Variational Hidden Conditional Random Fields with Beta Processes

Chen Luo, Shiliang Sun, Jing Zhao*
Department of Computer Science and Technology
East China Normal University
Shanghai, P. R. China

***Abstract*—Hidden conditional random fields (HCRFs) are an effective method for sequential classification. It extends the conditional random fields (CRFs) by introducing latent variables to represent the hidden states, which helps to learn the hidden structures in the sequential data. In order to enhance the flexibility of the HCRF, Dirichlet processes (DPs) are employed as priors of the state transition probabilities, which allows the model to have countable infinite hidden states. Besides DPs, Beta processes (BPs) are another kinds of prior models for Bayesian nonparametric modeling, which are more suitable for latent feature models. In this paper, we propose a novel Bayesian nonparametric version of the HCRF referred as BP-HCRF, which takes the advantages of the BPs on modeling hidden states. In the BP-HCRF, BPs are employed as priors for the state indicator variables for each sequence, and the modeled sequences can have different state spaces with infinite hidden states. We develop a variational inference approach for the BP-HCRF using the stick-breaking construction of BPs. We conduct experiments on synthetic dataset to demonstrate the effectiveness of our proposed model.**

***Keywords*—Hidden Conditional Random Fields; Beta Processes; Variational Inference; Sequential Classification**

I. Introduction

The conditional random field (CRF) is well known as an effective discriminative model for structural prediction [1]. It has been adopted for various applications, such as part-of-speech tagging in the natural language processing area. In order to capture the hidden structures of data, Quattoni et al. [2] proposed the hidden conditional random field (HCRF) which adapts the original CRF to a latent variable model for dealing with classification problems in structured domains (e.g., sequential data and images). The introducing of latent variables makes the HCRF be able to capture the hidden structures of data, which helps to better model data and further predict categories more accurately.

The original HCRFs often suffer from the limitation that the number of hidden states has to be predefined or determined through model selection. Generally, the best number of hidden states for specific data is selected from large amounts of candidate values by trial and error. The procedure of model selection is often time-consuming and will hinder the application of such models. As a generative counterpart of the HCRF, the hidden Markov model (HMM) has analogous limitations. In the study of the HMM, Dirichlet Processes (DPs) [3], which are infinite dimensional extensions of Dirichlet distribution, were naturally introduced as priors of the transition matrixes of countable infinite hidden states [4]–[6], and have been extended to handle large scale problems [7], [8]. Such resulting models are known as Bayesian nonparametric models where infinite parameters are involved. Inspired by such Bayesian nonparametric extensions of the HMM, DPs have also been introduced into the HCRF. The HCRFs based on DPs were proposed to have infinite hidden states, which are only applicable to discrete features [9]. In the work of Bousmalis et al. [9], due to the infinite number of hidden states, the Markov chain Monte Carlo (MCMC) sampling method was adopted to infer the model. Further, for overcoming the inefficiency and difficulties in identifying the convergence of the MCMC methods, a variational approach for the HCRF with DP mixtures (DPM-HCRF) was developed [10], where multiple DPs were involved to measure the compatibilities among the hidden states, class labels and observation sequences.

However, all of such DP-based models assume that the hidden state space is shared by the instances from different classes, which could be ill-suited for specific data sets. Another nonparametric prior model is known as Beta processes (BPs). The BP was first proposed in [11] for survival analysis and has become a popular prior model in the study of Bayesian nonparametric models [12]. It has also been introduced into the HMM [13], [14] and applied to trajectory recognition [15]. In such models, each sequence has an infinite dimensional random variable vector that indicates the probability of every hidden state occurring in this sequence. Each dimension of the vector corresponds to a hidden state, and each element of the vector is sampled from a Bernoulli distribution which is a draw of a BP. Inspired by the development of BPs in HMMs, we introduce the BP into the HCRF to propose a BP-based HCRF model called BP-HCRF, which can explicitly model the sequence labels by discrete variables and capture the hidden states by latent variables and state indicators with BP priors. In the BP-HCRF, different sequences may have different state spaces, which is more suitable for realistic situations.

