**COLLEGE OF ENGINEERING**



**Preliminary Exam Report**

**Review of Hierarchical Dirichlet Process and Infinite HMMs**

Amir Harati

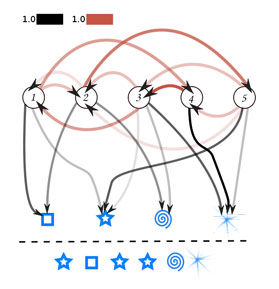
[Amir.harati@gmail.com](mailto:Amir.harati@gmail.com)

Institute for Signal and Information Processing,

Department of Electrical Engineering,

Temple University

Feb 2012



**Executive Summary**

In this report, I review three research papers concerning hierarchical Dirichlet processes (HDP), infinite hidden Markov models (HDP-HMM), and several inference algorithms for these models. I also use several other sources to support the material discussed in this report.

The three primary references used in this exam are "Hierarchical Bayesian Nonparametric Models with Applications,” “On-Line Learning for the Infinite Hidden Markov Model,” and “Sticky HDP-HMM with Application to Speaker Diarization". The basic materials in all of these papers involve HDP and HDP-HMMs. The first paper is a more general overview and contains several other models which are not related directly to the main theme of this report. Therefore this report is not organized based on any one of these references but contains most of the relevant material from those papers. In many cases I have corrected several errors ranging from typos to simple algorithmic/computational errors and have also used a unified notation through the report that I believe makes the presentations easier.

To make the report more self-contained I have added a background review of Dirichlet processes (DP), but this review is very short and readers may need to review some background papers before reading this report. After reviewing DPs, I start by introducing HDP and the reason that they are needed. Several properties of HDP are derived (both in the main report and the appendix) and some of its properties are justified. Two basic inference algorithms for HDPs are presented in detail.

HDP-HMM is introduced based on the general framework of hierarchical Dirichlet processes. The differences between this new model and HDP are emphasized and several of their properties are derived and explained. Three important inference algorithms are reviewed and presented in detail.

In writing this review, one of my primary intentions is to produce a self-sufficient document that can be used as a reference to implement some of the inference algorithms for HDP-HMM. Moreover, an interested reader can easily start from these and derive more general or application specific algorithms.

Table of Contents

[1 Introduction 1](#_Toc318303857)

[2 Background 1](#_Toc318303858)

[3 Hierarchical Dirichlet Process 2](#_Toc318303859)

[3.1 Stick-Breaking Construction 2](#_Toc318303860)

[3.2 Different Representations 4](#_Toc318303861)

[3.3 Chinese Restaurant Franchise 4](#_Toc318303862)

[3.3.1 Posterior and Conditional Distributions 5](#_Toc318303863)

[3.4 Inference Algorithms 6](#_Toc318303864)

[3.4.1 Posterior Sampling in CRF 7](#_Toc318303865)

[3.4.2 Augmented Posterior Representation Sampler 9](#_Toc318303866)

[3.5 Applications 10](#_Toc318303867)

[4 HDP-HMM 11](#_Toc318303868)

[4.1 CRF with Loyal Customers 12](#_Toc318303869)

[4.2 Inference Algorithms 12](#_Toc318303870)

[4.2.1 Direct Assignment Sampler 12](#_Toc318303871)

[4.2.2 Block Sampler 16](#_Toc318303872)

[4.2.3 Learning Hyper-parameters 18](#_Toc318303873)

[4.2.4 Online learning 21](#_Toc318303874)

[4.3 Applications 26](#_Toc318303875)

[5 Conclusion 26](#_Toc318303876)

[6 Reference 27](#_Toc318303877)

[Appendix A: Derivation of HDP Relationships 29](#_Toc318303878)

[A.1. Stick-Breaking Construction (Teh Y. , Jordan, Beal, & Blei, 2006) 29](#_Toc318303879)

[A.2. Deriving Posterior and Predictive Distributions 30](#_Toc318303880)

[Appendix B: Derivation of HDP-HMM Relationships 32](#_Toc318303881)

[B.1. Derivation of the posterior distribution for 32](#_Toc318303882)

# Introduction

Nonparametric Bayesian methods provide a consistent framework to infer the model complexity from the data. Moreover, Bayesian methods make hierarchical modeling easier and therefore open doors for more interesting and complex applications. In this report, we review hierarchical Dirichlet processes (HDP) and its applications to derive infinite hidden Markov models (HMM) or HDP-HMM. We also review three inference algorithms for the so called HDP-HMM in details.

