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

**David S. HippocampusCoauthor**

Department of Computer Science Affiliation

Cranberry-Lemon University Address

Pittsburgh, PA 15213 *email*

*hippo@cs.cranberry.lemon.edu*

|  |
| --- |
|  |

**Coauthor Coauthor Coauthor**

Affiliation Affiliation Affiliation

Address Address Address

*email email email*

(if needed)

**Abstract**

The Abstract paragraph should be indented ½ inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. **Abstract** must be centered, bold and in point size 12. Two line spaces precede the Abstract. The Abstract must be limited to one paragraph.

**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 applications 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 (but we can share or tie GMM mixture components... so this isn’t true.)

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 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 (DP) [??][??] 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 (HDP) 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 ordera 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 HDP 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. 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 ???. 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 HDP Emission Distributions**

In previous work [] [], 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 defined for the whole HMM.

The final model (without non-emitting states) is given in Eq. and shown graphically in ???. For comparison figure ??? (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 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 :





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**

These instructions apply to everyone, regardless of the formatter being used.

**4.1 Citations within the text**

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BibTeX style unsrt produces this.) As to the format of the references themselves, any standard reference style is acceptable, as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A.Anonymous".

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

Figure 2-(a) xxxxxx(b)xxxxxx

**4.2 Footnotes**

Indicate footnotes with a number in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

**4.3 Figures**

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; artwork should not be hand drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

|  |
| --- |
|  |

Figure 1: Sample Figure Caption

**4.4 Tables**

All tables must be centered, neat, clean and legible. Do not use hand drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Table 1: Sample table title

|  |  |
| --- | --- |
| **Part**  **Description** |  |
| Dendrite | Input terminal |
| Axon | Output terminal |
| Soma | Cell Body (contains cell nucleus) |

**5 Experiments**

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle that the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

**6 Discussion**

Please prepare PostScript or PDF files with paper size “US Letter,” and not, for example, “A4.” The -t letter option on dvips will produce US Letter files.

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

* You can check which fonts a PDF files uses. In Acrobat Reader, select menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program pdffonts which comes with xpdf and is available out-of-the-box on most Linux machines.
* The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf
* LaTeX users:
  + Consider directly generating PDF files using pdflatex (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.
  + Otherwise, please generate your PostScript and PDF files with the following commands:
  + dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
  + ps2pdf mypaper.ps mypaper.pdf
  + Check that the PDF files only contains Type 1 fonts.
* xfig “patterned” shapes are implemented with bitmap fonts. Use “solid” shapes instead.
* The \bbold package almost always uses bitmap fonts. You can try the equivalent AMS Fonts with command
  + \usepackage[psamsfonts]{amssymb}
  + or use the following workaround for reals, natural and complex:
  + \newcommand{\RR}{I\!\!R} %real numbers
  + \newcommand{\Nat}{I\!\!N} %natural numbers
  + \newcommand{\CC}{I\!\!\!\!C} %complex numbers
* Sometimes the problematic fonts are used in figures included in LaTeX files. The ghostscript program eps2eps is the simplest way to clean such figures. For black and white figures, slightly better results can be achieved with program potrace.
* MSWord 2007 and Windows users (via PDF file):
  + Install the Microsoft Save as PDF Office 2007 Add-in from
  + http://www.microsoft.com/downloads/details.aspx?displaylang=en&familyid=4d951911-3e7e-4ae6-b059-a2e79ed87041
  + Select "Save or Publish to PDF" from the Office or File menu
* MSWord and Mac OS X users (via PDF file):
  + From the print menu, click the PDF drop-down box, and select "Save as PDF…"
* MSWord and Windows users (via PS file):
  + To create a new printer on your computer, install the AdobePS printer driver and the Adobe PostScript Printer Description (PPD) file from
  + <http://www.adobe.com/support/downloads/detail.jsp?ftpID=204>
  + *Note:* You must reboot your PC after installing the AdobePS driver for it to take effect.
  + To produce the ps file, select "Print" from the MS app, choose the installed AdobePS printer, click on "Properties", click on "Advanced."
  + Set “TrueType Font” to be “Download as Softfont”
  + Open the “PostScript Options” folder
  + Select “PostScript Output Option” to be “Optimize for Portability”
  + Select “TrueType Font Download Option” to be “Outline”
  + Select “Send PostScript Error Handler” to be “No”
  + Click “OK” three times, print your file.
  + Now, use Adobe Acrobat Distiller or ps2pdf to create a PDF file from the PS file. In Acrobat, check the option “Embed all fonts” if applicable.

If your file contains Type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

**6.1 Margins in LaTeX**

Most of the margin problems come from figures positioned by hand using \special or other commands. We suggest using the command \includegraphics from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below

\usepackage[dvips]{graphicx} ...

\includegraphics[width=0.8\linewidth]{myfile.eps}

or

\usepackage[pdftex]{graphicx} ...

\includegraphics[width=0.8\linewidth]{myfile.pdf}

for .pdf graphics. See section 4.4 in the graphics bundle documentation (http://www.ctan.org/texarchive/macros/latex/required/graphics/grfguide.ps)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the \- command.

**Acknowledgments**

Use unnumbered third level headings for the acknowledgments. All acknowledgements go at the end of the paper. Do not include acknowledgements in the anonymized submission, only in the final paper.

**References**

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to ‘small’ (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited 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.