Since there is no tractable solution for the BP based models, some approximation methods are needed for the inference of the BP-HCRF. The key procedure of the approximate inference methods is the construction of BPs. A commonly used representation of the BP is the Indian buffet process (IBP) which is a marginal representation for beta-Bernoulli Process [16]. The IBP has been widely used in infinite latent feature models [17],

*Corresponding author. *Email address:* jzhao2011@gmail.com

[18]. The previous mentioned BP-based HMM [13], [14] also adopted the IBP. With the representation of the IBP, sampling methods for model inference are accessible while the more efficient deterministic approximate methods are not feasible, such as variational inference [19]. Recently, the stick-breaking construction of BPs and the variational inference algorithm were proposed [20] which is in a manner similar to the stick-breaking construction of DPs [10], [21]. The stick-breaking construction of BPs provides the feasibility of variational approaches for the HCRF with the BP prior. We will adapt the mean field variational methods to our proposed BP-HCRF by employing the stick-breaking construction of BPs.

The main contributions of this paper are as follows. First, the new BP-HCRF model is proposed, which adapts the original HCRF to a latent feature model with infinite latent features (i.e., hidden states). Second, a variational approach is developed for the proposed model. Finally, we demonstrate the effectiveness of the proposed model on a synthetic data set, and show its superiority than the original HCRF.

The rest of this paper is organized as follows. First, in Section II, we give an overview of the HCRFs and BPs. Then, in Section III, we present the proposed model, BP-HCRF. In Section IV, we describe the variational inference algorithm for the proposed BP-HCRF, and give details of the update for the latent variables. In addition, the optimization algorithm of model parameters is presented in Section V. Finally, we show the experimental settings and experimental results in Section VI, and make conclusions in Section VII.

II. Related Work

In this section, we give a brief overview on the necessary background, including the HCRF and the stick-breaking construction of BPs.

A. Hidden Conditional Random Fields

The hidden conditional random field (HCRF) is an undirected graphical model with a hidden state layer. Here we only consider a specific form that has a linear chain structure. Different from its generative counterpart, HMM, which a directed graphical model without class labels, the HCRF is a undirected graph model which explicitly models the class labels. Suppose that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ and y are an observation sequence and its label, respectively. The variable $\mathbf{s} = \{s_1, s_2, ..., s_T\}$ has the same length as the observation, with each one representing the corresponding hidden state for every $\mathbf{x}_t$. The HCRF is formulated as

$$p(y, \mathbf{s}|\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{X})} \mathcal{F}\left(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}\right), \tag{1}$$

where the potential function $\mathcal{F}\left(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}\right)$ is

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

and the partition function $Z(\mathbf{X})$ is

$$Z(\mathbf{X}) = \sum_{y,\mathbf{s}} \mathcal{F}\left(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}\right). \tag{3}$$

The model parameters are $\boldsymbol{\theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_e\}$ which are involved in the node, label and edge factors, respectively. The number of model parameters is growing with the dimension of the observation and the number of the hidden sates.

Attributed to the linear chain structure, $Z(\mathbf{X})$ as well as the marginal distributions of hidden states can be computed by the forward-backward algorithm efficiently. The maximum likelihood estimation is used to optimize the model parameters $\boldsymbol{\theta}$.

B. Stick-breaking Construction of Beta Processes

The BP specifies an infinite collection of atoms, and their weights have a degenerate beta distribution [20]. Since the draws of a Beta distribution are in the range of $[0, 1]$, it is amenable to represent the probabilities of occurrences of the hidden states.

