|  |
| --- |
|  |
| **Left-to-Right HDP-HMM with HDP Emission** |
|  |

**AnonymousAnonymous**

Affiliation Affiliation

Address Address

*email* *email*

|  |
| --- |
|  |

**Anonymous**

Affiliation

Address

email

**Abstract**

In this paper, we introduce a new nonparametric Bayesian HMM based on the well-known HDP-HMM model. Unlike the original HDP-HMM which is ergodic, our model has a left-to-right structure. We also introduce two approaches to add non-emitting states which are used to model the beginning and end of finite sequences. Finally, we extend the HDP-HMM definition and introduce HDP-HMM with HDP emissions. By experimentations, we will show the new model can outperform the ergodic model in certain conditions and can learn the underlying structure of the generative model.

**1 Introduction**

Hidden Markov models (HMMs) [] are among the most powerful statistical modeling tools and have found a wide range of applications in many pattern recognition area such as speech recognition [], machine vision [], genomics [] and finance []. HMMs are parameterized both in their topology (e.g. number of states) and emission distributions (e.g. Gaussian mixtures). Model comparison methods are traditionally used to select the adequate number of states and mixture components. However, these methods are computationally expensive and moreover there is no consensus on the optimum criterion for the selection [].

An infinite HMM has been developed in the last few years [][][] based on nonparametric Bayesian approaches. In this model, instead of defining a parametric prior over the transition distribution, a hierarchical Dirichlet process (HDP) prior is used. Hence, this model is known as an HDP-HMM model. HDP-HMM introduced in [] and [] is an ergodic model (a transition from an emitting state to all others is allowed). However, in many pattern recognition applications involving temporal structure, such as speech processing, a left-to-right topology is preferred or sometimes required []. For example, in continuous speech recognition applications we want to model speech units (e.g. phonemes), which evolve in a sequential manner, using HMMs. Since we are dealing with an ordered sequence (e.g. a word is an order sequence of phonemes) left-to-right model is preferred []. Moreover, speech data is not segmented into these units in advance and therefore in the training process we need to be able to connect these smaller models together into a larger HMM that models the entire utterance. Obviously, this task can easily be achieved using left-to-right HMMs.

If the data has a finite length the beginning and end of a sequence should be modeled as two additional discrete events – non-emitting initial and final states []. In the original HDP-HMM formulation [][] this problem is not addressed. Also, in the original HDP-HMM (and also parametric HMMs), the emission distribution for each state is estimated by data points mapped to that state. For example, if we use a Gaussian mixture model (GMM) to model the emission distribution, for every state we compute a separate GMM and components can’t be shared or re-used within a model. (comments has problem so I write here: I mean with in a model not between different models)

In this paper we propose a left-to-right HDP-HMM with non-emitting initial and final states. In our model, emission distributions are modeled using GMMs with an infinite number of components. Sharing components is achieved by using an HDP prior instead of Dirichlet process (DP) priors as in []. We review some background material in Section 2. Our contribution and proposed model is discussed in Section 3 and some experimental results are presented in Section 4.

**2 Background**

A Dirichlet process [??][??] is a discrete distribution that consists of countable infinite probability masses. A DP is denoted by *DP(α,H),* where *α* is the concentration parameter and *H* is the base distribution. The base distribution, *H*, is defined by [set1994]:



In this definition, is the unit impulse function at , and is referred to as an atoms [hdp2004]. The weights , are sampled through a stick-breaking construction []:



The sequence of *βk* sampled by this process satisfies the constraint  with probability 1 [?] and are denoted by *β~GEM(α)*. One of the main applications of DP is to define a nonparametric prior distribution on the components of a mixture model. For example, a DP can be used to define a Gaussian mixture model (GMM) with an infinite number of mixture components []. This is a useful model in many areas of science. For example, in speech recognition, an acoustic unit (a word or a phoneme) can be modeled using a GMM [].

A hierarchical Dirichlet process extends a Dirichlet process to grouped data []. In this case there are several related groups and the goal is to model each group using a mixture model while linking these models, for example via parameter sharing. One approach is to use a DP to define a mixture model for each group and place a global Dirichlet process, *DP(γ,H)*, as the common base distribution for all DPs []. An HDP is defined as:



where *H* provides the prior for the parameters and *G0* represents the average of the distribution of the parameters (e.g. means and covariances). For example, consider the problem of modeling acoustic units using a mixture model in which parameters of different units can be shared with each other.

