Variational Infinite Hidden Conditional Random Fields

Konstantinos Bousmalis, *Student Member, IEEE,* Stefanos Zafeiriou, *Member, IEEE,* Louis-Philippe Morency, *Member, IEEE,* Maja Pantic, *Fellow, IEEE,* and Zoubin Ghahramani, *Member, IEEE*

Abstract—Hidden Conditional Random Fields (HCRFs) are discriminative latent variable models which have been shown to successfully learn the hidden structure of a given classification problem. An Infinite Hidden Conditional Random Field is a Hidden Conditional Random Field with a countably infinite number of hidden states, which rids us not only of the necessity to specify a priori a fixed number of hidden states available but also of the problem of overfitting. Markov chain Monte Carlo (MCMC) sampling algorithms are often employed for inference in such models. However, convergence of such algorithms is rather difficult to verify, and as the complexity of the task at hand increases the computational cost of such algorithms often becomes prohibitive. These limitations can be overcome by variational techniques. In this paper, we present a generalized framework for infinite HCRF models, and a novel variational inference approach on a model based on coupled Dirichlet Process Mixtures, the HCRF–DPM. We show that the variational HCRF–DPM is able to converge to a correct number of represented hidden states, and performs as well as the best parametric HCRFs —chosen via cross–validation— for the difficult tasks of recognizing instances of agreement, disagreement, and pain in audiovisual sequences.

Index Terms—nonparametric models, discriminative models, hidden conditional random fields, dirichlet processes, variational inference

1 Introduction

Hidden Conditional Random Fields (HCRFs) [1] are discriminative models that learn the joint distribution of a class label and a sequence of latent variables conditioned on a given observation sequence, with dependencies among latent variables expressed by an undirected graph. HCRFs do not only learn hidden states that discriminate one class label from all the others, but also structure that is shared among labels. A limitation of the HCRFs is that finding the optimal number of hidden states for a given classification problem is not always intuitive, and learning the correct number of states is often a trial–and–error process involving cross–validation, that can be very computationally expensive. Even then, one has to be careful to avoid the trap of overfitting. These limitations motivated our proposal of an infinite HCRF model that allows its number of states to grow as necessary to fit the data.

Over the past decade, nonparametric methods have been successfully applied to many existing graphical models, allowing them to grow the number of latent states as necessary to fit the data. A prominent and well–studied example is the Infinite Hidden Markov Model (IHMM or HDP–HMM) [2], [3], [4], a Hierarchical Dirichlet Process–driven HMM with an infinite number of potential hidden states. Other notable examples include the first such model, the Infinite Gaussian Mixture Model [5], but also the more recent Infinite Factorial Hidden Markov Model [6], the Infinite Latent Conditional Random Fields[1] [8], the Mixture Dirichlet Process Markov Random Field (MDP–MRF) [9] and the Infinite Hidden Markov Random Field Model (IHMRF) [10]. Hidden Conditional Random Fields (HCRFs) are related to Hidden Markov Random Fields, in that both employ a layer of latent variables with an undirected graph specifying dependencies between those variables. However, there is the important difference that HMRFs model a joint distribution over latent variables and observations, whereas the HCRF is a discriminative sequential model with latent variables.

Infinite HCRFs were first presented in [11] and since exact inference for such a model with an infinite number of parameters is intractable, inference was based on a Markov chain Monte Carlo (MCMC) sampling algorithm. Although MCMC algorithms have been successfully applied on numerous applications, they have some significant drawbacks: they are notoriously slow to converge, it is hard to verify their convergence, and they often don't scale well to larger datasets and higher model complexity. Most importantly, the model presented in [11] is better suited for handling solely discrete features.

In this work, we consider a deterministic alternative to MCMC sampling algorithm for infinite HCRFs with a variational inference [12] approach. Variational inference will allow the model to converge faster, verify convergence and scale without a prohibitive computational cost. The model we

- *K. Bousmalis, S. Zafeiriou, and M. Pantic are with the Department of Computing, Imperial College London, SW7 2AZ, UK. L.-P. Morency is with the Institute for Creative Technologies, University of Southern California. Z. Ghahramani is with the University of Cambridge, Cambridge, UK. e-mails: {k.bousmalis, s.zafeiriou, m.pantic}@imperial.ac.uk, morency@ict.usc.edu, zoubin@eng.cam.ac.uk*

[1] To avoid confusion, note that these are not Latent-Dynamic Conditional Random Fields [7] with countably infinite hidden states, but an infinite mixture of latent Conditional Random Field models.

present in this paper allows a countably infinite number of hidden states, shared among labels, via the use of multiple Dirichlet Process Mixtures (DPMs). Specifically, we present a novel mean field variational approach that uses DPM constructions in the model potentials to allow for the representation of a potentially infinite number of hidden states. Furthermore, we show that our model, the HCRF–DPM, is a generalization of the model presented in [11] and is able to handle continuous features naturally.

HCRF models are well–suited for a number of problems, including object recognition, gesture recognition [1], speech modeling [13] and multimodal cue modeling for human behavior recognition [14]. The latter problem of classifying episodes of high–level emotional states based on nonverbal cues in audiovisual sequences of spontaneous human behavior is rather complex. Infinite models are particularly attractive for modeling human behavior as we usually cannot have a solid intuition regarding the number of hidden states in such applications. Furthermore, it opens up the way of analyzing the hidden states these models converge to, which might provide social scientists with valuable information regarding the temporal interaction of groups of behavioral cues that are different or shared in these behaviors. We therefore decided to evaluate our novel model on behavior analysis and specifically the real–world problems of recognizing instances of agreement, disagreement and pain in recordings of spontaneous human behavior. We expected that our HCRF–DPM would converge to a correct number of shared hidden states and perform at least as well as the best cross–validated finite HCRF.

In summary, we propose in this paper:

- A novel discriminative probabilistic model that is able to automatically determine its hidden structure without losing the flexibility of an HCRF learning the appropriate weights to fine-tune this structure. The proposed model can be considered a generalization of the model proposed in [11], in terms of scalability and ability to handle continuous observations, and of the model proposed in [1] in terms of automatically determining the hidden structure of the model.
- A novel variational inference procedure to learn such a model.

In the following section, we consicely present Dirichlet Processes and finite HCRFs. We present in Section 3 our variational HCRF–DPM model. Finally, we evaluate our model performance in Section 4.2, and conclude in Section 5.

2 Theoretical Background

Our HCRF–DPM model, like many other infinite models, relies on Dirichlet Process Mixtures. We present in this section a brief introduction to Dirichlet Processes and finite Hidden Conditional Random Fields. Along with the introduction to Dirichlet Processes we discuss the *Chinese Restaurant Analogy*, an analogy that has proved helpful in explaining Dirichlet Processes and their generalizations. For a concise but complete discussion of Dirichlet Processes the reader is advised to read [15], [16]. use the formulation from [16].

2.1 Dirichlet Processes

A Dirichlet Process (DP) is a distribution of distributions, parameterized by a scale parameter α and a probability measure Ξ, the basis around which the distributions $G \sim \mathrm{DP}(\alpha, \Xi)$ are drawn, with variability governed by the α parameter. Sethuraman [17] presented the so–called "stick–breaking" construction for DPs, which is based on random variables $(\beta'_k)_{k=1}^{\infty}$ and $(h_k)_{k=1}^{\infty}$, where $\beta'_k|\alpha, \Xi \sim Beta(1, \alpha)$ and $h_k|\alpha, \Xi \sim \Xi$:

$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l) \qquad G = \sum_{k=1}^{\infty} \beta_k \delta_{h_k}, \tag{1}$$

where δ is the Dirac delta function. By letting $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ we abbreviate this construction as $\boldsymbol{\beta}|\alpha \sim GEM(\alpha)$ [17].

Successive draws from G are conditionally independent given G. By integrating G out, the conditional distribution of a draw c_i given all past draws $\{c_1, c_2, \ldots, c_{i-1}\}$ is:

$$c_i|c_1, c_2, \ldots, c_{i-1}, \alpha, \Xi \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha}\delta_{h_k} + \frac{\alpha}{i-1+\alpha}\Xi, \tag{2}$$

where n_k is the number of times a draw was assigned h_k.

A useful analogy for understanding equation 2, and its explicit clustering effect, is the *Chinese Restaurant Process*. According to the metaphor, the DP is a chinese restaurant with an unlimited number of tables. c_i is the i^{th} customer, h_k is a table in the restaurant. A draw from a DP can then be described as follows: The i^{th} customer enters the restaurant, and sits at a table h_k with a probability proportionate to the number of existing customers n_k on the k^{th} table. The customer will refuse to sit on one of the K already occupied tables with probability proportional to α, in which case the restaurant provides a new table (a new state, drawn from Ξ) and the number of occupied tables in the restaurant is incremented.

A Dirichlet Process Mixture model is a hierarchical Bayesian model that uses a DP as a nonparametric prior:

$$\begin{aligned} G|\alpha, \Xi &\sim \mathrm{DP}(\alpha, \Xi), \\ c_t \mid G &\sim G \\ s_t &\sim p(s_t|c_t) \end{aligned} \tag{3}$$

where $(s_t)_{t=1}^{T}$ is a dataset of size T, governed by a distribution conditioned on $(c_t)_{t=1}^{T}$, auxiliary index variables that get assigned each to one of the clusters $(h_k)_{k=1}^{\infty}$. As new datapoints are drawn, the number of components in this mixture model grows. In the model we present in this paper, as we explain later, we employ a number of DP priors coupled together at the data generation level, i.e. s_t above is a function of auxiliary index variables drawn from all different DPs.

