Infinite Hierarchical Hidden Markov Models

Katherine A. Heller
Engineering Department
University of Cambridge
Cambridge, UK
heller@gatsby.ucl.ac.uk

Yee Whye Teh and Dilan Görür
Gatsby Unit
University College London
London, UK
{ywteh,dilan}@gatsby.ucl.ac.uk

Abstract

In this paper we present the Infinite Hierarchical Hidden Markov Model (IHHMM), a nonparametric generalization of Hierarchical Hidden Markov Models (HHMMs). HHMMs have been used for modeling sequential data in applications such as speech recognition, detecting topic transitions in video and extracting information from text. The IHHMM provides more flexible modeling of sequential data by allowing a potentially unbounded number of levels in the hierarchy, instead of requiring the specification of a fixed hierarchy depth. Inference and learning are performed efficiently using Gibbs sampling and a modified forward-backtrack algorithm. We present encouraging results on toy sequences and English text data.

1 Introduction

Hierarchical Hidden Markov Models (HHMM) are multiscale models of sequences where each level of the model is a separate Hidden Markov Model (HMM) emitting lower level HMMs in a recursive manner [Fine et al., 1998]. HHMMs are well-suited to the multiscale nature of many naturally occurring sequential data, and have been successfully applied across a wide spectrum of domains, including language and speech processing [Fine et al., 1998], information extraction [Skounakis et al., 2003], video structure discovery [Xie et al., 2002] and activity detection and recognition [Nguyen et al., 2005].

Inference and learning in the HHMM are carried out in a straightforward manner using extensions of both the standard forward-backward algorithm and the Baum-Welch algorithm. However, learning the model structure of HHMMs is significantly harder, due to the multitude of local optima and the inherent non-identifiability of such a flexible model.

We present a nonparametric Bayesian approach to HHMMs. Rather than assuming a hierarchy of finite depth and attempting to learn the appropriate depth, our model assumes an infinite number of levels in the HHMM at the outset. The assumption made in our model instead is that in any finite length sequence only a finite number of state transitions (over a finite number of levels) will be performed. This assumption allows for a computationally and statistically tractable alternative to model selection in HHMMs.

The graphical model representation of our infinite hierarchical HMM (IHHMM) consists of an infinite number of levels, where each level is a sequence of latent variables dependent on the level above, and where the observed sequence lies at the bottom level. This results in an approach to nonparametric Bayesian modeling that differs significantly from previous work, in that previous work has dealt only with models with a finite hierarchy. See Section 6 for further discussion.

The structure of the rest of this paper is as follows. In section 2 we review HHMMs. In section 3 we introduce and define the infinite hierarchical HMM and place it within the context of previous HHMM models. Section 4 derives a block Gibbs sampling based inference algorithm for the IHHMM, while section 5 presents the results of using the IHHMM on toy sequence and English text data, comparing it with previous approaches. Related work, including a variety of related method available for grammar learning, is discussed in section 6. Lastly we conclude with a discussion in section 7.

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

2 Hierarchical Hidden Markov Models

Hierarchical Hidden Markov Models (HHMMs) [Fine et al., 1998] are used for modeling the hierarchical structure of data found in many application domains, such as natural language [Fine et al., 1998] or music [Weiland et al., 2005]. HHMMs are an extension of HMMs where instead of being restricted to emitting single observations, the states of the HHMM are themselves HHMMs, which contain substates, and can emit strings of observations. States that emit strings of observations are called "abstract states", while those that have single emissions are called "production states". Internal HHMMs can be called recursively from abstract states, such that control is returned to the abstract state when it has completed running. Every HHMM can be represented with a standard HMM where the state of the HMM consists of the production state as well as the abstract states higher up the hierarchy of the HHMM. However, this results in an exponential number of states and parameters, and lacks the informative hierarchical structure of the HHMM and the ability to reuse the same internal model components in differing situations.

Murphy and Paskin [2001] show that the HHMM can be represented as a dynamic Bayesian network (DBN) which allows the application of a whole spectrum of learning and inference methods to the HHMM. In the DBN representation of the HHMM the observed symbol at time t, O_t, depends on the state variables at time t at all levels of the hierarchy $\mathbf{Q}_t = (Q_t^1 \ldots Q_t^L)$, where level L is the top level. These state variables represent the sequence of abstract states leading from the root node to the production state for O_t in the HHMM. Indicator variables F_t^l control completion of the HHMM at level l and time t and enforce higher level HMMs transitioning only when ones at lower levels have completed.