Taking $BP(\alpha, H_0)$, $H_0(\Omega) = \gamma$ as an example, the original representation of the stick-breaking construction of the BPs is as follows.

$$\begin{aligned} H &= \sum_{k=1}^{\infty}\sum_{j=1}^{C_j} \hat{V}_{ij}^{(i)} \prod_{l=1}^{i-1}(1 - \hat{V}_{ij}^{(l)})\delta_{\hat{w}_{ij}}, \\ \hat{V}_{ij}^{(l)} &\overset{iid}{\sim} Beta\left(1, \alpha\right), \\ C_i &\overset{iid}{\sim} Poisson\left(\gamma\right), \\ \hat{w}_{ij} &\overset{iid}{\sim} \frac{1}{\gamma} H_0. \end{aligned} \tag{4}$$

This representation describes the procedure of generating a draw of BP. First, in *round 1*, C_1 sticks are prepared, where C_1 obeys Poisson distribution. Let $j = \{1, 2, ..., C_1\}$ be the index of prepared sticks in this round. The jth stick is broken off at the position of $\hat{V}_{1j}^{(1)}$. Every $\hat{V}_{1j}^{(1)}$ for $j = \{1, 2, ..., C_1\}$ is independent and identically distributed with same beta distribution. After *round 1*, C_1 atoms with their weights are generated. Then in *round 2*, C_2 sticks are prepared. Different from that in *round 1*, all sticks would be broken off twice. Thus, C_2 sticks of length-$\hat{V}_{2j}^{(2)}(1 - \hat{V}_{2j}^{(1)})$ are generated. Repeating such operations for infinite times, we can obtain a draw of BP, which is a collection of broken sticks (i.e., atoms) with independent lengths (i.e., weights or measures).

Considering the feasibility of variational inference, there is another kinds of stick-breaking construction for the BPs which can be expressed as follows [20].

$$
\begin{aligned}
H &= \sum_{k=1}^{\infty} V_k e^{-T_k} \delta_{w_k}, \\
V_k &\overset{iid}{\sim} Beta\left(1, \alpha\right), \\
T_k &\sim Gamma\left(d_k - 1, \alpha\right), \\
\sum_{k=1}^{\infty} \mathbf{1}_{d_k}\left(r\right) &\overset{iid}{\sim} Poisson\left(\gamma\right), \\
w_k &\overset{iid}{\sim} \frac{1}{\gamma} H_0.
\end{aligned}
\tag{5}
$$

This construction is actually equivalent to the original construction expressed in Eq. 4 although it has different formulations [20]. The expression in Eq. 5 is much simpler which will lead to easier variational inference. Thus we employ the stick-breaking construction in [20] to construct the BP in our model.

III. Hidden Conditional Random Fields with Beta Processes

First, we exhibit the graphical representation of our proposed BP-HCRF in Fig. 1 to show the model assumptions. In Fig. 1, $\mathbf{X}$ and y represent a sequence and its label, respectively. The random variable $\mathbf{z} = \{z_1, z_2, ..., z_\infty\}$ is introduced to indicate whether a hidden state occurs in the sequence. We assume that $\mathbf{z}$ obeys the Bernoulli process

$$
\mathbf{z} \sim BeP(H), \tag{6}
$$

where H is a draw from a BP. By introducing H as the parameter of the Bernoulli process, the resulting process is a beta-Bernoulli process which is a collection of Bernoulli random variables. The Bernoulli random variables have the same atoms as the BP, and the weights of the atoms are the parameters of the Bernoulli distributions, i.e., z_k has a Bernoulli distribution with $p(z_k = 1) = V_k e^{-T_k}$.

Then, according to the assumptions as shown in Fig. 1, we introduce the joint distribution of BP-HCRF,

$$
p(y, \mathbf{s}, \pi, \mathbf{z} | \mathbf{X}, \theta) = p(y, \mathbf{s} | \mathbf{X}, \pi, \mathbf{z}, \theta) p(\boldsymbol{\pi}) p(\mathbf{z}). \tag{7}
$$

The specific expression of $p(\mathbf{z})$ is omitted for clarity, and

$$
p(y, \mathbf{s} | \mathbf{X}, \pi, \mathbf{z}, \theta) = \frac{1}{Z(\mathbf{X})} \mathcal{F}\left(y, \mathbf{s}, \pi, \mathbf{z}, \mathbf{X}, \theta\right), \tag{8}
$$

where $Z(\mathbf{X})$ is the normalizing constant that makes $p(\cdot)$ become a probability distribution. The formulation of $Z(\mathbf{X})$ is given by

$$
Z(\mathbf{X}) = \sum_{y, \mathbf{s}} \mathcal{F}\left(y, \mathbf{s}, \pi, \mathbf{z}, \mathbf{X}, \theta\right). \tag{9}
$$