2.2 Finite Hidden Conditional Random Fields

Hidden Conditional Random Fields (HCRF) —discriminative models that contain hidden states— are well–suited to a number of problems. Quattoni et al. [1] presented and used them to capture temporal dependencies across frames and recognize different gesture classes. They did so successfully by learning a state distribution among the different gesture classes in a discriminative manner, allowing them to not only uncover the distinctive configurations that uniquely identify each class, but also to learn a shared common structure among the classes. Conditional Random Fields and HCRFs can be defined in arbitrary graph structures but in our paper, driven by our application field, we assume data to be sequences that correspond to undirected chains. Our work, however, can be readily applied to tree–structured models.

We represent T observations as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$. Each observation at time $t \in \{1, \ldots, T\}$ is represented by a feature vector $\mathbf{f}_t \in \Re^d$, where d is the number of features, that can include any features of the observation sequence. We wish to learn a mapping between observation sequence $\mathbf{X}$ and class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of available labels. The HCRF does so by estimating the conditional joint distribution over a sequence of latent variables $\mathbf{s} = [s_1, s_2, \ldots, s_T]$, each of which is assigned to a hidden state $h_k \in \mathcal{H}$, and a label y, given $\mathbf{X}$. One of the main representational power of HCRFs is that the latent variables can depend on arbitrary features of the observation sequence. This allows us to model long range contextual dependencies, i.e., s_t, the latent variable at time t, can depend on observations that happened earlier or later than t.

An HCRF models the conditional probability of a class label given an observation sequence by:

$$p(y \mid \mathbf{X}, \boldsymbol{\theta}) = \sum_{\mathbf{s}} p(y, \mathbf{s} \mid \mathbf{X}, \boldsymbol{\theta}) = \frac{\sum_{\mathbf{s}} \mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta})}{\sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \mathbf{X}, \boldsymbol{\theta})}. \tag{4}$$

The model is discriminative because it doesn't model a joint distribution that includes input $\mathbf{X}$, but it only models the distribution of a label y conditioned on $\mathbf{X}$. The potential function $\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \in \Re$ is parameterized by $\boldsymbol{\theta}$, which measures the compatibility between a label y, a sequence of observations $\mathbf{X}$ and a configuration of the latent variables $\mathbf{s}$. This potential function in linear-chain finite HCRFs is defined as:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) = \exp\left\{ \sum_{t=1}^{T} \sum_{l \in L_1} \phi_{1,l}(y, s_t, \mathbf{X}) \theta_{1,l} + \sum_{t=2}^{T} \sum_{l \in L_2} \phi_{2,l}(y, s_t, s_{t-1}, \mathbf{X}) \theta_{2,l} \right\} \tag{5}$$

where L_1 is the set of node features, L_2 the set of edge features, $\phi_{1,l}$, $\phi_{2,l}$ are functions defining the features in the model, and $\boldsymbol{\theta}_{1,l}$, $\boldsymbol{\theta}_{2,l}$ are the components of $\boldsymbol{\theta}$, corresponding to node and edge parameters. Each of the ϕ_1 features depends on a single latent variable in the model; the ϕ_2 features depend on pairs of latent variables/nodes.

The graph of a linear–chain HCRF is a chain where each node corresponds to a latent variable s_t at time t. For such a model, the potential function is usually defined as:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \;=\; \exp\left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} \theta_x(s_t, i) f_t(i) + \theta_y(s_t, y) + \sum_{t=2}^{T} \theta_e(s_t, s_{t-1}, y) \right\} \tag{6}$$

In this case, our parameter vector $\boldsymbol{\theta}$ is made up of three components: $\boldsymbol{\theta} = \left[\boldsymbol{\theta_x}^T \; \boldsymbol{\theta_y}^T \; \boldsymbol{\theta_e}^T \right]^T$. Parameter vector $\boldsymbol{\theta_x}$ models the relationship between features of the observation sequence $\mathbf{f}_t$ and hidden states $h_k \in \mathcal{H}$ and is typically of length $(d \times |\mathcal{H}|)$. It can be modeled as a table with each row corresponding to one dimension of a single observation and every column to one hidden state. If the HCRF model has 10 input features and 3 hidden states, then the $\boldsymbol{\theta_x}$ parameter will be of size 30 (10×3). $\boldsymbol{\theta_y}$ models the relationship of the hidden states $h_k \in \mathcal{H}$ and labels $y \in \mathcal{Y}$ and is of length $(|\mathcal{Y}| \times |\mathcal{H}|)$. It can be modeled as a table with each row corresponding to one label and each column to a hidden state. If the model contains 3 hidden states and 2 labels, then the $\boldsymbol{\theta_y}$ will be of size 6 (2×3). $\boldsymbol{\theta_e}$ represents the links between hidden states. It is equivalent to the transition matrix in a Hidden Markov Model, but an important difference is that an HCRF keeps a matrix of "transition" weights for each label and $\boldsymbol{\theta_e}$ is of length $(|\mathcal{Y}| \times |\mathcal{H}| \times |\mathcal{H}|)$. If the HCRF model contains 3 hidden states and 2 labels, then the $\boldsymbol{\theta_e}$ will be of size 18 (2×3×3).

In this paper, we use the notation $\theta_x(h_k, \phi)$ to refer to the weight that measures the compatibility between the feature indexed by ϕ and state $h_k \in \mathcal{H}$. Similarly, $\theta_y(h_k, y)$ stand for weights that correspond to class y and state h_k, whereas $\theta_e(h_k, h', y)$ measure the compatibility of the label y with a transition from h' to h_k.

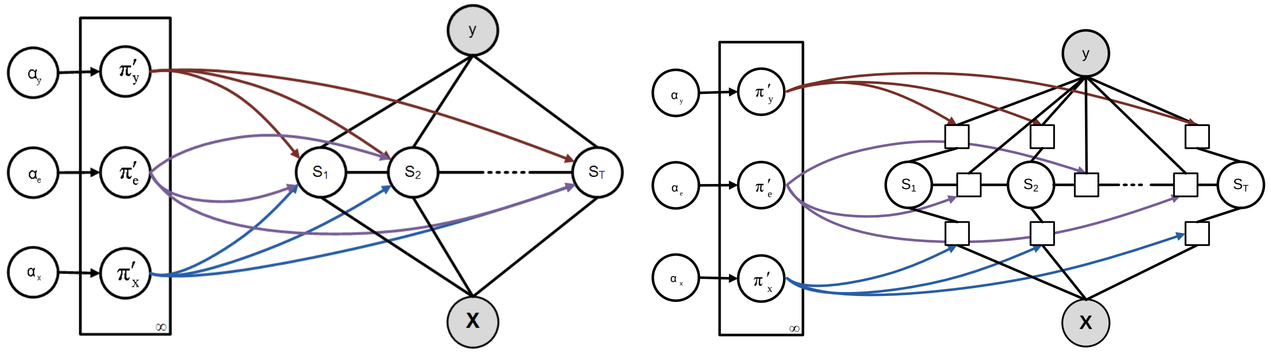
3 Hidden Conditional Random Fields with Coupled Dirichlet Process Mixtures

For an infinite HCRF we allow an unbounded number of potential hidden states in $\mathcal{H}$. This means, that for a timestamp t, latent variable s_t could get assigned to one of the infinitely many $h_k \in \mathcal{H}$. This becomes possible, by introducing random variables $\{\pi_x(h_k|i)\}_{k=1}^{\infty}$, $\{\pi_y(h_k|y)\}_{k=1}^{\infty}$, $\{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|}$ for an observation feature indexed by i, label y, and an assignment $s_{t-1} = h_a$. These new random variables are drawn by distinct processes that are able to model such quantities and are subsequently incorporated in the node and edge features of our HCRF. We present in this paper a model that uses Dirichlet Process Mixtures, an HCRF–DPM, to define these random quantities.[2] These variables, even though drawn by distinct processes, are coupled together by a common latent variable assignment in our graphical model. Figure 1 shows the graphical representations of our model. We redefine our potential function $\mathcal{F}$ from (6) as follows:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \;=\; \exp\left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} \theta_x(s_t, i) f_t(i) \log \pi_x(s_t|i) + \theta_y(s_t, y) \log \pi_y(s_t|y) + \sum_{t=2}^{T} \theta_e(s_t, s_{t-1}, y) \log \pi_e(s_t, y|s_{t-1}) \right\} \tag{7}$$

We assume that random variables $\boldsymbol{\pi} = \Big\{ \{\pi_x(h_k|i)\}_{k=1}^{\infty}, \{\pi_y(h_k|y)\}_{k=1}^{\infty}, \{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|} \Big\}$ are between 0 and 1. These are in effect the quantities that will allow the model

[2] We could however use others, like the Hierarchical Dirichlet Process and the Pitman–Yor process.



(a) HCRF–DPM Graphical Model

(b) HCRF–DPM Factor Graph

Fig. 1: Graphical representation of our Variational IHCRF driven by a number of Dirichlet Processes incorporated in the model potentials.

to 'select' an appropriate number of useful hidden states for a given classification task. $\mathbf{f}_t$ are positive features extracted from the observation sequence $\mathbf{X}$ and, as before, they can include arbitrary features of the input. We assume that $\boldsymbol{\theta}$ are positive parameters and, as in (6), they model the relationships between hidden states and features ($\boldsymbol{\theta}_x$), labels ($\boldsymbol{\theta}_y$) and transitions ($\boldsymbol{\theta}_e$). These positivity constraints for $\boldsymbol{\theta}$ and $\mathbf{f}$ are essential in this model, since the π-quantities are random variables and influence the probabilities of the hidden states: a negative parameter or feature would make an otherwise improbable state very likely to be chosen. Moreover, these constraints ensure compliance with the positivity constraints of our variational parameter updates (29)-(34), as we shall see later in this section. Finally, it is important to note that the positivity of $\boldsymbol{\theta}$ is not theoretically restrictive for our model due to the HCRF normalization factor $\frac{1}{Z(\mathbf{X})}$ in (4) where $Z(\mathbf{X}) = \sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \mathbf{X}, \boldsymbol{\theta})$.