Our Infinite Hierarchical Hidden Markov Model bears close resemblence to the HHMM described above. In addition to the infinite levels of the hierarchy, in certain aspects we opted to simplify the description of our model relative to that of Murphy and Paskin [2001], largely for clarity's sake. The resulting differences will be discussed in section 3.1.

3 The Infinite Hierarchical Hidden Markov Model

The Infinite Hierarchical Hidden Markov Model (IHHMM) is a nonparametric generalization of the HHMM which allows the HHMM hierarchy to have a potentially infinite number of levels. Let y_t be the observation at time t, s_t^l the state at time t and level $l = 1, 2, \ldots$, and z_t^l a binary variable indicating whether there is a completion of the HHMM at level $l-1$ right before time t. There is a state transition at level l exactly when the HHMM at level $l-1$ completes, thus z_t^l also indicates presence of a state transition from s_{t-1}^l to s_t^l (absence of a transition means $s_t^l = s_{t-1}^l$). For simplicity, all state and observation variables are discrete with finite cardinality. The conditional probability of z_t^l is:

$$P(z_t^l = 1 | z_t^{l-1}) = \begin{cases} \alpha_l & \text{if } z_t^{l-1} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where α_l is a parameter which controls the chance of a state transition at level l and $z_t^0 = 1$ for simplicity. Notice that there is an opportunity to transition at level l only if there was a transition at level $l-1$—this imposes the hierarchy on the state transitions of the IHHMM.

The structure of the variables z_t^l implies a number of properties regarding state transitions in the IHHMM. Firstly, the number of transitions at level $l-1$ before a transition at level l occurs is geometrically distributed with parameter α_l, which has a mean of $1/\alpha_l$. This implies that the expected number of time steps for which a state at level l persists in its current value is $1/\prod_{k=1}^{l} \alpha_k$. Thus we see that states at higher levels persist longer—we expect these states to capture longer range dependencies in the IHHMM, while states at levels below capture shorter range ones. Secondly, the first non-transitioning level at time t, L_t, has the following distribution:

$$P(L_t = l) = (1 - \alpha_l) \prod_{k=1}^{l-1} \alpha_k \tag{2}$$

Thus $1 - \alpha_l$ is the hazard rate at step l of L_t. If all $\alpha_l = \alpha$ are equal, L_t is geometrically distributed as well with parameter $1 - \alpha$. Note that L_t can take on arbitrarily large values. Thus the IHHMM allows for a potentially infinite number of levels in the hierarchy with a decreasing number of state transitions at higher levels. We take the top level of the hierarchy corresponding to a finite sequence of observations $(y_1, \ldots, y_T)$ to be the first level, L_*, where $z_t^{L_*} = 0, \forall t$. Note that as the number of time steps increases L_* increases as well.

The remainder of the generative process for the observation y_t given $z_t^{1:\infty}$ now proceeds quite similarly to the HHMM, where states s_t^l are generated from levels $L_t - 1$ down to 1:

$$P(s_t^l = a | s_{t-1}^l = b, s_t^{l+1} = c, z_t^l = 1) = A_{abc}^l \tag{3}$$

where A^l is a 3D state transition matrix at level l. States at levels $l \geq L_t$ persist: $s_t^l = s_{t-1}^l$. We assume

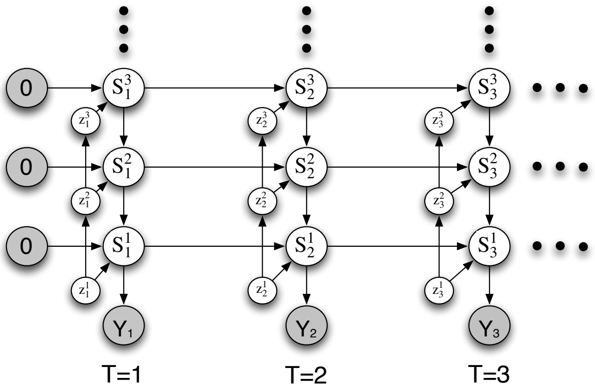

Figure 1: Graphical model for the IHHMM. Parameters of the model has been omitted for clarity.

that at time $t = 0$ states at all levels start at a special begining-of-sequence value $s_0^l = 0$. Finally, the data is generated from an emissions matrix:

$$P(y_t = o|s_t^1 = a) = E_{oa} \qquad (4)$$

The graphical model for the IHHMM is given in figure 4. We place beta priors on α_l, and symmetric Dirichlet priors on the transition A and emission E probabilities.