Particularly in our model, the variables $\boldsymbol{\pi} = \{\boldsymbol{\pi}_x, \boldsymbol{\pi}_y, \boldsymbol{\pi}_e\}$ measure the compatibility among the observation $\mathbf{X}$, label y and hidden state $\mathbf{s}$ which are involved in the node, label and edge factors, respectively. We have

$$
\begin{aligned}
\pi_x &= \{\pi_x(h_k | i)\}_{k=1, i=1}^{\infty, d}, \\
\pi_y &= \{\pi_y(h_k | y)\}_{k=1, y=1}^{\infty, |\mathcal{Y}|}, \\
\pi_e &= \{\pi_e(h_k, y | h_{k'})\}_{k=1, k'=1, y=1}^{\infty, \infty, |\mathcal{Y}|}.
\end{aligned}
\tag{10}
$$

For each type of latent variables, we assume that they are mutually independent and have the same prior distribution,

$$
\begin{aligned}
\pi_x(h_k | i) &\sim Beta(1, \alpha_x), \\
\pi_y(h_k | y) &\sim Beta(1, \alpha_y), \\
\pi_e(y, h_k | h_{k'}) &\sim Beta(1, \alpha_e).
\end{aligned}
\tag{11}
$$

Thus, the distribution of $\boldsymbol{\pi}$ can be factorized as

$$
p(\boldsymbol{\pi}) = p(\boldsymbol{\pi}_x) p(\boldsymbol{\pi}_y) p(\boldsymbol{\pi}_e), \tag{12}
$$

and further,

$$
\begin{aligned}
p(\boldsymbol{\pi}_x) &= \prod_{k=1, i=1}^{\infty, d} p(\pi_x(h_k | i)), \\
p(\boldsymbol{\pi}_y) &= \prod_{k=1, y=1}^{\infty, |\mathcal{Y}|} p(\pi_y(h_k | y)), \\
p(\boldsymbol{\pi}_e) &= \prod_{k=1, k'=1, y=1}^{\infty, \infty, |\mathcal{Y}|} p(\pi_e(h_k, y | h_{k'})).
\end{aligned}
\tag{13}
$$

Finally, by introducing the variables $\boldsymbol{\pi}$ and $\mathbf{z}$ into the potential function, $\mathcal{F}$ can be expressed as

$$
\begin{aligned}
&\mathcal{F}\left(y, \mathbf{s}, \pi, \mathbf{z}, \mathbf{X}, \theta\right) = \\
&\exp\left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} \theta_x\left(s_t, i\right) f_t\left(i\right) \log \pi_x\left(s_t | i\right) z_{s_t} \right. \\
&+ \sum_{t=1}^{T} \theta_y\left(s_t, y\right) \log \pi_y\left(s_t | y\right) z_{s_t} \\
&\left. + \sum_{t=2}^{T} \theta_e\left(s_t, s_{t-1}, y\right) \log \pi_e\left(s_t, y | s_{t-1}\right) z_{s_t} z_{s_{t-1}} \right\}.
\end{aligned}
\tag{14}
$$

From Eq. 14, we can see that the BP-HCRF is more powerful than the HCRF by introducing the probabilities of each hidden state $p(\mathbf{z})$ and the compatibilities between hidden states, labels and observations $\boldsymbol{\pi}$. The effect of $\boldsymbol{\pi}$ is similar with model parameters $\boldsymbol{\theta}$, because they are shared in all sequences. Meanwhile, every observation sequence $\mathbf{X}_i$ has a collection of latent variables $\mathbf{z}_i$ which repesents the hidden structure. Thus, $\mathbf{z}_i$ is a local factor which could provide more detailed information. The latent variable $\mathbf{z}$ affects the potential function as follows. When $q(z_k = 1)$ is smaller, the potential function will be smaller which reduces the contribution of unrelated hidden state h_k. Because $\log \boldsymbol{\pi}$ is negative, the model parameters $\boldsymbol{\theta}$ are restricted in R^+ and optimized through their logarithm, which ensures the satisfaction of such restriction.

