,z)/(p(s)p(z))]$. For each condition, we performed 50 runs of the sampler, and mutual information from the result of the 20,000th sweep is calculated. We ran the tests with several set of hyperparameters and used the best for each model. We used beam sampler [14] for iHMM.

Figure 7 shows the results of experiments. The mutual information from the results of the iHMM stayed around 3.6. They tend to fall in a local optimum. with around 15 hidden states, each of which corresponds to two original states (e.g. states 0 and 1 in Fig. 6 and so on). On the other hand, the 2-level HN-iHMM can produce the mutual information around 4.1 using around 50 states within 12 clusters. correctly identifying most of the states but still fail to distinguish some states (10 and 11 for example). The HN-iHMM model typically uses 8 clusters, around 30 subclusters and 60 states; the obtained mutual information went close to 4.7, which is the maximum (equal to the entropy of true states), meaning that the true states are completely identified.

Note that, unlike usual HMMs, spurious states observed in HN-iHMMs does little harm, because the states share the transition distribution with other states through clusters, In fact, the learned model gives better prediction even with spurious states (Table 1). Rather, by allowing spurious states, the HN-iHMM gives more chance to the sampler for searching with additional states, resulting a better inference. Note also that the provided length of the emission is not so long (500 symbols); Despite the complexity of the model, the required amount of data seems not increaased. This also indicates that the HN-iHMM can deal with sparseness of hidden states.

In another test, we applied an HN-iHMM to perform hierarchical clustering of words from the text of *Alice's Adventure in Wonderland*. The text is converted to lower case, punctuations are removed, and a special word EOS is put after every sentence to obtain a corpus with $T = 28,120$ words; we introduce a special word UNK (unknown) to replace every word that occurs less than 3 times, resulting $|\Sigma| = 1,078$ unique words in the text. For these experiments, we introduced prefix-based blocked Gibbs sampler, which tries to change all occurrence of a certain state z to another unused state z^* at once, and combined with step-wise Gibbs sampler described above. We also introduce the sampling of hyperparameters as done in previous work [12]. The last $\tau = 1,000$ words of the text are held out for testing, and trained the model with the remaining $T = 27,120$ words. The quality of prediction is evaluated by perplexity (reciprocal geometric mean of emission probability): $PPL = \exp \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \log p(x_t)$, where $p(x_t) = \frac{1}{J} \sum_{j=1}^{J} p(x_t|x_{1:t-1}, \boldsymbol{S}^{[j]})$ and $\boldsymbol{S}^{[j]}$ denotes the sampling state after j-th sweep of the Gibbs sampler.

Figure 8 shows the progress of inference for sampling processes. Each line corresponds to one sampling process. As shown in Fig. 8(a), the number of states inferred by 2-level HN-iHMM is

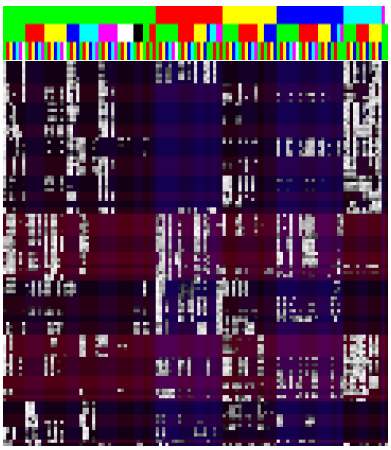

Figure 5: A transition matrix generated from 3-level HN-iHMM. Each row represents transition distribution from a source state (white cell indicates high probability). The top color bar and blue/red shadows indicate cluster configuration.

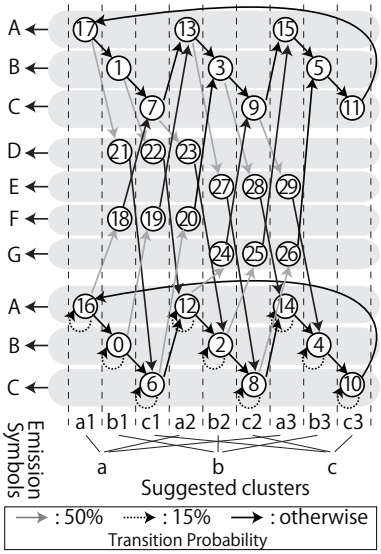

Figure 6: HMM that generates the artificial sequence. "Suggested clusters" are just for reader's help.