The HCRF–DPM model is an IHCRF where the quantities $\{\pi_x(h_k|i)\}_{k=1}^{\infty}$, $\{\pi_y(h_k|y)\}_{k=1}^{\infty}$, $\{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|}$ in (7) are driven by coupled DPMs. It is important to understand that for the DPMs driving the π_e quantities in the IHCRF edge features, h_k and y are treated as a single random variable –their product– $\omega_\mu = \{h_k, y\}$ that effectively has a state–space of size $|\mathcal{Y}| \times |\mathcal{H}|$, still an infinite number. According to the stick–breaking properties of DPs, we construct $\boldsymbol{\pi} = \{\boldsymbol{\pi}_x, \boldsymbol{\pi}_y, \boldsymbol{\pi}_e\}$ conditioned on a new set of random variables $\boldsymbol{\pi}' = \{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e\}$ that follow $Beta$ distributions:

$$\pi'_x(h_k|i) \sim Beta(1, \alpha_x),$$
$$\pi_x(h_k|i) = \pi'_x(h_k|i) \prod_{j=1}^{k-1} (1 - \pi'_x(h_j|i)) \qquad (8)$$

$$\pi'_y(h_k|y) \sim Beta(1, \alpha_y),$$
$$\pi_y(h_k|y) = \pi'_y(h_k|y) \prod_{j=1}^{k-1} \left(1 - \pi'_y(h_j|y)\right) \qquad (9)$$

$$\pi'_e(\omega_\mu|h_a) \sim Beta(1, \alpha_e),$$
$$\pi_e(\omega_\mu|h_a) = \pi'_e(\omega_\mu|h_a) \prod_{j=1}^{\mu-1} (1 - \pi'_e(\omega_\mu|h_a)) \qquad (10)$$

This process can be made clearer by examining figure 2, where we visualize the stick breaking construction of an HCRF–DPM model with 2 observation features, 3 labels, and 10 'important' hidden states. The π_e-sticks have an important —for the implementation of our model— difference to the π_x and π_y–sticks in that the hidden states are intertwined with the labels, with each stick piece representing an ω–state. This means there are $|\mathcal{Y}|$ such states corresponding to one h–state. This becomes particularly important later on when we calculate our variational updates.

By using (7) the sequence of latent variables $\mathbf{s} = \{s_1, ...s_T\}$ can then be generated by the following process:

1) Draw $\pi'_x|\alpha_x \sim Beta(1, \alpha_x)$, $\pi'_y|\alpha_y \sim Beta(1, \alpha_y)$, $\pi'_e|\alpha_e \sim Beta(1, \alpha_e)$
2) Calculate $\boldsymbol{\pi}$ from (8)-(10). Note that this will only need to be calculated for a finite number of hidden states, due to our variational approximation.
3) For the t^{th} latent variable, using (7) we draw

$$\begin{aligned} s_t|\{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e, s_{t-1}, y, \mathbf{X}\} \sim \\ Mult\Bigg(\exp\Big\{ \sum_{i=1}^{d} \theta_x(s_t, i) f_t(i) \log \pi_x(s_t|i) + \\ \theta_y(s_t, y) \log \pi_y(s_t|y) + \\ \theta_e(s_t, s_{t-1}, y) \log \pi_e(\{s_t, y\}|s_{t-1}) \Big\} \Bigg) \end{aligned} \qquad (11)$$

Rather than expressing the model in terms of $\boldsymbol{\pi}$, we use $\boldsymbol{\pi}' = \{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e\}$ resulting in the folowing joint distribution that describes the HCRF–DPM:

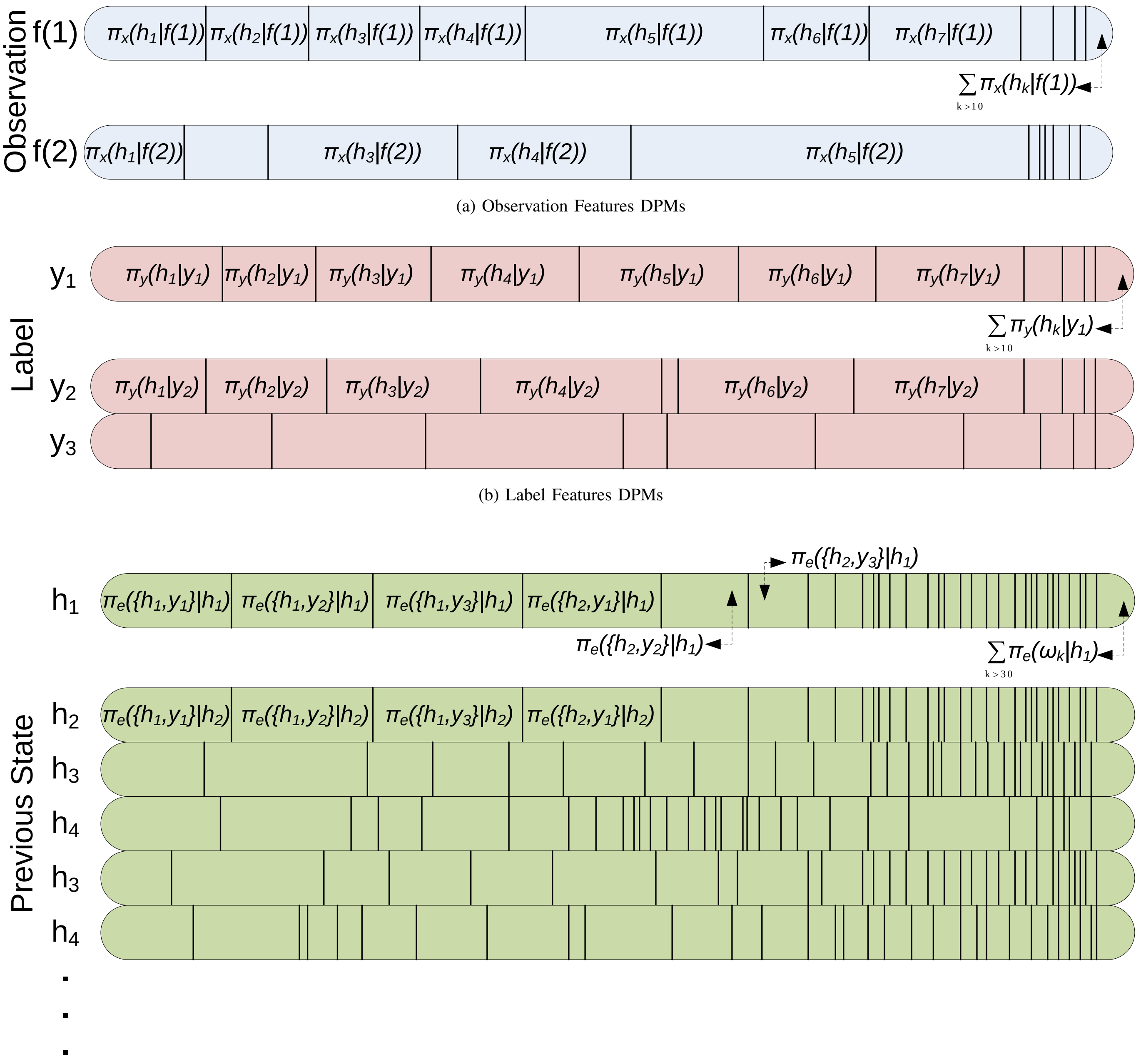
$$p(y, \mathbf{s}, \boldsymbol{\pi}'|\mathbf{X}, \theta) = p(y, \mathbf{s} \mid \boldsymbol{\pi}', \mathbf{X}, \theta) p(\boldsymbol{\pi}'_x) p(\boldsymbol{\pi}'_y) p(\boldsymbol{\pi}'_e) \quad (12)$$

with

$$p(y, \mathbf{s} \mid \boldsymbol{\pi}', \mathbf{X}, \theta) = \frac{1}{Z(\mathbf{X})} \mathcal{F}(y, \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}, \boldsymbol{\theta}) \qquad (13)$$

where $Z(\mathbf{X}) = \sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}, \boldsymbol{\theta})$. We assume independence of all π' variables above, so for example $p(\boldsymbol{\pi}'_x) = \prod_{k=1}^{\infty} \prod_{i=1}^{d} \pi'_x(h_k|i)$. We avoid explicitly writing out such expansions to make the paper easier to read.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2388228, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Observation Features DPMs

(b) Label Features DPMs

(c) Transition Features DPMs

Fig. 2: Visualization of the π-‘sticks’ used to construct the infinite states in our HCRF–DPM. The fictitious model presented here has 2 observation features $f(1), f(2)$, 3 labels y_1, y_2, y_3 and fewer than 10 important hidden states $h_1, h_2, h_3 \dots$. Each ‘stick’ sums up to 1, and the last piece always represents the sum of the lengths that correspond to all hidden states after the 10^{th} state. Notice that for the π_e-‘sticks’ this corresponds to 30 ω states. For example $\pi_e(h_1, y_3|h_2)$ controls the probability of transitioning from h_2 to h_1 in a sequence with label y_3. See text for more details.