This report is organized into five sections and two appendixes. Section two is a quick review of Dirichlet process. Section three is devoted to HDP and its inference algorithms and section four is focused on HDP-HMM and its inference algorithms. For the sake of readability, some of the mathematical details are presented in the appendix sections.

# Background

A Dirichlet process (DP) is a distribution over distributions, or more precisely over discrete distributions. Formally, a Dirichlet processis “defined to be the distribution of a random probability measureoversuch that for any finite measurable partitionofthe random distribution is distributed as finite dimensional Dirichlet distribution” (Teh Y. , Jordan, Beal, & Blei, 2006) :



A constructive definition for Dirichlet process is given by Sethuraman (Sethuraman, 1994) which is known as stick-breaking construction. This construction explicitly shows that draws from a DP are discrete with probability one.



can be interpreted as a random probability measure over positive integers and is denoted by. In both of these definitions, or base distribution, is the mean of the DP, andis the concentration parameter which can be understood as the inverse of variance.

Another way to look at the DP is through the Polya urn scheme. In this approach, we have to consider i.i.d. draws from a DP and consider the predictive distribution over these draws (Teh Y. , Jordan, Beal, & Blei, 2006):



In the urn interpretation of equation , we have an urn with several balls of different colors in it. We draw a ball and put it back in the urn and add another ball of the same color to the urn. With probability proportional towe draw a ball with a new color. To make the clustering property more clear, we should introduce a new set of variables that represent distinct values of the atoms. Letto be the distinct values andbe the number of associated with. We would now have:



Another useful interpretation of is the Chinese restaurant process (CRF). In CRF we have a Chinese restaurant with infinite number of tables. A new customer comes into the restaurant and can either sit around one of the occupied tables with probability proportional to the number of people already sitting there or start a new table with probability proportional to. In this metaphor, each customer is a data point and each table is a cluster.

# Hierarchical Dirichlet Process

A Hierarchical Dirichlet Process (HDP) is the natural extension of a Dirichlet process for problems with multiple groups of data. Usually, data is split into groups a priori. For example, consider a collection of documents. If words are considered as data points, each document would be a group. We want to model data inside a group using a mixture model. However, we are also interested to tie groups to each other, i.e. to share clusters across all groups. Let’s assume that we have an indexed collection of DPs with a common base distribution. Unfortunately this simple model cannot solve the problem since for continues  different  necessary have no atoms in common. The solution is to use a discrete  with broad support. In other words,  is itself a draw from a Dirichlet process. HDP is defined by (Teh & Jordan, 2010) equation .



In this definition provides prior distribution for factor. governs the variability of  around andcontrols the variability of around . , and are hyper-parameters of HDP.

## Stick-Breaking Construction

Because is a Dirichlet distribution it has a stick-breaking representation:



Where  and. Since support of is contained in within the support of  we can write a similar equation to for:



Then we have:





We also have:



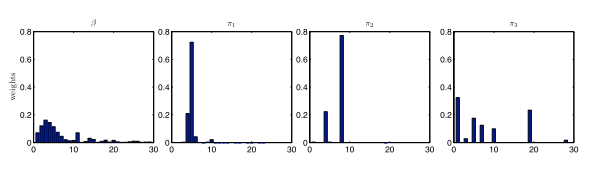


Figure - Stick Breaking Construction for HDP: The left panel shows a draw of, while the right three show independent draws conditioned on  (Teh & Jordan, 2010).

Stick-breaking construction for HDP and are derived in ‎A.1. Figure 1 demonstrates stick-breaking and cluster sharing of HDP.

## Different Representations

Definition shows the first representation of HDP. Another representation can be obtained by introducing an indicator variable as shown in equation .

Figure 2 shows the graphical models of both of these representations.