Table 1: Perplexity (reciprocal geometric mean of symbol emission probability) for 100-length held-out data. Smaller values are better.

model	perplexity
iHMM	2.34
2-level HN-iHMM	2.08
3-level HN-iHMM	2.03

Table 2: Learned clusters from *Alice's Adventure in Wonderland*. From each of 15 largest clusters, up to 5 most frequent states are chosen, and up to 5 words from each state are shown in a row (words less than 8 occurrences are omitted). Numbers indicate the frequency of the state occurrences and word emissions within the inferred state sequence, while numbers within brackets indicate the number of level-2 states in the cluster.

Level 1 (clusters)	Level 2 (states)	emission symbols with occurrence counts				
A	a 1577	1309–EOS	130–and	59–but	18–or	12–now
1598[6]	b 11	11–or				
B	a 486	169–she	75–and	52–it	42–alice	37–he
1438[22]	b 339	241–she	35–it	24–he	18–they	16–alice
	c 119	51–it	34–there	10–that	8–this	
	d 100	54–it	15–she	10–there	9–he	
	e 95	78–alice				
C	a 464	456–the				
1321[19]	b 235	235–the				
	c 224	220–the				
	d 158	151–the				
	e 48	46–the				
D	a 640	180–at	64–for	53–on	46–about	44–by
1035[17]	b 129	114–of				
	c 94	66–of				
	d 77	64–of				
	e 36	20–to				
E	a 677	149–as	107–and	84–that	65–EOS	58–but
975[16]	b 86	27–so	15–then	13–however	9–but	
	c 72	58–as				
	d 39	18–and	9–when			
	e 33	24–that				
F	a 702	395–a	203–the	26–an	18–this	12–another
960[15]	b 72	62–a	10–an			
	c 67	54–a	8–an			
	d 53	44–a				
	e 30	27–a				
G	a 489	222–i	157–you	26–they	24–we	17–i'll
814[19]	b 104	97–i				
	c 35	30–you				
	d 32	29–you				
	e 31	21–you	10–not			
H	a 519	184–in	117–with	64–into	36–like	33–to
695[12]	b 84	71–in	10–with			
	c 30	16–in	10–after			
	d 25	15–in				
	e 9	8–in				
I	a 468	460–to				
625[10]	b 82	79–to				
	c 31	19–and	10–to			
	d 15	14–to				
J	a 106	52–know	9–like	8–mean	8–wonder	
571[36]	b 50	17–wish	16–think			
	c 40	14–do	11–say			
	d 36	9–talk				
	e 30	18–see				
K	a 199	47–would	41–must	24–might	20–could	12–may
535[21]	b 63	32–don't	16–should			
	c 55	17–can				
	d 50	14–can't	12–could			
	e 37	15–never				
L	a 126	22–king	19–queen	11–dormouse	11–gryphon	10–mouse
533[18]	b 104	18–king	16–gryphon	14–queen	11–duchess	11–hatter
	c 104	24–hatter	19–gryphon	14–caterpillar	10–king	10–cat
	d 75	14–queen	14–dormouse	12–mouse	8–cook	
	e 29	8–queen				
M	a 416	70–down	68–up	41–out	33–back	30–off
527[21]	b 32	17–on				
	c 15	12–on				
	d 11	8–off				
N 429[10]	a 393	344–and	18–UNK			
O 426[12]	a 398	320–said	26–thought	15–cried	–	

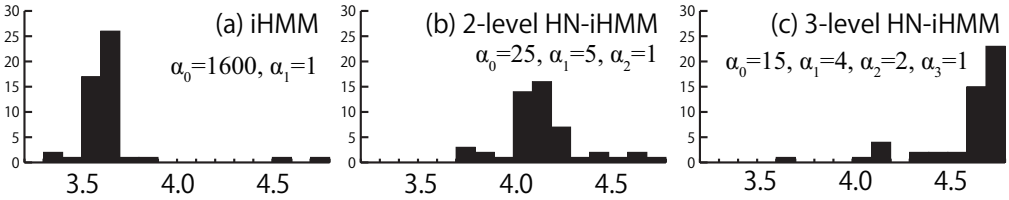

Figure 7: Results of inference for the artificial sequence. X-axis denotes mutual information between the sampled states and true hidden states, and Y-axis denotes frequency.