3.1 Relation to the HHMM

Most obviously the IHHMM is a nonparametric extension of the HHMM and allows for an unbounded hierarchy depth. This is a valuable property because the best number of levels to include in the hierarchy of an HHMM is usually unknown a priori. Utilizing a nonparametric method which allows the number of levels to be inferred directly from the data can improve the performance of the model while avoiding costly model comparisons.

The IHHMM also differs from the HHMM in a number of other respects. We did not require the chance of transition z_t^l to depend on the state s_{t-1}^{l-1}. In other words, the completion of an internal HHMM in the model is not governed by the internal HHMM itself (e.g. by transitioning into a termination state), but rather by an independent process. We allow states s_t^l to always depend on s_{t-1}^l, regardless of whether there were transitions at higher levels of the IHHMM, whereas in the HHMM s_t^l only depends on s_t^{l-1} if there was a transition into s_t^{l-1}. This might be a better assumption for certain kinds of data (for example a video sequence), but not for others (for example the beginning of sentences might be modeled better by not being conditioned on the end of the previous sentence). Finally, we allow states s_t^l to depend only on s_t^{l-1} but not on higher level states (similarly for the observations y_t). All of these differences can be reconciled with the HHMM at additional complexity in specifying the IHHMM model architecture, and at a cost of complicating the exposition and detracting attention from the main idea of the IHHMM. We show some examples of using some of these differences in Section 5.

4 Inference and Learning

We perform inference and learning in the IHHMM using Gibbs sampling and a modified forward-backtrack algorithm. Our inference algorithm iterates between sampling the state values of the IHHMM with fixed parameters and then relearning the parameters conditioned on the state values. These two steps are explained in detail below.

Sampling State Values with Fixed Parameters: Resampling the state values at all levels of the IHHMM conditioned on the current sampled parameter values is done by Gibbs sampling the state trajectory at each level of the hierarchy, starting at the bottom, conditioned on all the other levels (though only the levels immediately above and below are relevant) using a modified forward-backtrack algorithm. For a given level of the hierarchy, l, we compute forward messages, starting from $t = 1$ and going forward to $t = T$:

$$\begin{aligned}\Gamma_t(s_t^l, z_t^l) &\triangleq P(s_t^l, z_t^l, s_{1:t}^{l-1}, z_{1:t}^{l+1} | s_{1:t}^{l+1}, z_{1:t}^{l-1}) \\ = &\sum_{s_{t-1}^l, z_{t-1}^l} \Gamma_{t-1}(s_{t-1}^l, z_{t-1}^l) P(z_t^l | z_t^{l-1}) P(z_t^{l+1} | z_t^l) \\ &\times P(s_t^l | s_{t-1}^l, s_t^{l+1}, z_t^l) P(s_t^{l-1} | s_{t-1}^{l-1}, s_t^l, z_t^l) \qquad (5)\end{aligned}$$

These terms can all be simply computed. The first term is the preceding message Γ_{t-1}, the next two terms are given by equation (1), and the last two terms are given by the state transition matrix, A. For level $l = 1$ we replace the transition term $P(s_t^{l-1}|s_{t-1}^{l-1}, s_t^l, z_t^l)$ with the emission probability $P(y_t|s_t^1)$. A backward pass can now resample s_t and z_t, starting at $t = T$ and going back to $t = 1$:

$$\begin{aligned}s_t^l, z_t^l \sim &P(s_t^l, z_t^l | s_{t+1}^l, z_{t+1}^l, s_{1:T}^{l\pm1}, z_{1:T}^{l\pm1}) \\ &\propto \Gamma_t(s_t^l, z_t^l) P(s_{t+1}^l | s_t^l, s_{t+1}^{l+1}, z_{t+1}^l) \qquad (6)\end{aligned}$$

The last term is easy to compute and given by the transition matrix A (or emission matrix if $l = 1$). When the top level is reached it is also resampled by creating a new level above it with all states having a state value of 1 (this property always defines the top level). If state transitions did occur after resampling the top level, then the level above becomes the new top level and is retained. Analogously if the level below the current top level has no state transitions it becomes the new top level. Thus the number of levels in the hierarchy can grow (or shrink) in an unbounded manner during sampling.