## Chinese Restaurant Franchise

The Chinese restaurant franchise (CRF) is the natural extension of Chinese restaurant process for HDPs. In CRF, we have a franchise with several restaurants and a franchise wide menu. The first customer in restaurant *j* sits at one of the tables 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. In this metaphor, restaurants correspond to groups and customerin restaurant** corresponds to (customers are distributed according to). Tables are i.i.d. variables distributed according toand finally foods are i.i.d. variables distributed according to. If customerat restaurantsits at tableand that table serves dish, we will have. In another way, each restaurant represents a simple DP and therefore a cluster over data points. At the franchise level we have another DP but this time clustering is over tables.



Figure -(a) HDP representation of (b) Alternative indicator variable representation (Teh, Jordan, Beal, & Blei, 2004)

Now let introduce several variables that will be used throughout this paper. is the number of customers in restaurant , seated around table,and who eat dish.is the number of tables in restaurant serving dishand is the number of unique dishes served in the entire franchise. Marginal counts are denoted with dots. For example,is the number of customers in restauranteating dish.

### Posterior and Conditional Distributions

CRF can be characterized by its state which consists of the dish labels, the tables  and dishes . As a function of the state of the CRF, we also have the number of customers, the number of tables, customer labels and table labels (Teh & Jordan, 2010). The posterior distribution ofis given by:



Where is the total number of tables in the franchise andis the total number of tables serving dish. Equation shows the posterior for.is the total number of customers in restaurant andis the total number of customers in restauranteating dish.



Conditional distributions can be obtained by integrating outandrespectively. By integrating outfrom we obtain:



And by integrating outfrom we obtain:



A draw from can be obtained using and a draw from can be obtained using .





From and we see that the posterior of is a mixture of atoms corresponding to dishes and an independent draw from andis a mixture of atoms at and an independent draw from (Teh & Jordan, 2010).

In an HDP each restaurant represents a DP and so  since the number of clusters scales logarithmically. On the other hand, equation shows is a DP over tables and so . This shows that HDP represents a prior belief that the number of clusters grows very slowly (double logarithmically) when increasing the number of data points but faster (logarithmically) in the number of groups (Teh & Jordan, 2010).

All of the above relationships are derived in ‎A.2.

## Inference Algorithms

First, let introduce several notations.,,,and.Wheredenotes the mixture component associated with the observation.To indicate the removal of a variable from a set of variables or a count, we use a superscript; for exampleoris the number of customers (observations) in restaurant (group)seated at tableexcluding. We also assumehas the densityandhas the densityand is conjugate to. The conditional density ofunder mixture componentgiving all data excluding is defined as:



wheredenotes the set of indices of the data item currently associated with dish. For the conjugate case, we could obtain a closed form for this likelihood function. Particularly if emission distributions are Gaussian with unknown mean and covariance, the conjugate prior is a normal-inverse-Wishart distribution (Sudderth, 2006) denoted byand. Given some observations for component, of the mixture (in this case =) from a multivariate Gaussian distribution, the posterior still remains in the normal-inverse-Wishart family and its parameters are updated using:



In practice there are some efficient ways (using Cholesky decomposition) to update these equations and allows fast likelihood evaluation for each data point. Finally, marginalizinginduces a multivariate t-student distribution with  degree of freedom (is the dimension of data points):



Assuming  can be approximated by moment-match Gaussian (Sudderth, 2006).

### Posterior Sampling in CRF

Equations and show how we can produce samples from the prior overand, using the proper likelihood function and how this framework can provide us with necessary tools to sample from the posterior given the observations . In this first algorithm, we sample index andusing a simple Gibbs sampler.

In this algorithm we assumed that emission distributions are Gaussian but using other conjugate pairs is the same. The following list shows a single iteration of the algorithm, but to obtain reliable samples we have to run this many times. Notice that the most computational costly operation is the calculation of the conditional densities. Moreover, the number of events that could happen for each iteration of the algorithm is.

1. Given the pervious state assignment,and:

**Sample t:**

1. For alldo the follow sequentially
2. Compute the conditional density forusing for.
3. Calculate the likelihood for using:



1. Samplefrom the multinomial probability



1. If the sampled value ofis, obtain a sample ofby:



1. Ifthen increment.
2. Update the cached statistics.
3. If a tablebecomes unoccupied delete the correspondingand, if as a result some mixture component becomes unallocated, delete that mixture component too.