Comparison with previous work

It is important at this stage to compare our model described by (7) with the MCMC model (IHCRF–MCMC) presented in [11]. The latter work defined potentials for each of the relationships between hidden states and features, labels and transitions and the potential function $\mathcal{F}$ as their product along the model chain:

$$\mathcal{F}(y,\mathbf{s},\mathbf{X}) = \mathcal{F}_x(\mathbf{s},\mathbf{X})\mathcal{F}_y(y,\mathbf{s})\mathcal{F}_e(y,\mathbf{s}) \tag{14}$$

$$\mathcal{F}_x(\mathbf{s},\mathbf{X}) = \prod_{t=1}^{T}\prod_{i=1}^{d}\pi_x(s_t|i)^{f_t(i)} \tag{15}$$

$$\mathcal{F}_y(y,\mathbf{s}) = \prod_{t=1}^{T}\pi_y(s_t|y) \tag{16}$$

$$\mathcal{F}_e(y,\mathbf{s}) = \prod_{t=2}^{T}\pi_e(y,s_t|s_{t-1}) \tag{17}$$

The quantities $\boldsymbol{\pi}_x, \boldsymbol{\pi}_y, \boldsymbol{\pi}_e$ above are conceptually the same as in our model, except for the fact that in [11] they have Hierarchical Dirichlet Process (HDP) priors instead of DP priors, as we do in this paper.[3]

The potential function (14) above can be rewritten as follows:

$$\begin{aligned}\mathcal{F}(y,\mathbf{s},\mathbf{X}) \;=\; \exp\Bigg\{&\sum_{t=1}^{T}\sum_{i=1}^{d} f_t(i)\log\pi_x(s_t|i) + \\ &\log\pi_y(s_t|y) + \sum_{t=2}^{T}\log\pi_e(s_t,y|s_{t-1})\Bigg\}\end{aligned} \tag{18}$$

A comparison between (18) and (7) makes it clear that our model is a generalization of the IHCRF presented in [11], which assumes, according to our framework, that $\boldsymbol{\theta}$-parameters are set to 1. The introduction of these parameters is not redundant, but allows for more powerful and flexible models. Also, when dealing with classification problems involving continuous observation features using (7) for the potential function of an infinite HCRF is more suitable than (18), as we show in the experimental section. In those cases it is known that θ–parameters are of particular importance as they are able to capture the scaling of each input feature. The former model is not guaranteed to perform well unless some non–trivial normalization is applied on the observation features.

3.1 Variational Inference for the HCRF–DPM

Since inference on our model (12) is intractable, we need to approximate the marginal probabilities along the chain of our graphical model, and the π–quantities in (7). We shall do so with a mean–field variational inference approach. The basic idea of such an approach is to restructure our quantities computation into an optimization problem. We can then simplify our optimization which depends only on a number of so–called variational parameters. Solving for those will give us updates for a coordinate descent algorithm which will converge to an approximation of the quantities we wish to calculate. We use the following approximation for the joint distribution of our model:

$$q(y,\mathbf{s},\boldsymbol{\pi}'|\mathbf{X}) = q(y,\mathbf{s}|\mathbf{X})q(\boldsymbol{\pi}'_x)q(\boldsymbol{\pi}'_y)q(\boldsymbol{\pi}'_e) \tag{19}$$

where,

$$\begin{aligned} q(y,\mathbf{s}|\mathbf{X}) &= q(y,s_1|\mathbf{X})\prod_{t=2}^{T} q(y,s_t|s_{t-1},\mathbf{X}) \\ &= \prod_{i=1}^{d} q(s_1|i)\prod_{y'\in\mathcal{Y}} q(s_1|y') \\ &\quad \prod_{t=2}^{T}\prod_{i=1}^{d} q(s_t|i)\prod_{y'\in\mathcal{Y}}\Big(q(s_t|y')\Big)q(s_t,y|s_{t-1}) \end{aligned} \tag{20}$$

Each individual approximate $q(\pi'_x), q(\pi'_y), q(\pi'_e)$ follows a $Beta$ distribution with variational parameters $\boldsymbol{\tau}_x, \boldsymbol{\tau}_y, \boldsymbol{\tau}_e$ respectively. Explicitly, for features indexed by i, labels indexed by y, and hidden states indexed by k, k':

$$q(\pi'_x(h_k|i)) = Beta\left(\tau_{x,1}(k,i), \tau_{x,2}(k,i)\right) \tag{21}$$

$$q(\pi'_y(h_k|y)) = Beta\left(\tau_{y,1}(k,y), \tau_{y,2}(k,y)\right) \tag{22}$$

$$q(\pi'_e(y,h_k|h_{k'})) = Beta\left(\tau_{e,1}(y,k,k'), \tau_{e,2}(y,k,k')\right) \tag{23}$$

In order to make inference tractable we approximate all $\boldsymbol{\pi}$ variables by employing a truncated stick–breaking representation which approximates the infinite number of hidden states with a finite number L [15]. This is the crux of our variational approach, and it effectively means that we set a truncation threshold L, above which the above quantities are set to 0: $\forall k > L,\;\; q(\pi'_x(h_k|i)) = 0, \;\; q(\pi'_y(h_k|y)) = 0,$ $q(\pi'_e(y,h_k|h_{k'})) = 0$. Note that using this approximation is statistically rather different from using a finite model: an HCRF–DPM simply approximates the infinite number of states and will still reduce the number of usefull hidden states to something smaller than L. It will be easier to understand how by examining figure 2 in our supplementary material, where we show how a finite HCRF with 50 hidden states compares to an HCRF–DPM with $L = 50$. It is finally important to stress that by constraining our θ–parameters and observation features to be positive, we effectively make the number of the θ–parameters that matter finite: changing a θ–parameter associated with a hidden state $k > L$ will not change our model, as one can see in (7). Note that the choice of L has to be the same during training and inference.

3.2 Model Training

A trained variational HCRF–DPM model is defined as the set of optimal parameters $\boldsymbol{\theta}^*$ and optimal variational parameters $\boldsymbol{\tau}^*$. In this work we obtain these with a training algorithm (see Alg. 1 for a summary) that can be divided in two distinct phases: *(i)* the optimization of our variational paramaters through a coordinate descent algorithm using the updates derived below and *(ii)* the optimization of parameters $\boldsymbol{\theta}$ through

[3] Using HDP priors allows separate DPMs to be linked together via an identical base probabilistic measure, which is itself a DP. It would be interesting to use such priors for our model, but we were able to obtain satisfactory results without introducing higher complexity and additional hyperparameters into the Variational IHCRF we experimented with. Notice that our model allows for such flexibility: using HDP priors would simply change the updates for our variational coordinate descent algorithm.

a gradient ascent method. Although it would be possible to have a fully Bayesian model with $\boldsymbol{\theta}$ being random variables in our model, inference would become more difficult. Moreover, having a single value for our $\boldsymbol{\theta}$ parameters is good for model interpretability and makes the application of a trained model to test data much easier.

Algorithm 1 Model Training for Variational HCRF–DPM

```
Initialize $s_{x,1}, s_{x,2}, s_{y,1}, s_{y,2}, s_{e,1}, s_{e,2}$
Randomly initialize $\alpha_x, \alpha_y, \alpha_e, \boldsymbol{\theta}, \boldsymbol{\tau}$
Initialize $nbItrs$, $nbVarItrs$
$itr = 0$
$converged = FALSE$

while (not $converged$) and ($itr < nbItrs$) do
  $varItr = 0$
  $varConverged = FALSE$
  while (not $varConverged$) and
  ($varItr < nbVarItrs$) do
    {Phase 1: Optimize variational parameters $\boldsymbol{\tau}$}
    Calculate $\forall t\ q(s_t|\mathbf{X}, y, s_{t-1})$ by using (35)–(41)

    Compute approximate marginals $q(s_t = h_k|i)$,
    $q(s_t = h_k|y)$, and $q(s_t = h_k, y, s_{t-1} = h_{k'})$ by using
    a forward–backward algorithm.

    Hyperparameter posterior sampling for $\alpha_x, \alpha_y, \alpha_e$ by
    using (45)

    Calculate Kullback-Liebler divergence $KL(varItr)$ by
    using (27)

    Update $\boldsymbol{\tau}$ by using (29)-(34)

    $varConverged = \frac{KL(varItr) - KL(varItr-1)}{KL(varItr)} < \epsilon$

    $varItr = varItr + 1$
  end while
  {Phase 2: Optimize parameters $\boldsymbol{\theta}$}
  Gradient ascent to find $\boldsymbol{\theta}(iteration)$ by using a quasi–
  Newton method with (46)–(48) and an Armijo backtrack-
  ing line search with projected gradients to keep $\boldsymbol{\theta}$ non–
  negative
  $converged = \sum \left(|\boldsymbol{\theta}(itr) - \boldsymbol{\theta}(itr-1)|\right) < \epsilon'$

  $itr = itr + 1$
end while
```

Although it is possible to have a fully Bayesian model with $\boldsymbol{\theta}$ being random variables in our model, inference would become more difficult. Moreover, having a single value for our parameters is good for interpretability of our model, and makes the application of a trained model to test data much easier.