Sampling Parameters Given the Current State: Parameters are initialized as draws from the prior. State transition and emission variables are given symmetric Dirichlet priors. Given the state trajectories sampled in the previous step counts of state transitions and emissions can be computed and used to calculate the posterior distribution (also Dirichlet) from which the transition and emission parameters are redrawn.

Notice that the running time of each iteration of this Gibbs sampler is $O(TL_*)$. This is in contrast with the $O(T^3)$ algorithm of Fine et al. [1998] and the $O(TK^{L_*})$ algorithm of Murphy and Paskin [2001]. However it may take a larger number of iterations to converge.

Predicting New Observations:
Given the current state of the IHHMM (consisting of the state variables $s^{1:L_*}_{1:T}$, transition indicators $z^{1:L_*}_{1:T}$, and parameters $\alpha_{1:L_*}$, $A^{1:L_*}$ and E), we can efficiently compute the probability of each possible symbol being the next observation y_{T+1} using dynamic programming. This probability requires summing over the infinitely many states $s^{1:\infty}_{T+1}$ and transition indicators $z^{1:\infty}_{T+1}$. We will show that we can compute it using $O(L_*)$ computations. Consider the following set of recursions starting from $l = L^* - 1$ and going down to $l = 1$:

$$\begin{aligned}
&\Phi_l(s^l_{T+1}) \triangleq P(s^l_{T+1}|z^{l-1}_{T+1}=1, s^{1:L_*}_{1:T}, z^{1:L_*}_{1:T}) \\
=&P(z^l_{T+1}=0|z^{l-1}_{T+1}=1)P(s^l_{T+1}|s^l_T, s^{l+1}_{T+1}, z^l_{T+1}=0)+ \\
&\sum_{s^{l+1}_{T+1}} P(z^l_{T+1}=1|z^{l-1}_{T+1}=1)P(s^l_{T+1}|s^l_T, s^{l+1}_{T+1}, z^l_{T+1}=1) \\
&\quad \times P(s^{l+1}_{T+1}|z^l_{T+1}=1, s^{1:L_*}_{1:T}, z^{1:L_*}_{1:T}) \\
=&(1-\alpha_l)\mathbb{I}(s^l_{T+1}=s^l_T)+ \\
&\alpha_l \sum_{s^{l+1}_{T+1}} P(s^l_{T+1}|s^l_T, s^{l+1}_{T+1}, z^l_{T+1}=1)\Phi_{l+1}(s^{l+1}_{T+1}) \qquad (7)
\end{aligned}$$

where we have suppressed dependence on the parameters for clarity, and $\mathbb{I}(\cdot) = 1$ if its argument is true and 0 otherwise. The recursion can be initialized at the top level L_* with $\Phi_{L_*}(s^{L_*}_{T+1}) = 1/N_{L_*}$ where N_{L_*} is the number of states at level L_*. This is because level L_* is the first level for which no transition has been encountered, thus integrating out A^{L_*} the state $S^{L_*}_{T+1}$ will be uniformly distributed among the N_{L_*} possible states. Finally, the probability of observing y_{T+1} can be computed from $\Phi_1(s^1_{T+1})$:

$$\begin{aligned}
&P(y_{T+1}|s^{1:L_*}_{1:T}, z^{1:L_*}_{1:T}) \\
=&P(z^1_{T+1}=0)P(y_{T+1}|s^1_{T+1}=s^1_T)+ \\
&P(z^1_{T+1}=1)\sum_{s^1_{T+1}} P(y_{T+1}|s^1_{T+1})\Phi_1(s^1_{T+1}) \qquad (8)
\end{aligned}$$