An alternative representation, which is useful in deriving the inference algorithms, is Chinese restaurant franchise (CRF) representation []. In a CRF, we have a franchise with several restaurants and a franchise wide menu. The first customer (observed data) in the restaurant *j* (group *j*) sits at one of the tables (clusters) and orders an item from the menu. Other customers either sit at one of the occupied tables and eat the food served at that table or sit at a new table and order their own food from the menu. Moreover, the probability of sitting at a table is proportional to the number of customers already seated at that table. However, if a customer starts his own table (with probability proportional to *α*), he orders a food from the menu with probability proportional to the number of tables serving that food in the franchise, or alternately orders a new food item with a probability proportional to *γ*.

An HDP-HMM [?][?][?] is an HMM with an unbounded number of states. In a typical ergodic HMM, the number of states is fixed so a matrix of dimension *N* states by *N* transitions per state is used to represent the transition probabilities. In an HDP-HMM, the transition matrix is replaced by an infinite, but discrete transition distribution, modeled by an HDP for each state. This lets each state have a different probability distribution for its transitions while the set of reachable states would be shared among all states. Fox et al. [?] extends the definition of HDP-HMM to HMMs with state persistence by introducing a sticky parameter κ. The definition for HDP-HMM is given by:



The state, mixture component and observation are represented by *zt*, *st* and *xt* respectively. The indices *j* and *k* are indices of the state and mixture components respectively. The base distribution that links all DPs together is represented by *β* and can be interpreted as the expected value of state transition distributions. The transition distribution for state *j* is a DP denoted by *πj* with a concentration parameter *α*. Another DP, *ψj*, with a concentration parameter *ϭ*, is used to model an infinite mixture model for each state (*zj*). The distribution *H* is the prior for the parameters *θkj*. If we want the posterior distribution over the parameters to remain in the same family as the prior, then *H* should be chosen to be a conjugate prior to the observation likelihood. Since the likelihood is a multivariate normal, the conjugate prior is normal inverse Wishart distribution.

**3 Left-to-Right HDP-HMM with HDPM Emission**

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain []. The state of a Markov chain at time *t* is denoted by *zt* and an observation is denoted by  where *F* is the emission distribution (e.g., a Gaussian mixture) and *st* is a mixture component index. In an HMM, there is a probability distribution to transit into state *zt*. In infinite HMMs, this transition distribution should have infinite support and as discussed previously is modeled using HDP. For state *j* this transition distribution is denoted by *πj*:



From Eq. we can see transition distribution has no topological restriction and therefore Eq. defines an ergodic HMM. In this section we introduce a left-to-right HDP-HMM with initial and final non-emitting states. Moreover, we replace DP with HDP to model multimodal emission distributions that allow states to share mixture components.

**3.1 Left-to-Right Transitional Distribution**

In order to obtain a left-to-right topology we need to force the base distribution of the Dirichlet distribution in Eq. to only contain atoms to the right of the current state. This mean *β* should be modified so the probability of transiting to states left of the current state (i.e. states visited so far) becomes zero. For state *j* we define *Vj={Vji}* as:



where *i* is index for all states. Then we can modify *β* by multiplying it to this vector:



Therefore to obtain a left-to-right HDP-HMM, we simply replace with *β* in Eq.. The rest of the definition remains the same. Also notice that defining a different type of topology can be achieved by defining an appropriate *Vj* .

**3.2 Initial and Final Dummy States**

In many applications (such as continuous speech recognition), a left-to-right HMM begins from and ends with non-emitting states. These states are required to model the beginning and end of finite length sequences. Adding non-emitting initial state is trivial: the probability of transition into the initial state is one and the probability distribution of transition from this state is equal to *πinit* which is the initial probability distribution for an HDP-HMM without non-emitting states. However, adding a final non-emitting state is more complicated. In the following we will discuss two approaches to achieve this.

**3.2.1 Maximum Likelihood Estimation**

Consider state *zi* depicted in Figure 1. As this figure shows the outgoing probabilities for any state can be classified into three categories: (1) a self-transition (*P1*), (2) a transition to all other states (*P2*), and (3) a transition to a final non-emitting state (*P3*). Moreover, we have *P1+P2+P3=1*. Suppose that we obtained *P2* from the inference algorithm. We will need to reestimate *P1* and *P3* from the data. This problem is in fact equivalent to the problem of tossing a coin until we obtain the first tails (each head is equal to a self-transition and the first tails triggers a transition to the final state). This problem can be modeled using a geometric distribution []:



Figure 1- Outgoing probabilities for state *Zi*



Eq.  shows the probability of *K* heads before the first tails. In this equation *1-ρ* is the probability of heads (success). We also have:



Suppose we have a total of *N* examples but for just *M* examples the state *zi* is the last state of the model (*SM*). It can be shown [] that the maximum likelihood estimation is obtained by:



where *ki* are the number of self-transitions for state *i*. Notice that if *zi* never happens to be the last state (*M=0*), *P3=0*.