**Sample K:**

1. For alldo the following sequentially:
2. Compute the conditional density forusing for.
3. Samplefrom a multinomial distribution:



1. If then increment.
2. Update the cached statistics.
3. If a mixture component becomes unallocated delete that mixture component

Equation is obtained from . From we see the prior probability that takes a previously used value is proportional to and the prior probability of taking a new value is proportional to.The likelihood forgivenfor some previously usedis and the likelihood for is given by . By multiplying these priors and likelihoods we can obtain the posterior distribution . In the same way, and can be obtained by multiplying the likelihoods and priors given by .

### Augmented Posterior Representation Sampler

In the previous algorithm, the sampling for all groups is coupled which makes the derivation of the CRF sampler for certain models difficult (Teh Y. , Jordan, Beal, & Blei, 2006). This happens becausewas integrated out. An alternative approach is to sample. More specifically, we will use and to sample fromandrespectively. This algorithm contains two main steps. First we sample cluster indices (instead of tables and dishes) and then we sampleand. Equation shows in order to sample fromwe should first sample. This completes the second algorithm.

1. Given the pervious state assignment for,andfrom previous step

**Sample Z**

1. For alldo the follow sequentially
2. Compute the conditional density forusing for.
3. Sample  using:



1. If a new componentis chosen, the corresponding atom is initiated using and setand.



1. Update the cached statistics.

**Sampling m**

1. Sampleusing whereis the Stirling number of the first kind. Alternatively, we can sample by simulating a CRF, which is more efficient for large.



**Sampling and**

1. Sampleandusing and respectively.

In this algorithm we have used alternative representation given by . Particularly, we can see . By combining this with we can obtain a conditional prior probability function forand by multiplying it with the likelihoods we can obtain . is the stick-breaking step for the new atom and follows the steps described in section ‎3.1. In particular the third line in is obtained by replacing the remaining stickwith a unit stick. The validity of this approach is shown in (Pitman, 1996).For a proof of look at (Antoniak, 1974). It should be noted that computing this equation is generally very costly and we can alternatively simulate a CRF to sample.

## Applications

Among the several applications of HDP, we will only review two of them in this section. It should be noted that the following section, HDP-HMM, is by itself an application of the general HDP framework.

One of the most cited applications of HDP is in the field of information retrieval (IR) (Teh & Jordan, 2010). A state of the art but heuristic algorithm which is very popular in IR applications is the “term-frequency inverse document frequency” (tf-idf) algorithm. The intuition behind this algorithm is that the relevance of a term to a document is proportional to the number of times that term occurred in the document. However, terms that occur in many documents should be down weighted. It has been shown that HDP provides a justification for this intuition (Teh & Jordan, 2010) and an algorithm based on HDP outperforms all state of the art algorithms.

Another application cited extensively in literature is topic modeling (Teh & Jordan, 2010). In topic modeling, we want to model documents with a mixture model. Topics are defined as probability distributions across a set of words while documents are defined as probability distributions across different topics. At the same time we want to share topics among documents within a corpus. So each document is a group with its own mixing proportions but components (topics) are shared across all documents using an HDP model.

# HDP-HMM

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain (Rabiner, 1989). In the following discussion we will denote the state of the Markov chain at time  with  and the state-specific transition distribution for stateby.The Markovian structure means. Observations are conditionally independent given the state of the HMM and are denoted by.

HDP-HMM is an extension of HMM in which the number of states can be infinite. The idea is relatively simple; at each statewe should be able to go to an infinite number of states so the transition distribution should be a draw from a DP. On the other hand, we want reachable states from one state to be shared among all states so these DPs should be linked together. The result is an HDP. In an HDP-HMM each state corresponds to a group (restaurant) and therefore, unlike HDP in which an association of data to groups is assumed to be known a priori, we are interested to infer this association. The major problem with original HDP-HMM is the state persistence. HDP-HMM has a tendency to make many redundant states and switch rapidly among them (Teh Y. , Jordan, Beal, & Blei, 2006). This problem is solved by introducing a sticky parameter to the definition of HDP-HMM (Fox E. , Sudderth, Jordan, & Willsky, 2011). Equation shows the definition of a sticky HDP-HMM with unimodal emissions.is a sticky hyper-parameter and generally can be learned from data. Original HDP-HMM is a special case with. From this equation we can see for each state (group) we have a simple unimodal emission distribution. This limitation can be addressed using a more general model defined in . In this model, a DP is associated with each state and a model with augmented stateis obtained. Figure 3 shows a graphical representation.