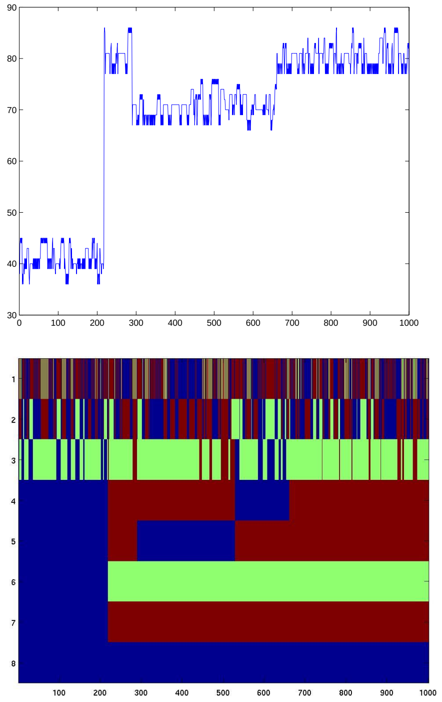

Figure 2: **top:** Data generated from a IHHMM model. **bottom:** The sampled states of the IHHMM used to generate the data (levels in reverse order with emission level at top). To generate data from the model with specified hyperparameters, we start at time $t = 0$ with a single level, with state $= 1$ and sample indicator variables of each level for time $t = 1$ until we sample a zero indicator for a level. Given the number of turned on levels, we sample the states of each level starting from the top, and finally generate the observation. We generate the following observations similarly by first initiating the levels in the hierarchy, then sampling states of each level and finally emitting the observation. Note that the big jumps in the dataset correspond to state transitions at higher levels of the hierarchy.

5 Demonstrations

We begin by generating data from the IHHMM. The IHHMM model allows us to capture the underlying hierarchical structure in the data. To have a better intuition about what we mean by ”underlying hierarchical structure in the data”, we have generated data from the model, shown in Figure 2. Here we allow observations to depend on transitions at all levels of the generated hierarchy. We see that there are low-level fluctuations in the data as well as fewer larger jumps. The state transitions at the lower levels of the hierarchy produce the fine scale fluctuations while higher level structure produces bigger jumps. Note that mod-

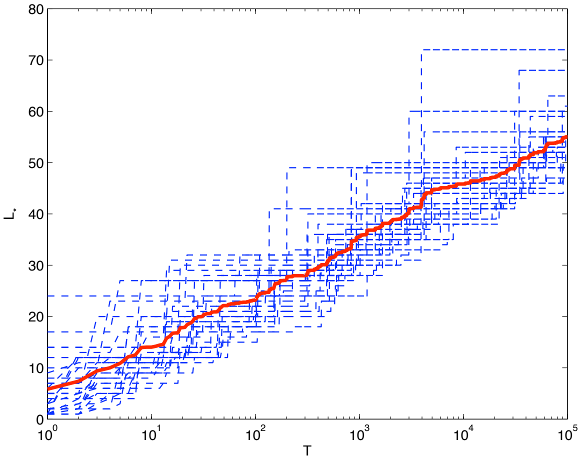

Figure 3: Number of levels L_* used by the IHHMM as a function of sequence length T. Each of 30 dashed blue lines is an independent run with all $\alpha_l = .8$ while the bold red line is the average.

eling this data with an HMM would require exponentially many additional states than is necessitated in an HHMM.

We also looked at the rate at which the model increases the number of levels it used. In Figure 3 we plotted the number of levels used by the model as a function of the sequence length T. We see that in this case where $\alpha_l = .8$, the model increases its number of levels logarithmically in T.

To demonstrate that the model can successfully capture the structure in hierarchical data, we performed experiments on two sets of toy data. Both toy datasets consist of concatenation of an increasing and decreasing series of integers. The first dataset consists of repeats of integers increasing from 1 to 7, followed by repetitions of integers decreasing from 5 to 1, repeated twice. The second dataset is the first data concatenated with another series of repeated increasing and decreasing sequences of integers. We use 7 states in the model for this data, at all levels of the IHHMM. Note in Figure 4(a) that the model needs two levels of hierarchy to express the structure in the data. The state transition on the second level corresponds to the transition from repetitions of increasing integers to the repetitions of decreasing numbers. In Figure 4(b), the model still finds the same structure on the same part of the data and adds another level of hierarchy to model the second half of the data part of which comes from a different series. We computed the predictive log probability of the next integer in the sequence for both the IHHMM and vanilla (Gibbs sampling based) HMM over 10 sequences, each being run 10 times. Each prediction averaged over 20 samples taken after burn in. The mean predictive likelihood for the HMM was 0.25 versus 0.31 for the IHHMM. We also computed predictive likelihood scores for a fixed level HHMM and found results to be comparable with the IHHMM (0.30 for 2-4 levels of hierarchy). This is to be expected since most of the structure of the data can be captured with the first two levels of the IHHMM. On points where there is a transition at the third level, we did find that the IHHMM outperforms the two-level HHMM, but there are very few of these points since the data set size is small. In order to eliminate the possiblity of sampling noise in these results we plan to perform a deeper evaluation when running the IHHMM on large scale application data.