Phase 1: Optimization of variational parameters $\boldsymbol{\tau}$

Now that we have defined an approximate model distribution in (20), we can approximate the necessary quantities $q(s_t)$, $q(s_t, y)$, $q(s_t, s_{t-1})$, $q(s_t, s_{t-1}, y)$, $\log q(\boldsymbol{\pi}_x)$, $\log q(\boldsymbol{\pi}_y)$, $\log q(\boldsymbol{\pi}_e)$ for our inference. These approximations, as one can see later in this section, depend solely on our variational parameters $\boldsymbol{\tau}$. We calculate those by minimizing the reverse Kullback-Liebler divergence (KL) between approximate and actual joint distributions of our model, (12) and (20), using a coordinate descent algorithm:

$$KL\left[q(y, \mathrm{s}, \boldsymbol{\pi}'|\mathbf{X},) \mid\mid p(y, \mathrm{s}, \boldsymbol{\pi}'|\mathbf{X}, \boldsymbol{\theta})\right] = \tag{24}$$

$$\int_{\boldsymbol{\pi}'} \sum_{y,\mathrm{s}} q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}') \log \frac{q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}')}{p(y, \mathrm{s}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\pi}') p(\boldsymbol{\pi}')} d\boldsymbol{\pi}' = \tag{25}$$

$$\int_{\boldsymbol{\pi}'} \sum_{y,\mathrm{s}} q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}') \log \frac{Z(\mathbf{X}) q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}')}{\mathcal{F}(y, \mathrm{s}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}')} d\boldsymbol{\pi}' \tag{26}$$

Since the normalization factor $Z(\mathbf{X}) = \sum_{y,\mathrm{s}} \mathcal{F}(y, \mathrm{s}, \mathbf{X})$ is a constant for a given observation sequence, the reverse Kullback–Liebler divergence becomes:

$$KL[q||p] = \log Z(\mathbf{X}) - \langle \log \mathcal{F}(y, \mathrm{s}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}') \rangle_{q(y,\mathrm{s},\boldsymbol{\pi}'|\mathbf{X})} + \langle \log q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}') \rangle_{q(y,\mathrm{s},\boldsymbol{\pi}'|\mathbf{X})} \tag{27}$$

where $\langle \cdot \rangle_q$ is the expectation of $\cdot$ with respect to q. Thus, the energy of the configuration of our random variables y, s, and $\boldsymbol{\pi}'$ is $\log \mathcal{F}(y, \mathrm{s}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}')$ and the free energy of the variational distribution:

$$\mathcal{L}(q) = -\langle \log \mathcal{F}(y, \mathrm{s}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}') \rangle_{q(y,\mathrm{s},\boldsymbol{\pi}'|\mathbf{X})} + \langle \log q(y, \mathrm{s}|\mathbf{X}) q(\boldsymbol{\pi}') \rangle_{q(y,\mathrm{s},\boldsymbol{\pi}'|\mathbf{X})} \tag{28}$$

Since $\log Z(\mathbf{X})$ is constant for a given observation sequence, minimizing the free energy $\mathcal{L}(q)$ minimizes the KL divergence. And since KL$[q||p]$ is positive, the free energy $\mathcal{L}(q) \geq -\log Z(\mathbf{X})$. Therefore KL is minimized at 0 when $\mathcal{L}(q) = \log Z(\mathbf{X})$.

We will obtain the variational updates for the two groups of latent variables $q(y, \mathrm{s}|\mathbf{X})$ and $q(\boldsymbol{\pi}')$ by setting the partial derivative with respect to each group of $\mathcal{L}(q)$ to 0 and solving for the approximate distribution of each group of latent variables. The updates for the $Beta$ parameters of $q(\boldsymbol{\pi}')$ from (21)-(23) are:

$$\tau_{x,1}(k,i) = 1 + \sum_t f_t[i]\theta_x(k,i)q(s_t = h_k) \tag{29}$$

$$\tau_{x,2}(k,i) = \alpha_x + \sum_t f_t[i] \sum_{b>k} \theta_x(b,i)q(s_t = h_b) \tag{30}$$

$$\tau_{y,1}(k,y) = 1 + \sum_t \theta_y(k,y)q(s_t = h_k) \tag{31}$$

$$\tau_{y,2}(k,y) = \alpha_y + \sum_t \sum_{b>k} \theta_y(b,i)q(s_t = h_b) \tag{32}$$

$$\tau_{e,1}(y,k,k') = 1 + \sum_t \theta_e(k,k',y)q(s_t = h_k, s_{t-1} = h_{k'}, y) \tag{33}$$

$$\begin{aligned}\tau_{e,2}(y,k,k') =& \\ &\alpha_e + \sum_t \sum_{y_l>y} \theta_e(k,k',y_l)q(s_t = h_k, s_{t-1} = h_{k'}, y_l) + \\ &\sum_{b>k,y_l} \theta_e(b,k',y_l)q(s_t = h_b, s_{t-1} = h_{k'}, y_l)\end{aligned} \tag{34}$$

Quantities $q(s_t = h_k)$, $q(s_t = h_k)$, and $q(s_t = h_k, y, s_{t-1} = h_{k'})$ can be obtained by the forward–backward algorithm. The latter requires only conditional approximate likelihoods $q(s_t = h_k|i, y, h_{k'})$, which can be be calculated by setting the derivative of $\mathcal{L}(q)$ with respect to $q(y, \mathbf{s}|\mathbf{X})$ to zero:

$$\begin{aligned}&q(s_t = h_k|i, y, h_{k'}) \propto \\ &\quad \exp\Big\{ f_t(i)\theta_x(k,i)\Big(\langle \log \pi'_x(s_t = h_k|i)\rangle_{q(\boldsymbol{\pi}')} + \\ &\quad \sum_{j=1}^{k-1} \langle \log(1 - \pi'_x(s_t = h_j|i))\rangle_{q(\boldsymbol{\pi}')}\Big) \\ &\quad \theta_y(k,y)\Big(\langle \log \pi'_x(s_t = h_k|y)\rangle_{q(\boldsymbol{\pi}')} + \\ &\quad \sum_{j=1}^{k-1} \langle \log(1 - \pi'_y(s_t = h_j|y))\rangle_{q(\boldsymbol{\pi}')}\Big) \\ &\quad \theta_e(k,k',y)\Big(\langle \log \pi'_e(s_t = h_k, y|s_{t-1} = h_{k'})\rangle_{q(\boldsymbol{\pi}')} + \\ &\quad \sum_{j=1}^{k-1} \langle \log(1 - \pi'_e(s_t = h_j, y|s_{t-1} = h_{k'}))\rangle_{q(\boldsymbol{\pi}')}\Big)\Big\}\end{aligned} \tag{35}$$

Since all $\boldsymbol{\pi}'$ follow a $Beta$ distribution, the expectations above are known:

$$\langle \log \pi'_x(s_t = h_k|i)\rangle = \Psi(\tau_{x,1}(k,i)) - \Psi(\tau_{x,1}(k,i) + \tau_{x,2}(k,i)) \tag{36}$$

$$\begin{aligned}&\langle \log(1 - \pi'_x(s_t = h_k|i))\rangle = \\ &\quad \Psi(\tau_{x,2}(k,i)) - \Psi(\tau_{x,1}(k,i) + \tau_{x,2}(k,i))\end{aligned} \tag{37}$$

$$\begin{aligned}&\langle \log \pi'_y(s_t = h_k|y)\rangle = \\ &\quad \Psi(\tau_{y,1}(k,y)) - \Psi(\tau_{y,1}(k,y) + \tau_{y,2}(k,y))\end{aligned} \tag{38}$$

$$\begin{aligned}&\big\langle \log\big(1 - \pi'_y(s_t = h_k|y)\big)\big\rangle = \\ &\quad \Psi(\tau_{y,2}(k,y)) - \Psi(\tau_{y,1}(k,y) + \tau_{y,2}(k,y))\end{aligned} \tag{39}$$

$$\begin{aligned}&\langle \log \pi'_e(s_t = h_k, y|h_{k'})\rangle = \\ &\quad \Psi(\tau_{e,1}(y,k,k')) - \Psi(\tau_{e,1}(y,k,k') + \tau_{e,2}(y,k,k'))\end{aligned} \tag{40}$$

$$\begin{aligned}&\langle \log(1 - \pi'_e(s_t = h_k, y|h_{k'}))\rangle = \\ &\quad \Psi(\tau_{e,2}(k,y)) - \Psi(\tau_{e,1}(k,y) + \tau_{e,2}(k,y))\end{aligned} \tag{41}$$

where $\Psi(\cdot)$ is the digamma function.

The scaling parameters $\alpha_x, \alpha_y, \alpha_e$ can have a significant effect on our HCRF–DPM model, as they control the growth of the used hidden states. It is suggested in [15] that for DPMs one should place a $Gamma(s_1, s_2)$ prior on these parameters and integrate over them. Since our model uses a number of DPMs, we include posterior updates for these scaling parameters as part of our variational coordinate descent algorithm. In this work, we use a different scaling parameter for each DPM, but with a common prior. The variational distribution for the scaling parameter $\alpha_{x,i}$ corresponding to the DPM for feature i is

$$q(\alpha_{x,i}) = Gamma\,(w_{1,x}, w_{2,x,i}) \tag{42}$$

where

$$w_{1,x} = s_{1,x} + L - 1 \tag{43}$$

$$w_{2,x,i} = s_{2,x} - \sum_{k=1}^{L-1} \langle \log(1 - \pi'_x(k,i))\rangle_q \tag{44}$$

and we replace the α_x values in (30) with the respective expectation:

$$\langle \alpha_{x,i}\rangle_q = \frac{w_{1,x}}{w_{2,x,i}} \tag{45}$$

The posterior updates for the rest of the scaling parameters are obtained in a similar fashion and so they are omitted for brevity.