Figure 3-Graphical model of HDP-HMM (Fox E. , Sudderth, Jordan, & Willsky, 2011)

## CRF with Loyal Customers

The metaphor for the Chinese restaurant franchise for sticky HDP-HMM is a franchise with loyal customers. In this case each restaurant has a special dish which is also served in other restaurants. If a customer is going to restaurantthen it is more likely that he eats the specialty dish there. His children also go to the same restaurant and eat the same dish. However, if eats another dish () then his children go to the restaurant indexed byand more likely eat their specialty dish. Thus customers are actually loyal to dishes and tend to go to restaurants where their favorite dish is the specialty.

## Inference Algorithms

### Direct Assignment Sampler

This sampler is adapted from (Fox E. , Sudderth, Jordan, & Willsky, 2011) and (Fox E. , Sudderth, Jordan, & Willsky, 2010). In this section we present the sampler for HDP-HMM with DP emission. This algorithm is very similar to the second inference algorithm for HDP presented in ‎3.4.2. The algorithm is divided into two steps: the first step is to sample the augmented stateand the second is to sample.

In order to sample  we need to have the posterior. By inspecting Figure 3 and using the chain rule we can write the following relationship for this posterior.



The reason that we have summed overin the last line is because we are interested to calculate the likelihood for each state. This equation also tells us that we should first sample the stateand then conditioned on the current state, sample the mixture component for that state. In ‎B.1. we will derive the following relationships for the component of . is written for Gaussian emissions but we can always use the general relationship for an arbitrary emission distribution.







After samplingwe have to samplebut from we see that we need to know the distribution of the number of tables considering dish (). The approach is to first find the distribution of tables serving dish (). In this algorithm, instead of using the approach based on Stirling numbers, we can obtain this distribution by a simulation of the CRF, and then adjust this distribution to obtain the real distribution of considered dishes by tables. To review the reason that this adjustment is necessary, we should notice thatintroduces a non-informative bias to each restaurant so customers are more likely to select the specialty dish of the restaurant. In order to obtain the considered dish distribution we should reduce this bias from the distribution of the served dish. This can be done using an override variable. Supposetables are serving dishin restaurant. If then since the served dish is not the house specialty but if then there is probability that tables are overridden by the house specialty. Suppose that is the override variable with prior , we can write:



The sum of these Bernoulli random variables is a binomial random variable and finally we can calculate the number of tables that considered ordering dishby .

1. Given a previous set of and
2. For all.
3. For each of thecurrently instantiated states compute:

* The predictive conditional distributions for each of the  currently instantiated mixture components for this state, and also for a new component and for a new state.





* The predictive conditional distribution of the HDP-HMM state without knowledge of the current mixture component.



1. Sample:



1. Sample conditioned on:



1. If increase theand transform as



1. Ifincrement.
2. Update the cache. If there is a state withor removeand decrease. If remove the componentand decrease.
3. Sample auxiliary variables by simulating a CRF:
4. For eachsetand. For each customer in restauranteating dish(), sample:



1. Incrementand if increment.
2. For each,sample the override variables in restaurant:



1. Set the number of informative tables in restaurant:



1. Sample:



1. Optionally sample hyper-parametersand.

### Block Sampler

The problem with the direct assignment sampler mentioned in the previous section is the slow convergence rate since we sample states sequentially. The sampler can also group two temporal sets of observations related to one underlying state into two separate states. However, in the last sampling scheme we have not used the Markovian structure to improve the performance. In this section a variant of forward-backward procedure is incorporated in the sampling algorithm that enables us to sample the state sequenceat once. However, to achieve this goal, a fixed truncation level should be accepted which in a sense reduces the model into a parametric model (Fox E. , Sudderth, Jordan, & Willsky, 2011). However, it should be noted that the result is different from a classical parametric Bayesian HMM since the truncated HDP priors induce a shared sparse subset of thepossible states (Fox E. , Sudderth, Jordan, & Willsky, 2011). In short, we obtain an approximation to the nonparametric Bayesian HDP-HMM with maximum number of possible states set to. However, for almost all applications this should not cause any problem if we set reasonably high. The approximation used in this algorithm is the degreeweak limit approximation to the DP (Ishwaran & Zarepour, 2002) which is defined as:



Using is approximated as (Fox, Sudderth, Jordan, & Willsky, Supplement to " A Sticky HDP-HMM with Application to Speaker Diarization", 2010):



Similar to we can write:



And posteriors are (similar to ):



In is the number of transitions from stateto stateand is the same as .

Finally an orderweak limit approximation is used for the DP prior on the emission parameters:



The forward-backward algorithm for the joint sample  andgiven can be obtained by:



The right side of equation has two parts: forward and backward probabilities (Rabiner, 1989).The forward probability includes  and backward probability includes. It seems that the authors in this work approximate the forward probabilities with, and for backward probabilities we have:



As a result we would have (Fox E. , Sudderth, Jordan, & Willsky, 2010) :



where for Gaussian emission for components are given by 

The algorithm is as follows (Fox E. , Sudderth, Jordan, & Willsky, 2010):

1. Given the previous and.
2. For, initialize messages to 
3. Forand compute



1. Sample the augmented state sequentially and start from:

Set andforand

For all compute:



1. Sample augmented state:



1. Increase andand add to the cached statistics.



1. Sample  similar to the previous algorithm
2. Update :



1. For :

* Sample and:



* For  sample:



1. Set and
2. Optionally sample hyper-parametersand.

### Learning Hyper-parameters

Hyper-parameters includingandcan also be inferred like other parameters of the model (Fox, Sudderth, Jordan, & Willsky, 2010).

#### Posterior for

Consider the probability of data to sit behind table:



This equation can be written by considering equation and . From this equation we can say customer table assignment follows a DP with concentration parameter. Antoniak (Antoniak, 1974) has shown that if  then the distribution of the number of unique values of  resulting from draws from has the following form:



Where is the Stirling number of the first kind. Using these two equations the distribution of the number of tables in the restaurantis as follows:



The posterior overis as follows:



The reason for the last line is that is not a function of and therefore can be ignored.

By substitution of  and also by considering that  we obtain:



Finally by considering the fact that we have placed a prior on we can write:

   
Wherecan be either one or zero. For marginal probabilities we obtain:







#### Posterior of

Similar to the discussion for if we want to find the distribution of the unique number of dishes served in the whole franchise we would have. Therefore for the posterior distribution of we can write:



By considering the fact that that prior overiswe can finally write:



And finally for the marginal distributions we have:







#### Posterior of

The posterior foris obtained in a similar way to. We use two auxiliary variablesand and the final marginalized distributions are:







It should be noted that in cases where we use auxiliary variables we prefer to iterate several times before moving to the next iteration of the main algorithm.

#### Posterior of

By definition  and by considering the fact that the prior on is and we can write:



### Online learning

The last two approaches are based on batch learning methodology. One problem with these methods is the need to run the whole algorithm for the whole data set when new data points become available. More than that, for large datasets we might face some practical constraints such as memory size. Another alternative approach is to use sequential learning techniques which essentially let us update models once a new data point becomes available. The algorithm that we are describing here is adapted from (Rodriguez, 2011), but the main idea for a general case is published in (Carvalho, Johannes, Lopes, & Polson, 2010) and (Carvalho, Lopes, Polson, & Taddy, 2010). For Bayesian problems different versions of particle filters are used to replace batch MCMC methods. For further information about particle filters refer to (Cappe, Godsill, & Moulines, 2007). It should be noted that this algorithm is developed for the non-sticky () HDP-HMM with one mixture per state but generalization to sticky HDP-HMM with DP emissions is straightforward.