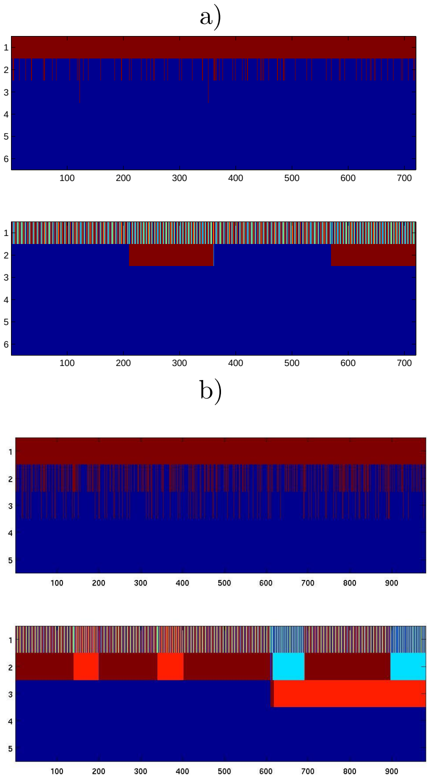

Figure 4: Results on toy number data. The first and third plot from the top show the z values (determining whether there is a state transition; red means 1 while blue means 0) for each of two data sets. The second and bottom plots show the state values for the final sample of IHHMM inference (one color for each state value). The model successfully captures the hierarchical structure in the data. A higher level initiation (i.e. transition from the default state to another state at a higher level) is possible only when its indicator variable is turned on.

We also ran the IHHMM on spectral data from Han-

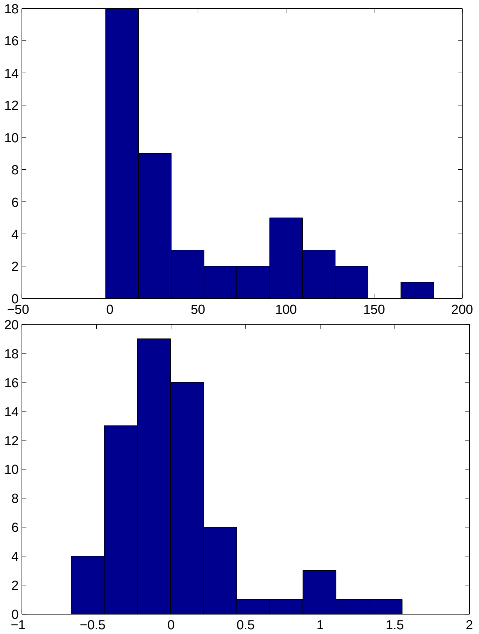

Figure 5: Alice in Wonderland letters data set. Top: The difference in log predictive likelihood between the IHHMM and a standard HMM learned using the EM algorithm. Bottom: The difference in log predictive likelihood between the IHHMM and a one level HMM learned using Gibbs sampling. The long tail signifies that there are letters which can be better predicted by learning long range correlations through the use of higher hierarchical levels and infered by the IHHMM model.

del's Hallelujah chorus. The data is shown at the bottom of figure 6. For this dataset, the model learns a single level of the hierarchy since it infers that adding more levels does not benefit inference for this particular data set. We can see in the middle (state) plot in figure 6 that the structure corresponding to the high frequency regions of the data are discovered by the model. These high frequency regions correspond to two sung 'Hallelujahs' in the music. We also note that the IHHMM does not discover higher levels of structure in this data, implying that the model in its current form (with discrete emissions, etc.) might not be appropriate for this type of data. A similar result was also found when running our algorithm on other pieces of music.

Lastly, we used the IHHMM model to infer characters in written text. We used characters from the first two chapters of 'Alice in Wonderland' as given by the data set from Project Gutenberg. We ran the IHHMM algorithm over the first N characters (where N ranged between 1000 and 3000) and then predicted on the N+1st character. We used 7 latent states per level, an alpha value of 0.5 and performed 200 iterations of Gibbs sampling per run. The value of alpha was held fixed in order to improve mixing of the sampler (and thereby removing some flexibility from the model). Our predictions averaged over 20 samples taken after a burn in period. We also ran a standard HMM using Gibbs sampling on this data (analogous to alpha always being 0 in the IHHMM), and a standard HMM which is learned using the Expectation-Maximization algorithm (EM). A histogram of the differences in log predictive likelihood between the IHHMM and each of these two competing algorithms are given in figure 5 for many runs of the algorithms and subsequent predictions of the next character in the text. The IHHMM obviously performs much better than the HMM with EM (the maximum value over 5 initializations was taken for each prediction). The histogram for the IHHMM versus the one level Gibbs HMM shows a long tail where the IHHMM is performing significantly better on some letters. This is due to the fact that while many letters can be well predicted with one level of hierarchy structure, there some letters can be better predicted by learning long range correlations through the use of higher hierarchical levels and infered by the IHHMM model. The heavy tails in both plots are significant, but the EM HMM often performs quite poorly due to the fact that it is a maximum likelihood based algorithm and is prone to overfitting. The mean differences in both plots are positive, demonstrating that the IHHMM gives superior performance on this data.