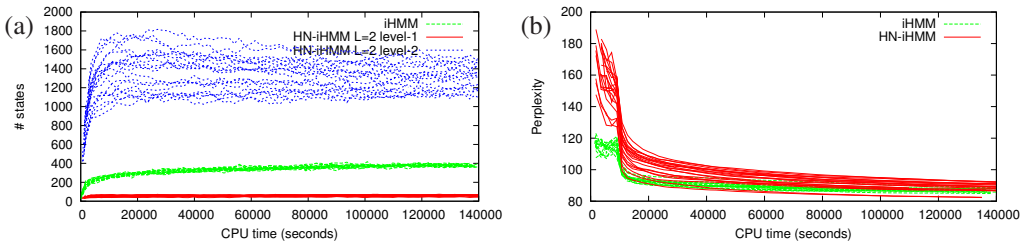

Figure 8: Test results on *Alice's Adventure in Wonderland*

larger than that of iHMM. but the number of level-1 states (that is, the number of clusters) stays around the range from 50 to 100. The plot also indicates that the results are not mixed well during the given CPU time, partly because we introduced hyperparameter sampling, that slows down mixing. At this moment perplexity does not differ significantly (Fig. 8(b)).

Table 2 shows the learned clusters taken from a snapshot at the 103,000th sweep of a Gibbs sampling. You can see states with similar syntactic features are gathered into clusters. For example, clusters D and H correspond to propositions, J to root form of verbs, K to the auxiliary verbs, and L to nouns; States within a cluster (such as Ja, Jb, ...) correspond to different subtypes. It is interesting to see that the model finds domain-specific knowledge, such as distinguishing nominative cases of narrative sentences (cluster B) from nominative cases within dialogs (cluster G). Moreover, apparently redundant states within a cluster (such as states in clusters C and F) seems to play the role to convey such information over non-adjacent words, while sharing basic grammatical rules between states within the cluster. This result indicates that the HN-iHMM can be applied for unsupervised learning of hierarchical clusters from a sequence.

6 Related work

The infinite hierarchical hidden Markov model (iHHMM) [7] is a nonparametric Bayesian model of hierarchical HMM (HHMM) [4]. The difference between our work and iHHMM lies in the use of hierarchy. The HHMM, as well as the iHHMM, uses a hierarchy to capture the temporal structure of the sequence. Upper levels in the hierarchy correspond to a slower change in a temporal sequence, e.g., the states at the lowest level may correspond to letters, their upper states correspond to words, and phrases, and so on. On the other hand, in our work, hierarchy is used to capture nested clusters of states, in which lower levels corresponds to minor categories. Thus the iHHMM and our work extends the iHMM in different directions; it would be possible to consider an HMM model that combines both extensions.

7 Conclusion

We proposed a new nonparametric Bayesian prior, called the hierarchical nested infinite hidden Markov model (HN-iHMM), which models hierarchical state clustering in hidden Markov models. This is a combination of the hierarchical Dirichlet process, for modeling the similarity of distributions over destination states, and the nested Dirichlet process, for modeling the similarity of distributions over source states. We also presented an inference algorithm of the HN-iHMM based on the Restricted Collapsed Draws sampling algorithm. The experiments showed that the proposed model showed flexibility of the proposed model.

References

[1] A. Anonymous. Restricted collapsed draws from hierarchical chinese restaurant process. Submitted; to be appeared on arXiv, 2011.

[2] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584. MIT Press, 2002.

[3] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[4] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[5] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 312–319, 2008.

[6] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[7] Katherine Heller, Yee Whye Teh, and Dilan Gorur. Infinite hierarchical hidden Markov models. *Journal of Machine Learning Research, Workshop and Conference Proceedings Series*, 5:224–231, 2009. (AISTATS 2009).

[8] Julian Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[9] Abel Rodríguez, David B. Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[10] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[11] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort et al., editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

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

[13] Yee Whye Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report Technical Report TRA2/06, School of Computing, NUS, 2006.

[14] J. van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1088–1095, 2008.