#### Particle learning (PL) Framework for mixtures

PL is proposed in (Carvalho, Johannes, Lopes, & Polson, 2010) and (Carvalho, Lopes, Polson, & Taddy, 2010) and is a special formulation of augmented particle filters. A general mixture model that PL is supposed to infer can be represented by:



In this set of equations, the first line is the observation equation and the second line is the state evolution which, in case of mixtures, indicates which component is assigned to the observations and

In order to estimate states and parameters we should define an “essential state vector”where andare the state sufficient statistics andis the parameter sufficient statistics (Carvalho, Johannes, Lopes, & Polson, 2010). After observing, particles should be updated based on:





A particle approximation to is:



Using this approximation we can generate propagated samples from the posterior  to approximate. Sufficient statistics can be updated using a deterministic mapping and finally parameters should be updated using sufficient statistics. The main condition for this algorithm to be possible is the tractability ofand. The PL algorithm is as follows:

1. Resample particleswith weights 
2. Propagate new states using
3. Update state and parameter sufficient statistics deterministically 
4. Samplefrom

After one sequential run through the data we can use smoothing algorithms to obtain. The algorithm is repeatedto generatesample paths.

1. Samplefrom output of particle filter

For 

1. Sample 

#### PL for HDP-HMM

In this algorithm we use whereis the state,is the number of states at time,is the number of transitions from stateto stateat time,is the sufficient statistics for state at time  and other variables have the same definitions as previous sections (the last three are auxiliary variables which are used to infer hyper-parameters.).

From and by setting we can write:



In this equation  is the next state  that we have not seen yet. Because of this we should integrate this out by summing over all possibilities. The result is:



After normalization we can write:





The algorithm is as follow:

1. Compute where  and we have:



Where  is similar to .

1. Sample from



1. Propagate the particles to generate:
2. Sample :



1. Update the number of states:



1. Update the sufficient statistics:



1. If :



1. Hyper-parameters update:
2. Sample :



Alternatively we can simulate a CRF instead of computing Stirling numbers.

1. Sample by first sample:



1. Samplingusing auxiliary variables:



1. Resample:



After finishing this inference step, we can optionally smooth the states using an algorithm similar to the one discussed in ‎4.2.4.1. However one drawback of this algorithm is the fact that paths forare coupled together since we integrate outfor the inference algorithm. To improve particle diversity we can sample transition probability and emission parameters explicitly. The smoothing algorithm would be as follows:

1. Sample
2. Sample



1. Sample from the .
2. Sample using a single run of Forward-Backward algorithm applications. Use andas parameters for the Forward-Backward algorithm.

## Applications

One of the applications of HDP-HMM, which is extensively discussed in (Fox E. , Sudderth, Jordan, & Willsky, 2011), is speaker diarization. In this application, we are interested to segment an audio file into time intervals associated with different speakers. If the number of speakers is known a priori a classic HMM can be used and each speaker can be modeled as different states of HMM. However, in real world applications the number of speakers is not known and therefore nonparametric models are a natural solution. It has been shown in (Fox E. , Sudderth, Jordan, & Willsky, 2011) that HDP-HMM can produce results comparable to other state of the art systems.

Another application which is cited as an application of HDP-HMM is word segmentation (Teh & Jordan, 2010). In this problem, we have an utterance and we are interested to segment it into words. Each word can be represented as a state in a HDP-HMM and transition distributions can define a grammar over words.

# Conclusion

In this report, we have investigated hierarchical Dirichlet processes and its application to extend HMMs into infinite HMMs. We also reviewed two inference algorithms for HDP and three inference algorithms for HDP-HMM.

HDP-HMM seems to be a good candidate for many applications which traditionally use HMMs. Using a nonparametric Bayesian approach could help us to automatically learn the complexity of the models from the data instead of relying on heuristic tuning methods. Moreover, the framework can provide a generic and simple approach to organize all models (i.e. different HMMs in a speech recognizer) in a well-defined hierarchy and tie parameters of different models using Bayesian hierarchical methods. The definition of HDP-HMM (with DP emission) can also be altered to include another HDP that links DP emissions of different states together (to link different components of mixture models together.) Another area of work is in inference algorithms. We have presented three algorithms based on Gibbs sampling. It seems block and sequential samplers have some interesting properties that make them reasonable candidates for big datasets. Particularly it seems easy to build a parallel implementation of sequential sampler which can be an important factor for large scale problems. Studying other kinds of inference methods like variational methods or parallel implementation of these algorithms can be a subject of further research.