6 Related Work and Extensions

The problem of modelling multi-scale structure in sequential data has been well studied. We describe a number of previously proposed models besides the HHMM and contrast them with the IHHMM. In addition, the version of the IHHMM presented in this paper is of the most basic architecture. Inspired by some of the related work, we also describe a number of variations and extensions within the same IHHMM framework. We expect different variations to be suitable for different types of data.

6.1 Stochastic Grammars

Probabilistic context free grammars form a large class of approaches to multi-scale structure learning that is especially prevalent in the linguistics and bioinformatics literature. Grammar induction methods include bottom-up approaches which reduce redundancies by identifying common subsequences [Stolcke and Omohundro, 1994], as well as nonparametric Bayesian approaches using MCMC or variational inference [Liang et al., 2007, Johnson et al., 2007, Finkel et al., 2007].

Grammar-based approaches define distributions over

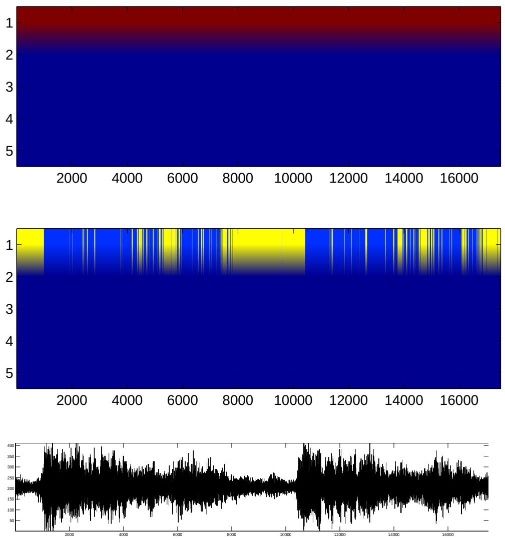

Figure 6: The IHHMM run on spectral data from Handel's Hallelujah chorus. Top plot are z values (red is 1, blue is 0), middle plot is state values, and bottom plot is the original data. Time is on the x-axis. The IHHMM discovers two high frequency regions in the data (corresponding to two hallelujahs in the music - shown on top).

finite (but potentially arbitrarily long) sequences, while our model gives a well-defined distribution over infinitely long sequences. Another difference is that each state in a grammar typically describes a "phrase" or specific sequence of states in the level below, while in the IHHMM each state describes a Markov chain of states below.

A significantly different variation on the IHHMM is inspired by the phrasal structure of stochastic grammars. Consider the distribution over state sequences in a level of the IHHMM given a state of the level above. We can use a factorized distribution where each state is independent of other states in the sequence *but dependent on its location in its sequence.* This allows each state in the IHHMM to encode for a "phrase" or specific sequence of states in the level below.

6.2 Nonparametric Bayesian Models

The IHHMM is an example of a nonparametric Bayesian model for sequential data. Related models include the infinite HMM [Beal et al., 2002], which has a single state variable per time point with infinite cardinality, and the infinite factorial HMM [Van Gael et al., 2008], which has an infinite collection of independently evolving Markov chains. On the other hand, the IHHMM has an infinite number of levels, each being a Markov chain dependent on the chain above. The relationship among these three nonparametric time series models highlights an interesting development in the nonparametric Bayes literature represented by the IHHMM. Specifically, the novel mechanism to derive the IHHMM allows for a graphical representation with unbounded depth while maintaining tractability, while past models have simpler graphical representations with only a few levels.

The obvious extension to the model presented here is to make the number of states at each level infinite as well a la the infinite HMM. We can achieve this in a straightforward manner by using a hierarchical DP to share the set of next states among the transition probabilities [Teh et al., 2006].