Compared with the DPM-HCRF [10], the BP-HCRF has different formulations of the variables $\{\pi_x, \pi_y, \pi_e\}$ with the

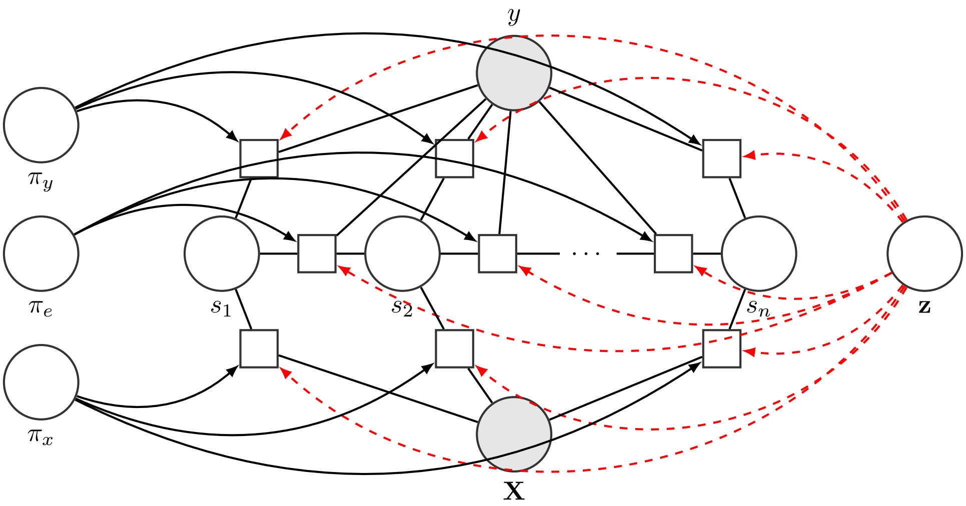

Fig. 1: The graphical representation of the BP-HCRF.

DPM-HCRF, and the BP-HCRF will be more flexible, because there are less restrictions on such variables. Specifically, $\boldsymbol{\pi}$ are the draws of multiple DPs in the DPM-HCRF [10] where the summation of the weights of each draw from a DP should be one.

IV. Variational Inference Algorithm

We adopt the mean-field variational inference method to approximate the posterior distribution of the hidden variables. The joint distribution of an observation sequence and its related hidden variables in our model is

$$\begin{aligned} p(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X}, \boldsymbol{\theta}) =& p(\alpha)p(\gamma)p(\mathbf{d}|\gamma)p(\pi) \\ & \prod_{k=1}^{\infty} p(V_k|\alpha)p(T_k|d_k, \alpha)p(z_k|V_k, d_k, T_k) \\ & p(y, \mathbf{s}|X, \pi, \mathbf{z}), \end{aligned} \tag{15}$$

where $\mathbf{z}$ and $\mathbf{s}$ are local factors. The other latent variables such as $\boldsymbol{\pi}$, α and γ as well as the model parameters are shared by all observation sequences. The variational distribution is assumed as

$$\begin{aligned} & q(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X}, \boldsymbol{\theta}) \\ =& q(\alpha)q(\gamma)q(\pi) \prod_{k=1}^{K} q(d_k)q(V_k)q(T_k)q(z_k)q(y, \mathbf{s}|\mathbf{X}), \end{aligned} \tag{16}$$

where

$$\begin{aligned} & q(y, \mathbf{s}|\mathbf{X}) \\ =& q(y, s_1|\mathbf{x}_1) \prod_{t=2}^{T} q(y, s_t|s_{t-1}, \mathbf{x}_t) \\ =& \prod_{i=1}^{d} q(s_1|i)q(s_1|y) \prod_{t=2}^{T} \prod_{i=1}^{d} q(s_t|i)q(s_t|y)q(s_t, y|s_{t-1}). \end{aligned} \tag{17}$$