**3.2.2 Bayesian Estimation**

Another approach to estimate *ρ* is to use a Bayesian framework. Since a beta distribution is the conjugate distribution for geometric distribution [], we can put a beta distribution with hyperparameters *(a,b)* as the prior and therefore obtain a posterior as [diconis] []:



where *M* and *SM* are same as the previous section. Hyperparameters *(a,b)* can also be estimated using a Gibbs sampler if required [Quintana].

**3.3 HDPM Emission Distributions**

In previous works [] [], emission distributions for each state of an HDP-HMM were modeled using a Dirichlet process mixture (DPM) as shown in Eq. . While this model is reasonably flexible, each data point is strictly associated with a single state and hence statistical estimation of each parameter would be less reliable. This is a more serious problem for HDP-HMMs with a left-to-right topology since these models will discover more states. As a result the available data for estimating the emission distribution for each state would be more limited. The solution proposed here to address this problem is to replace the DPM with an HDP mixture (HDPM) defined for the whole HMM.

The final model (without non-emitting states) is given in Eq. and shown graphically in Figure 2-(b). For comparison Figure 2-(a) shows the graphical model of the original HDP-HMM [].



**3.4 Modified Block Sampler**

A block sampler for HDP-HMM with a multimodal emission distribution has been introduced by Fox et al. []. In this section we review the modifications of this algorithm needed for our new model. The interested reader should refer to [][] for additional details. The basic idea is to jointly sample the state sequence *Z1:T* given the observations, model parameters and transition distribution *πj*. A variant of forward-backward procedure [] is utilized that allows us to exploit the Markovian structure of the HMM. However it requires approximation of the theoretically infinite distributions with a “degree L weak limit” approximation that truncates a DP into a Dirichlet distribution with L dimensions []:



The sampling of transition distribution is similar to []. The only difference is to replace *β* with given in Eq. . Using a similar approximation we can write the following prior distributions for the global weightsand state specific weights  of HDPM emissions:





where is the order of approximation in this case. For the posterior distribution we can write:





where *Mjk*is the number of tables (clusters) in restaurant (state) *j* that serves dish (mixture component) *k*; is total number of tables in the franchise that serves dish *k*. is the number of observations in state *j* that are assigned to component *k*. Estimating transition probabilities to final non-emitting state can be achieved at the last step and after estimating other parameters.

**4 Experiments**

**Synthetic data.** In the first experiment, we generate data from a left-to-right HMM with four states and without non-emitting states. The emission distribution for each state is a GMM with up to three components consists of two dimensional normal distributions. Three sequences with lengths 600,650 and 650 are generated for training. Ergodic HDP-HMM, left-to-right HDP-HMM with DPM emissions and left-to-right HDP-HMM with HDPM emissions have been studied and compared. A normal inverse Wishart (NIW) prior is used on the mean and covariance. The truncation levels are set to 10 for both the number of states and mixture components. Pseudocounts is set to 0.1 and mean and scale matrix are set to the empirical mean and covariance and finally degree of freedom is set to five.

Figure 3-(a) shows the average likelihoods for different models over heldout data by averaging five independent chains. Figure (3)-b shows the structure of the generative model. Figure 3-(c) and Figure 3-(d) show discovered structure for ergodic HDP-HMM and left-to-right HDP-HMM respectively. As it can be seen from these figures, left-to-right HDP-HMM with HDPM emissions can discover the correct structure while the ergodic HDP-HMM finds a more simplified HMM. Moreover, we can see using HDP emissions can improves the likelihood. While left-to-right HDP-HMM with DPM emissions can find the structure close to the correct one (not shown here), its likelihood is slightly less than ergodic HDP-HMM. However, left-to-right HDP-HMMs with HDPM emissions show better likelihood than ergodic model. It is also interesting to note that the likelihoods of models discovered by all HDP-HMM algorithms are better than the likelihood of the generative model itself (which is the upper bound for the parametric models).