6.3 Other Extensions

Other straightforward extensions on the IHHMM are to allow more complex dependencies within the model. For example, the next state can not just depend on the previous state and the state on the level above, but also on states further up the hierarchy as well. Another variation is to allow higher order Markov chains where each state depends on a number of previous states. Yet another variation is to allow the probability of whether there is a transition to depend on the current states of the IHHMM, rather than using a state-independent probability. We can also replace the emission distribution over discrete symbols with other distributions, e.g. mixture of Gaussians to model continuous data.

6.4 Efficient Inference Algorithms

For practical applications of the IHHMM, another important avenue of research is to develop efficient inference algorithms that can be applied to large data sets. Effective initializations need to be developed. For example, we can develop greedy layer-wise initialization where we train the IHHMM with just one layer, fix the first layer while we train the second layer, followed by the third, fourth etc. This allows the model to learn lower level structure first (typically easier) before higher level structure. Better MCMC samplers or variational approximations need to be developed, as well as MAP inference using a Viterbi-like algorithm. One possible approach might be to use a slice sampler to dynamically limit the number of levels of the hierarchy, so that the inside-outside algorithm can be applied on the finite number of levels, thus allowing for potentially much larger steps than the level-wise Gibbs sampler proposed in this paper.

7 Discussion

We proposed the infinite hierarchical hidden Markov model, a hierarchical model for sequential data that operates on multiple time scales. The IHHMM is a nonparametric Bayesian model that allows for an infinite number of levels in the hierarchy. We proposed a Gibbs sampler for the IHHMM that samples each level efficiently using a modified forward-backtrack algorithm.

Nonparametric Bayesian models have recently caught the attention of the machine learning and statistics communities as elegant alternatives to traditional structure learning approaches. The first heavily utilized nonparametric Bayesian models were the Dirichlet process and other clustering-based models, which, from a structure learning standpoint, assume a finite number of variables with infinite cardinality [Rasmussen, 2000, Teh et al., 2006]. The Indian buffet Process takes the next step, being a latent feature model with an infinite number of independent latent features, but only a finite number of which will be active for any given data item [Griffiths and Ghahramani, 2006]. The IHHMM represents the logical next step, whereby an infinite number of dependent latent variables interact in a complex fashion to produce the data, but only a finite amount of computation is needed to perform learning and inference in the model. This is not to say that the IHHMM subsumes these or other previous models. Instead, we wish to highlight an interesting development in the nonparametric Bayesian literature.

There are many exciting avenues for future research. Firstly it will be interesting to see how the model performs for various applications on large scale data, such as video segmentation, speech recognition, and synthetic music compositions. We have described a variety of potential extensions and variations of the model, as well as highlighted the need for more efficient inference algorithms. We are particularly keen on the grammatical variation of the IHHMM. In most of our experiments we have found that the Markov structure within each level was able to capture the regularities in the level below. The grammatical variation assumes that each state is independent given the higher level state, thus forces all regularities to be captured by the hierarchical structure. We believe this should allow the model to capture interesting and deep hierarchical structure inherent in many data sets.

References

M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32:41–62, 1998.

J. R. Finkel, T. Grenager, and C. D. Manning. The infinite tree. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Neural Information Processing Systems*, 2006.

M. Johnson, T. L. Griffiths, and S. Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2007.

K. Murphy and M.A. Paskin. Linear time inference in hierarchical HMMs. In *Neural Information Processing Systems*, 2001.

N. Nguyen, D. Phung, S. Venkatesh, and H.H. Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. In *InternationalConference on Computer Vision and Pattern Recognition*, 2005.

C. Rasmussen. The infinite Gaussian mixture model. In *Neural Information Processing Systems*, 2000.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1994.

M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden Markov models for information extraction. In *International Joint Conference on Artificial Intelligence*, 2003.

A. Stolcke and S. Omohundro. Inducing probabilistic grammars by bayesian model merging. In *Grammatical Inference and Applications*, pages 106–118. Springer, 1994.

Y. W. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. In *Neural Information Processing Systems*, 2006.

J. Van Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

M. Weiland, A. Smaill, and P. Nelson. Learning musical pitch structures with hierarchical hidden Markov models. Technical report, University of Edinburgh, 2005.

L. Xie, S. Chang, A. Divakaran, and H. Sun. Learning hierarchical hidden Markov models for video structure discovery. Technical report, Columbia University, 2002.