Phase 2: Optimization of parameters $\boldsymbol{\theta}$

We find our optimal parameters $\boldsymbol{\theta}^* = \arg\max \log p(y|\mathbf{X}, \boldsymbol{\theta})$ based on a training set by using a common HCRF quasi–Newton gradient ascent method (LBFGS), which requires the gradient of the log–likelihood with respect to each parameter. These gradients for our IHCRF are:

$$\begin{aligned}\frac{\partial \log p(y|\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_x(k,i)} &= \sum_t p(s_t = h_k|y, \mathbf{X}, \boldsymbol{\theta}) f_t(i) \log \pi_x(h_k|i) - \\ &\sum_{y' \in \mathcal{Y}, t} p(s_t = h_k, y'|\mathbf{X}, \boldsymbol{\theta}) f_t(i) \log \pi_x(h_k|i)\end{aligned} \tag{46}$$

$$\begin{aligned}\frac{\partial \log p(y|\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_y(k,y)} &= \sum_t p(s_t = h_k|y, \mathbf{X}, \boldsymbol{\theta}) \log \pi_y(h_k|y) - \\ &\sum_{y' \in \mathcal{Y}, t} p(s_t = h_k, y'|\mathbf{X}, \boldsymbol{\theta}) \log \pi_y(h_k|y)\end{aligned} \tag{47}$$

$$\begin{aligned}&\frac{\partial \log p(y|\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_e(k,k',y)} = \\ &\quad \sum_t p(s_t = h_k, s_{t-1} = h_{k'}|y, \mathbf{X}, \boldsymbol{\theta}) \log \pi_e(h_k, y|h_{k'}) \\ &\qquad - \sum_{y' \in \mathcal{Y}, t} p(s_t = h_k, s_{t-1} = h_{k'}, y'|\mathbf{X}, \boldsymbol{\theta}) \log \pi_e(h_k, y|h_{k'})\end{aligned} \tag{48}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2388228, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Transition Matrix of the HMM producing sequences for Label 1 with states S1, S2, S3 and S4

HMM-1	S1	S2	S3	S4
S1	0.4	0.4	0.1	0.1
S2	0.1	0.4	0.4	0.1
S3	0.1	0.1	0.4	0.4
S4	0.4	0.1	0.1	0.4

TABLE 2: Transition Matrix of the HMM producing sequences for Label 2 with states S1, S2, S3 and S4

HMM-2	S1	S2	S3	S4
S1	0.1	0.7	0.1	0.1
S2	0.1	0.1	0.7	0.1
S3	0.1	0.1	0.1	0.7
S4	0.7	0.1	0.1	0.1

TABLE 3: Mean and variance for the Gaussian states of each HMM

	S1	S2	S3	S4
HMM-1μ	0.1	2	5	15
HMM-1σ	0.4	0.8	0.12	0.56
HMM-2μ	0.1	2	-10	-13
HMM-2σ	0.4	0.8	0.8	0.8

We make this gradient ascent tractable by using the variational approximations for the intractable quantities in the above equations. However, there is a significant difference with other CRF and HCRF models that use such techniques to find optimal parameters: we are constrained to only positive θ-parameters, as this is an assumption we have to make for our truncated stick–breaking process. Since we are using a quasi–Newton method with Armijo backtracking line search, we can use the gradient projection method of [18], [19] to enforce this constrain. Finally, it is important to stress here that, although our model includes parameters that are not treated probabilistically, we have not seen signs of overfitting in our experiments (see Fig. 4).

Computational Complexity

The computational complexity of one iteration for the IHCRF–MCMC model that is used by [11] is in fact $\mathcal{O}(TL^2)$, where T is the length of the sequence and L is the number of represented states, as it is a forward filtering-backwards sampling algorithm. In our variational method an inference step is $\mathcal{O}(TL^2)$, where T is the length of the sequence and L the the number of available states. In an optimal implementation this could be a lot lower in practice by choosing to ignore the use of hidden states that have a probability of being chosen close to 0. In fact, a big advantage of our variational method is that it is a lot faster during inference. This is because the IHCRF–MCMC needs to aggregate a large number of samples during inference: after training only the hyperparameters for that model are fixed, and the parameters are sampled anew every time. In contrast, the method we present here learns fixed parameters that are used for the forward-backward algorithm.

4 Experimental Results

4.1 Performance on a Synthetic Dataset with Continuous Features

In an effort to demonstrate the ability of our HCRF–DPM to model sequences with continuous features correctly, we created a synthetic dataset, on which we compared its performance to that of the IHCRF–MCMC model [11]. The simple dataset was generated by two HMMs, with 4 Gaussian hidden states initialized with the transition matrices, means and standard deviations as shown in Tables 1–3. Two of the states were shared between the two HMMs, resulting in a total of 6 unique hidden states, out of a total of 8 for the two labels.

We trained 10 randomly initialized models of the finite HCRF, IHCRF–MCMC and HCRF–DPM on 100 training sequences and chose in each case the best one based on their performance on an evaluation set of 100 different sequences. The performance of the models was finally evaluated by comparing the F1 measure achieved on a test set of 100 other sequences. All sets had an equal number of samples from each label. The IHCRF–MCMC model was unable to solve this simple two–label sequence classification problem with continuous-only input features: it consistently selected Label 1. On the other hand, the finite HCRF and the new HCRF–DPM model were successful in achieving a perfect F1 score of 100% on the test set (see Table 4).

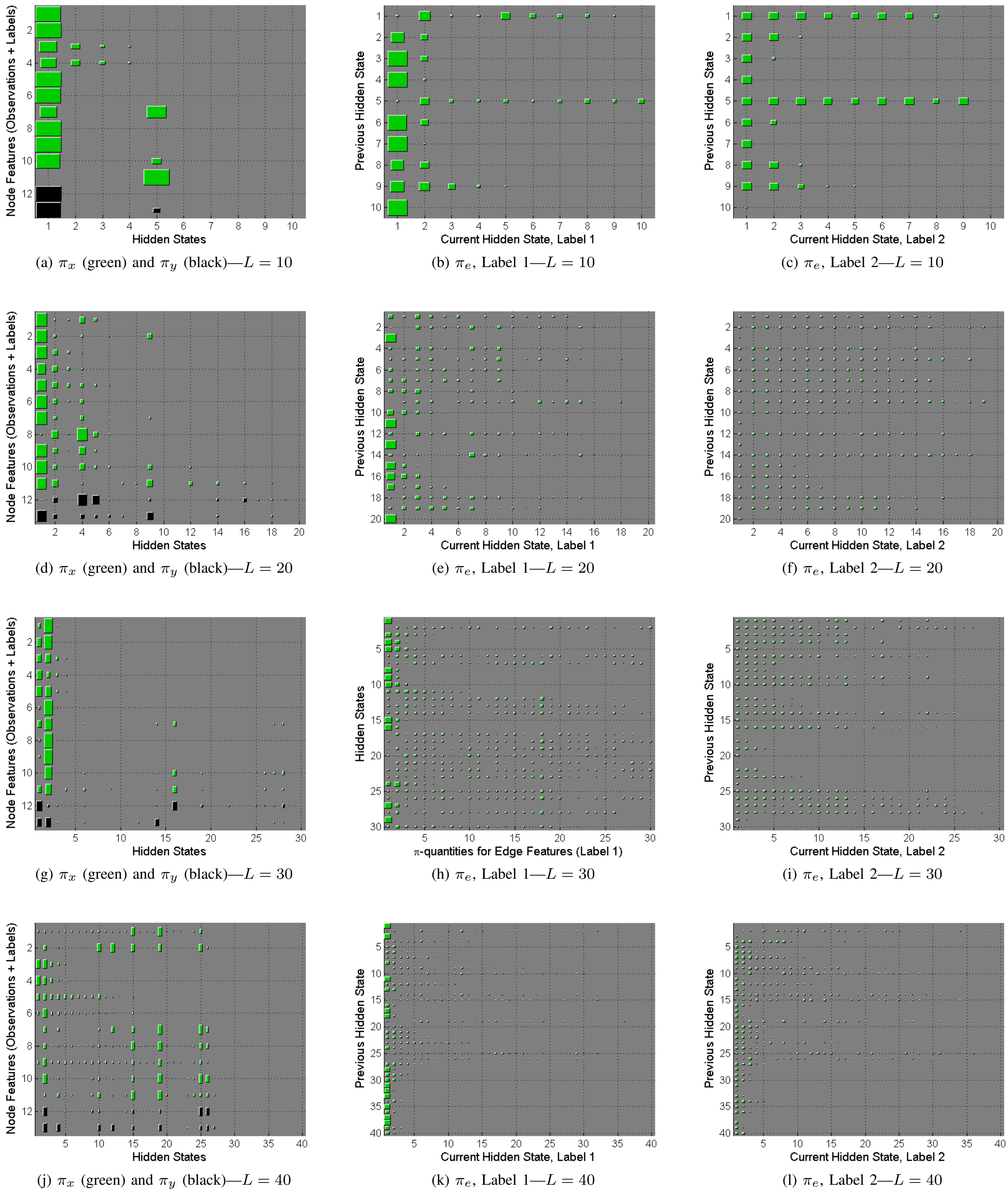
4.2 Application to the Audiovisual Analysis of Human Behavior

The problem of automatically classifying episodes of high–level emotional states, such as pain, agreement and disagreement, based on nonverbal cues in audiovisual sequences of spontaneous human behavior is rather complex [20]. Although humans are particularly good at interpreting such states, automated systems perform rather poorly. Infinite models are particularly attractive for modeling human behavior as we usually cannot have a solid intuition regarding the number of hidden states in such applications. Furthermore, it opens up the way of analyzing the hidden states these models converge to, which might provide social scientists with valuable information regarding the temporal interaction of groups of behavioral cues that are different or shared in these behaviors. We therefore decided to evaluate our novel approach on behavior analysis