# Reference

Antoniak, C. (1974). Mixtures of Dirichlet Process with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 1152-1174.

Cappe, O., Godsill, S. J., & Moulines, E. (2007, May). An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. *Proceedings of the IEEE, 95*(5), 899-924.

Carvalho, C. M., Johannes, M., Lopes, H. F., & Polson, N. (2010). Particel Learning and Smoothing. *Statistical Science*, 88-106.

Carvalho, C. M., Lopes, H. F., Polson, N. G., & Taddy, M. A. (2010). Particle Learning for General Mixtures. *Bayesian Analysis*, 709-740.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2010). *Supplement to " A Sticky HDP-HMM with Application to Speaker Diarization".* doi:10.1214/10-AOAS395SUPP

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011). A Sticky HDP-HMM with Application to Speaker Diarization. *The Annalas of Applied Statistics, 5*, 1020-1056.

Ishwaran, H., & Zarepour, M. (2002, June). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 269-283.

Pitman, J. (1996). Random Discrete Distributions Invariant under Size-Biased Permutation. *Advances in Applied ProbabilityApplied Probability Trust*, 525--539.

Rabiner, L. R. (1989, February). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Rodriguez, A. (2011, July). On-Line Learning for the Infinite Hidden Markov. *Communications in Statistics: Simulation and Computation, 40*(6), 879-893.

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

Sudderth, E. B. (2006). chapter 2 of "Graphical Models for Visual Object Recognition and Tracking". Cambridge, MA: Massachusetts Institute of Technology.

Teh, Y., & Jordan, M. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In N. Hjort, C. Holmes, P. Mueller, & S. Walker, *Bayesian Nonparametrics: Principles and Practice.* Cambridge, UK: Cambridge University Press.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). *Hierarchical Dirichlet Processes.* Technical Report 653 UC Berkeley.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 1566-1581.

1. Derivation of HDP Relationships
   1. Stick-Breaking Construction (Teh Y. , Jordan, Beal, & Blei, 2006)

In this lemma we show 

We know:



Using this equation we can write:







We know .Let be a random partition on, define  from and the definition of DP we have:



Using , and we obtain:



Since this is correct for every finite partition of positive integers we conclude that 

Now for a partition and aggregation property of Dirichlet distribution we obtain:



Using the neutrality property of a Dirichlet distribution:



Notice that and also defining  we have:



Using ‎Lemma A.1.1



* 1. Deriving Posterior and Predictive Distributions

Derivation of equation:

Sinceis a random distribution, we can draw from it. Let assumeare i.i.d. draws from. takes value insince is a distribution over.Letbe a finite measureable partition ofand. By using the conjugacy and also definition of DP we can write:



This shows the posterior ofis a DP with concentration parameter equal toand mean of. We also know  so we can write:



Derivation of equation:

Similar to the last proof, letbe a finite measureable partition ofand



Sois a DP with concentration and mean and finally we can write:



Derivation of equation:

For:





From and :



Derivation of Equation is very similar to the above lines and we just need to calculate the expectation ofinstead of.

Deration of Equation

We know that  and also we know that we can writehas two parts; one is a draw form a DP and the other is a draw from a multinomial distribution:



Letbe a partition onwherecontains the wholeexcept spikes located atandcontains spikes. We can write:



And as a result we can write:



Derivation of equation follows the same lines. 

1. Derivation of HDP-HMM Relationships
   1. Derivation of the posterior distribution for





Derivation of:

By using the chain rule and graphical model of Figure 3 we can write:



For and we can write:



For 



Using the fact that has a multinomial distribution,, and by using ‎Lemma B.1.1:



In the equation above, denotes the number of transitions from state to state. Using , , , and after some algebra we can obtain:



Equation can be obtained similar to . Notice in this case that for each state we have a DP and therefore numbers of data points for DP are all data points associated with that state.

Equation can be obtained similar to . The only difference is that we only consider observations assigned to stateand.