(a)

(b)

|  |  |
| --- | --- |
| Description: C:\Users\amir\Documents\My Dropbox\Projects\preliminary exam\fig3.jpg |  |

Figure 2-(a) Graphical model for ergodic HDP-HMM [] (b) Graphical model for the proposed left-to-right HDP-HMM with HDPM emissions.

**TIMIT Classification.** The TIMIT read speech data set [] is one of the most cited evaluation data used to compare new speech recognition algorithms. The data segmented manually into phonemes and therefore is a natural choice to evaluate phoneme classification algorithms. TIMIT contains 630 speakers from eight main dialect of American English []. The total numbers of utterances are 6300 where 3990 utterances are the standard training set and 150 utterances are core test set. We followed the standard practice of building models for 48 phonemes and then map them into 39 phonemes []. The first 12 Mel-Frequency Cepstral Coefficients (MFCC) plus energy and their first and second derivatives (delta and delta-delta) features have been used to convert speech data into 39 dimensional streams. We have used analysis Hamming window of 25 msec and frame rate of 10 msec. In this experiment, left-to-right HDP-HMMs with Gaussian and DPM emissions have been used. From previous experiments we expect left-to-right HDP-HMM with HDP emissions gave better results but due to computational reasons we have not completed experiments related to those models and the results for them will published in the future. We have used non-conjugate priors and placed a Gaussian prior on the mean and inverse-Wishart (IW) prior on the covariance matrix. Truncation levels are set to 10.

Table 1- Classification Error rates for different Algorithms

|  |  |
| --- | --- |
| Model | Classification Error Rate |
| Parametric HMM [] (10 mix.) | 27.8% |
| Left-to-Right HDP-HMM with Gaussian emissions | 26.7% |
| Left-to-Right HDP-HMM with DPM emissions | 25.1% |

|  |  |  |
| --- | --- | --- |
| (a) | | |
| (b) | (c) | (d) |

Figure 3-(a) Log-likelihoods of ergodic model, left-to-right HDP-HMM with DPM emissions and left-to-right model with HDP emissions are compared to each other. (b) The generative model structure. (c) The structure learnt by ergodic HDP-HMM. (d) The structure learnt by left-to-right HDP-HMM.

Table 1 compare the classification error of left-to-right models and parametric models. Since the maximum number of mixtures components set to 10, we have examined parametric HMMs with 10 components per state. As this table shows even left-to-right HDP-HMM with Gaussian emissions outperforms the parametric model.

Figure 4 shows the discovered structure for phonemes */aa/* and */sh/* using the proposed model. As the amount of data increases the system can learn a more complex model for the same phone. It is also important to note that structure learned for each phone is different and reflects underlying differences between phones. Also note that the learned structures models multi-modalities by learning several parallel left-to-right paths; i.e. in Figure 4-(c), S1-S2, S1-S3 and S1-S4 show three parallel models.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 4- An automatically derived model structure (without the first and last dummy states) for (a) */aa/* with 175 examples (b) */sh/* with 100 examples (c) */aa/* with 2256 examples and (d) */sh/* with 1317 examples using left-to-right HDP-HMM model. The data used in this illustration was extracted from the training portion of the TIMIT Corpus.







**6 Conclusion**

(c)

(a)

(b)

(d)

In this paper we introduced a left-to-right HDP-HMM with HDPM emissions. Through experimenting we have shown that the new model can successfully learn the underlying structure when the data is generated using generative left-to-right model. Moreover, it has been shown than the likelihood of the learnt model is higher than the ergodic model. In this paper we have also introduced two approaches to add non-emitting initial and final states to the left-to-right HDP-HMM model, and finally we present the modifications needed in the block sampler to implement the inference algorithm for the new model. Using some experiments on TIMIT data we have shown that the proposed model outperforms parametric HMMs and can learn a multimodal (e.g. speaker specific) structures from the data.

One of the current problems of the HDP-HMM model (including left-to-right model) is that the inference algorithm is still computationally expensive. It is become a serious problem when we are dealing with big datasets such as speech or video; therefore our next task is to improve the inference algorithm specifically for left-to-right HDP-HMMs with HDPM emissions using its specific properties and structure; i.e. the fact that number of possible transitions reduces as we observe more data.

Another possible direction is to replace HDP emissions with more general hierarchical structures such as Dependent Dirichlet Process [] or Analysis of Density (AnDe) Model []. It has been shown that AnDe model is the appropriate model for problems involves sharing statistical strength among multiple set of density estimators [] [].

**Acknowledgments**

This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854.

**References**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hiippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.