TABLE 4: F1 measure achieved by our HCRF-DPM vs. the best, in each fold of each problem, finite HCRF and IHCRF-MCMC. **Synthetic:** Two–label classification for an HMM–generated dataset with continuous–only features **ADA2:** Two–label classification for the Canal9 Dataset of agreement and disagreement; **ADA3:** Three-label classification for the Canal9 Dataset; **PAIN2:** Two–label classification for the UNBC dataset of shoulder pain; **PAIN3:** Three–label classification for the UNBC dataset

Dataset	Finite HCRF	IHCRF–MCMC	Our HCRF–DPMs
Synthetic	100.0%	33.3%	**100.0**%
ADA2	58.4%	61.2%	**76.1**%
ADA3	50.7%	**60.3**%	49.8%
PAIN2	83.9%	88.4%	**89.2**%
PAIN3	53.9%	57.7%	**59.0**%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2388228, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) π_x (green) and π_y (black)—$L = 10$

(b) π_e, Label 1—$L = 10$

(c) π_e, Label 2—$L = 10$

(d) π_x (green) and π_y (black)—$L = 20$

(e) π_e, Label 1—$L = 20$

(f) π_e, Label 2—$L = 20$

(g) π_x (green) and π_y (black)—$L = 30$

(h) π_e, Label 1—$L = 30$

(i) π_e, Label 2—$L = 30$

(j) π_x (green) and π_y (black)—$L = 40$

(k) π_e, Label 1—$L = 40$

(l) π_e, Label 2—$L = 40$

Fig. 3: Hinton Diagrams of π-quantities in node and edge features of variational HCRF-DPM models with $L = 10$ on the first row (a-c), $L = 20$ on the second (d-f), $L = 30$ on the third (g-i), $L = 40$ on the fourth (j-l) for ADA2. The first column presents the π-quantities for node features: π_x for observation features in green, π_y for labels in black. The second and third columns present the π_e-quantities for labels 1 and 2 respectively. See text for additional details

and specifically the recognition of agreement, disagreement and pain in recordings of spontaneous human behavior. We expected that our HCRF–DPM models would find a good number of shared hidden states and perform at least as well as the best cross–validated finite HCRF.

In this work we used an audiovisual dataset of spontaneous agreement and disagreement and a visual dataset of pain to evaluate the performance of the proposed model on four classification problems: *(1)* ADA2, agreement and disagreement recognition with two labels (agreement vs. disagreement); *(2)* ADA3, agreement and disagreement recognition with three labels (agreement vs. disagreement vs. neutral); *(3)* PAIN2, pain recognition with two labels (strong pain vs. no pain); and *(4)* PAIN3, pain recognition with three labels (strong pain vs. moderate pain vs. no pain). We show that *(1)* our model is capable of finding a good number of useful states; and *(2)* HCRF–DPMs perform better than the best performing finite HCRF and HCRF–MCMC models in all of these problems with the exception of ADA3, where the performance of the HCRF–DPM is similar to that of the finite model.

The audiovisual dataset of spontaneous agreement and disagreement comprises of 53 episodes of agreement, 94 episodes of disagreement, and 130 neutral episodes of neither agreement or disagreement. These episodes feature 28 participants and they occur over a total of 11 real political debates from *The Canal9 Database of Political Debates*[4] [21]. As the debates were filmed with multiple cameras, and edited live to one feed, the episodes selected for the dataset were only the ones that were contained within one personal, close–up shot of the speaker. We used automatically extracted prosodic features (continuous), based on previous work on agreement and disagreement classification, and manually annotated visual features, the hand and head gestures hypothesized relevant according to literature [22] (binary). The 2 prosodic features used were F0 and Energy, and the 9 gestures used in our experiments are the 'Head Nod', 'Head Shake', 'Forefinger Raise', 'Forefinger Raise–Like', 'Forefinger Wag', 'Hand Wag', 'Hand Chop', 'Hands Scissor', and 'Shoulder Shrug' (see [22] for details). We encoded each gesture in a binary manner, based on its presence at each of the 5,700 total number of video frames, with each sequence ranging from 30 to 120 frames. The prosodic features were extracted with the publicly available software package *OpenEar* [23]. We compared the finite HCRFs and the IHCRF–MCMC to our HCRF–DPM based on the F1 measure they achieved. In each case, we evaluated their performance on a test set consisting of sequences from 3 debates. We ran all models with 60 random initializations, selecting the best trained model each time by examining the F1 achieved on a validation set consisting of sequences from 3 debates. It is important to stress that each sequence belonged uniquely to either the training, the validation, or the testing set.

The database of pain we used was the *UNBC-McMaster Shoulder Pain Expression Database*[5] [24], which features 25 subjects–patients spontaneously expressing various levels of elicited pain in a total of 200 video sequences. The database was coded for, among others, pain level per sequence by expert observers on a 6–point scale from 0 (no pain) to 5 (extreme pain). Furthermore, each of the 48,398 video frames in the database was coded for each of the observable facial muscle movements–Action Units (AUs) according to the Facial Action Coding System (FACS) [25] by expert FACS coders. In our experiments we encoded each of the possible 45 AUs in a binary manner, based on their presence. We labeled sequences coded with 0 as 'no pain', sequences coded with 1–2 as 'moderate pain', and those coded as 3–5 as 'strong pain'. For our experiments, we compared the finite HCRFs and the IHCRF–MCMC to our HCRF–DPM based on the F1 measure they achieved. We evaluated the performance of the models on 25 different folds (leave–7–subjects–out for testing). In each case we concatenated the predictions for every test sequence of each fold and calculated the F1 measure for each label. The measure we used was the average F1 over all labels. We ran both HCRF and HCRF-DPM experiments with 10 random initializations, selecting the best model each time by examining the F1 achieved on a validation set consisting of the sequences from 7 subjects. In every fold our training, validation and testing sets comprised not only of unique sequences but also of unique subjects.

For all four tasks, in addition to the random initializations the best HCRF model was also selected by experimenting with different number of hidden states and different values for the HCRF L2 regularization coefficient. Specifically, for each random initialization we considered models with 2, 3, 4, and 5 hidden states and an L2 coefficient of 1, 10, and 100. This set of values for the hidden states was selected after preliminary results deemed a larger number of hidden states only resulted in severe overfitting for all problems. We did not use regularization for our HCRF-DPM models and all of them had their truncation level set to $L = 10$ and their hyperparameters to $s_1 = 1000$ and $s_2 = 10$. Finally, our finite HCRF models were trained with a maximum of 300 iterations for the gradient ascent method used [1], whereas our HCRF-DPM models were trained with a maximum of 1200 variational coordinate descent iterations and a maximum of 600 iterations of gradient ascent. All IHCRF–MCMC models were trained according to the experimental protocol of [11]. They had their initial number of represented hidden states set to $K = 10$, they were trained with 100 sampling iterations, and were tested by considering 100 samples.

In an attempt to clearly show how a variational HCRF-DPM functions differently from a finite HCRF, we compared the learned potentials of an HCRF with 50 hidden states for the 2–label (dis)agreement recognition problem to the learned equivalent potentials of an HCRF–DPM with an upper bound of hidden states set to $L = 50$. An HCRF uses all 50 states roughly equally, whereas the learned potentials for HCRF–DPM are a lot more sparse with only a few number of hidden states used, due to the nonparametric prior on the $\boldsymbol{\pi}$-quantities (see relevant figure in the supplementary material provided with this paper).

In figures 3 we show the learned nonparametric π parts of the features of the best HCRF–DPM ADA2 model, based

[4] Publicly available at http://canal9-db.sspnet.eu/

[5] Publicly available at http://www.pitt.edu/∼jeffcohn/PainArchive/

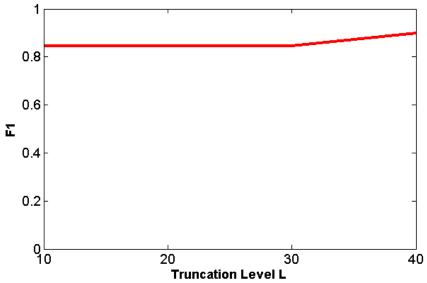


Fig. 4: HCRF–DPM F1 measure (higher F1 means higher perfomance) achieved on the validation set of ADA2. Our model does not show signs of overfitting: the F1 achieved on the validation set does not decrease as the truncation level L, and thus the number of $\boldsymbol{\theta}$–parameters, increases.

on F1 achieved on our validation set, for $L = 10, 20, 30$ and 40. Each row is a separate DPM, with the DPMs for the edge potentials spanning across labels. Recall from figure 2 that these quantities have to sum to 1 across each row. As one can see in these figures, paying particular attention to the first column (node features), the number of hidden states essentially utilized seems to be less than 10 in all cases. Figure 5 visualizes the learned nonparametric quantities of our HCRF–DPM features for PAIN2 with $L = 10$. As one can clearly see, the model uses only a small number of shared hidden states. An increase to L increases the number of quantities we need to estimate, and we also need to increase our number of random initializations to find a suitable one for our model. $L = 10$ therefore seems to be a reasonable value that allows the proper balance between computation time and accuracy.

Since we have introduced parameters $\boldsymbol{\theta}$ it is sensible to test our methodology for signs of overfitting. The only value linked with the number of our parameters is our truncation level L: their number increases as we increase L. In figure 4 we show the F1 measure achieved on the validation set of ADA2 for HCRF–DPMs with L=10, 20, 30, 40. This graph is a strong indication that HCRF–DPMs do not show signs of overfitting. We would see such signs if by increasing L the performance (F1 measure) for our validation set would decrease. However, as we see here, performance on the validation set remains roughly the same as we increase L.

Table 4 shows the average over all labels of the F1 measure on the test sets for all four of our problems. Since the nonparametric model structure is not specified a priori but is instead determined from our data, the HCRF–DPM model is more flexible than the finite HCRF and is able to achieve better performance in all cases with the exception of the 3–label classification problem of agreement/disagreement (ADA3), where the HCRF–DPM seems to perform almost equally well with the finite model. The HCRF–DPM performed better than the IHCRF–MCMC in all problems with the exception of ADA3. An analysis of an IHCRF–MCMC model trained for ADA3 shows that the model ignored the two continuous dimensions and used only the binary features to model the dataset, which evidently resulted in slightly better performance.

5 Conclusion

In this paper we have presented a variational approach to learning an infinite Hidden Conditional Random Field, the HCRF–DPM, a discriminative sequential model with a countably infinite number of hidden states. This deterministic approach overcomes the limitations of sampling techniques, like the one presented in [11]. We have also shown that our model is in fact a generalization of the one presented in [11] and is able to handle sequence classification problems with continuous features naturally. In support of the latter claim, we conducted an experiment with a Gaussian HMM–generated synthetic dataset of continuous–only features which showed that HCRF–DPMs are able to perform well on classification problems where the IHCRF–MCMC fails. Furthermore, we conducted experiments with four challenging tasks of classification of naturalistic human behavior. HCRF–DPMs were able to find a good number of shared hidden states, and to perform well in all problems, without showing signs of overfitting.

Acknowledgements

The current work of the team is funded in part by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work by Stefanos Zafeiriou is funded in part by the EPSRC project EP/J017787/1 (4DFAB).

References

[1] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1848–1852, 2007.

[2] M. Beal, Z. Ghahramani, and C. Rasmussen, "The infinite hidden Markov model," in *Proc. NIPS*, 2002, pp. 577–584.

[3] J. V. Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden markov model," in *ICML*, 2008, pp. 1088–1095.

[4] E. Fox, E. Sudderth, M. Jordan, and A. S. Willsky, "An HDP–HMM for systems with state persistence." in *ICML*, 2008.

[5] C. Rasmussen, "The infinite gaussian mixture model," in *Proc. NIPS*, 2000.

[6] J. Van Gael, Y. Teh, and Z. Ghahramani, "The infinite factorial hidden markov model," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1697–1704, 2009.

[7] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent–dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[8] Y. Jiang and A. Saxena, "Infinite latent conditional random fields for modeling environments through humans." in *Robotics: Science and Systems*, 2013.

[9] P. Orbanz and J. Buhmann, "Nonparametric bayes image segmentation," *Int'l Journal in Computer Vision*, vol. 77, pp. 25–45, 2008.

[10] S. Chatzis and G. Tsechpenakis, "The infinite hidden markov random field model," *IEEE Trans. Neural Networks*, vol. 21, no. 6, pp. 1004–1014, 2010.

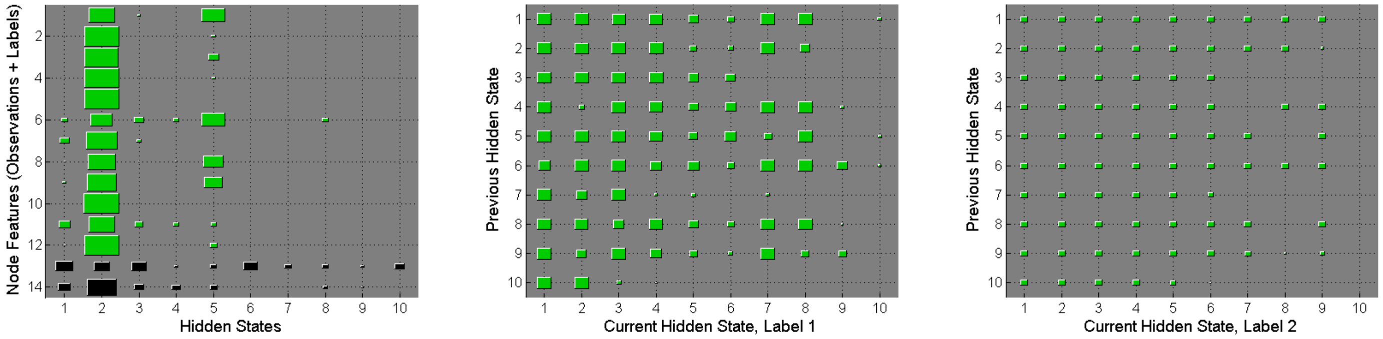
[11] K. Bousmalis, S.Zafeiriou, L.-P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 170–177, 2013.

[12] Z. Ghahramani and M. Beal, "Propagation algorithms for variational bayesian learning," *Advances in Neural Information Processing Systems*, vol. 13, pp. 507–513, 2001.

[13] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, vol. 2, no. 1. Citeseer, 2005, p. 1.

[14] K. Bousmalis, L.-P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE AFGR*, 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2388228, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) π_x (green) and π_y (black)—$L = 10$

(b) π_e, Label 1—$L = 10$

(c) π_e, Label 2—$L = 10$

Fig. 5: Hinton Diagrams of π-quantities in node and edge features of variational HCRF-DPM models with $L = 10$ for PAIN2. The first column presents the π-quantities for node features: π_x for observation features in green, π_y for labels in black. The second and third columns present the π_e-quantities for labels 1 and 2 respectively. See text for additional details

[15] D. Blei and M. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and B. D.M., "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2006.

[17] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[18] D. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *IEEE Trans. on Automatic Control*, vol. 21, pp. 174–184, 1976.

[19] ——, *Nonlinear Programming*. Athena Scientific, 1999.

[20] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[21] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proc. IEEE ACII*, vol. 2, 2009, pp. 96–99.

[22] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Proc. IEEE ACII*, 2009, pp. 1–9.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR — Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. IEEE ACII*, 2009, pp. 1–6.

[24] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Proc. IEEE AFGR*, 2011, pp. 57 –64.

[25] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system," Salt Lake City: Research Nexus, 2002.

Konstantinos Bousmalis Konstantinos Bousmalis received his B.S. degree in Computer Science from Lafayette College, Easton, PA, USA in 2005 and his M.Sc. in Artificial Intelligence from the University of Edinbugh, Edinburgh, UK in 2007. He is due to finish his Ph.D. degree from Imperial College London, London, UK in 2014. His research interests include Bayesian nonparametrics, nonnegative matrix factorization, conditional random fields among others. Konstantinos is a reciepent of a Google Europe Ph.D. Fellowship (2011-2014).

Stefanos Zafeiriou (M09) received the B.Sc. and Ph.D. degrees (Hons.) in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2007, respectively. He is currently a Lecturer with the Department of Computing, Imperial College London, London, U.K., where he was awarded one of the prestigious Junior Research Fellowships. He received various scholarships and awards during his undergraduate, Ph.D. and post-doctoral studies. He has co-authored more than 60 technical papers, including more than 30 papers in the most prestigious journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the International Journal of Computer Vision, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, the IEEE TRANSACTIONS ON NEURAL NETWORKS, Data Mining and Knowledge Discovery, and Pattern Recognition. Dr. Zafeiriou is an Associate Editor of the Image and Vision Computing Journal and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSPART B: CYBERNETICS. He has served as a Program Committee Member for a number of the IEEE international conferences.

Louis-Philippe Morency is a Research Assistant Professor in the Department of Computer Science at the University of Southern California (USC) and leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab) at the USC Institute for Creative Technologies. He received his Ph.D. and Master degrees from MIT Computer Science and Artificial Intelligence Laboratory. In 2008, Dr. Morency was selected as one of "AI's 10 to Watch" by IEEE Intelligent Systems. He has received 7 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational models of human communication dynamics. For the past two years, Dr. Morency has been leading a DARPA-funded multi-institution effort called SimSensei which was recently named one of the years top ten most promising digital initiatives by the NetExplo Forum, in partnership with UNESCO.

Maja Pantic Maja Pantic (M98, SM06, F12) is Professor in Affective and Behavioural Computing at Imperial College London, Department of Computing, UK, and at the University of Twente, Department of Computer Science, the Netherlands. She received various awards for her work on automatic analysis of human behaviour including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor in Chief of Image and Vision Computing Journal and as an Associate Editor for both the IEEE Trans. Pattern Analysis and Machine Intelligence and the IEEE Trans. Cybernetics.

Zoubin Ghahramani Zoubin Ghahramani is Professor of Information Engineering at the University of Cambridge, where he leads a group of about 30 researchers. He studied computer science and cognitive science at the University of Pennsylvania, obtained his PhD from MIT in 1995, and was a postdoctoral fellow at the University of Toronto. His academic career includes concurrent appointments as one of the founding members of the Gatsby Computational Neuroscience Unit in London, and as a faculty member of CMU's Machine Learning Department for over 10 years. His current research focuses on nonparametric Bayesian modelling and statistical machine learning. He has also worked on applications to bioinformatics, econometrics, and a variety of large-scale data modelling problems. He has published over 200 papers, receiving 25,000 citations (an h-index of 68). His work has been funded by grants and donations from EPSRC, DARPA, Microsoft, Google, Infosys, Facebook, Amazon, FX Concepts and a number of other industrial partners. In 2013, he received a $750,000 Google Award for research on building the Automatic Statistician. He serves on the advisory boards of Opera Solutions and Microsoft Research Cambridge, on the Steering Committee of the Cambridge Big Data Initiative, and in a number of leadership roles as programme and general chair of the leading international conferences in machine learning: AISTATS (2005), ICML (2007, 2011), and NIPS (2013, 2014). More information can be found at http://mlg.eng.cam.ac.uk.