In the framework of variational inference, we optimize the approximate variational distribution by minimizing the Kullback-Liebler divergence

$$KL[q(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X})||p(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X})], \tag{18}$$

which is equivalent to optimizing the variational lower bound given by

$$\begin{aligned} \mathcal{L}(q(\cdot)) =& -\mathbb{E}_{q(y,\mathbf{s},\pi,\mathbf{z}|X)} \log \mathcal{F}\left(y, \mathbf{s}, \pi, \mathbf{z}, \mathbf{X}, \boldsymbol{\theta}\right) p(\pi)p(\mathbf{z}) \\ & + \mathbb{E}_{q(y,\mathbf{s},\pi,\mathbf{z}|\mathbf{X})} \log q(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X}) + const, \end{aligned} \tag{19}$$

where the term $\log Z(\mathbf{X})$ is absorbed into the const, because its value does not affect the optimization of the variational distribution $q(y, \mathbf{s}, \pi, \mathbf{z}|\mathbf{X})$.

We define the variational distribution of each latent variable as

$$\begin{aligned} q(\pi_{\mathbf{x}}(h_k|i)) &= Beta(\mu_{x,1}(k, i), \mu_{x,2}(k, i)), \\ q(\pi_{\mathbf{y}}(h_k|y)) &= Beta(\mu_{y,1}(k, y), \mu_{y,2}(k, y)), \\ q(\pi_{\mathbf{e}}(h_k, y|h_{k'})) &= Beta(\mu_{e,1}(k, k', y), \mu_{e,2}(k, k', y)), \\ q(z_k) &= Bernoulli(\phi_{k1}, \phi_{k2}), \\ q(d_k) &= Multinomial(d_k|\psi_k), \\ q(T_k) &= Gamma(T_k|u_k, v_k), \\ q(V_k) &= Beta(V_k|a, b), \\ q(\alpha) &= Gamma(\alpha|\kappa_1, \kappa_2), \\ q(\gamma) &= Gamma(\gamma|\tau_1, \tau_2). \end{aligned} \tag{20}$$

For sequential classification tasks, each factor of $q(y, \mathbf{s}|\mathbf{X})$ is a discrete distribution, see the factorization in Eq. 17. The index k in the variational distribution is in the range of $\{1, 2, ..., K\}$ where K is the truncation level of the BP. In other words, the maximum count of hidden states (i.e., sticks) is limited to K. Another truncation R is the total rounds of the stick-breaking construction which means that the stick-breaking procedure is performed at most R rounds. Accordingly, the dimension of $\mathbf{d}$ is set to R.

Given the above assumptions, the update for the related latent variables can be derived as follows.

A. Update for $q(\boldsymbol{\pi})$

The $\boldsymbol{\pi}_x$, $\boldsymbol{\pi}_y$ and $\boldsymbol{\pi}_e$ are updated individually through taking partial derivative of $\mathcal{L}(q(\cdot))$ with respect to $q(\boldsymbol{\pi})$ and then setting such derivatives to zero. The resulting optimal posterior distributions of $\boldsymbol{\pi}$ are still beta distributions. The updating equations of $\boldsymbol{\pi}_x, \boldsymbol{\pi}_y$ and $\boldsymbol{\pi}_e$ are given below, respectively.

For $\boldsymbol{\pi}_x$, the resulting optimal variational distribution is

$$\begin{aligned} & q(\pi_x(h_k|i)) \\ =& Beta\Big(\sum_t f_t(i)\theta_x(h_k,i)q(s_t=h_k|i)+1, \\ & \sum_t \sum_{h_j\neq h_k} f_t(i)\theta_x(h_k,i)q(s_t=h_j|i)+\alpha_x\Big). \end{aligned} \tag{21}$$

For $\boldsymbol{\pi}_y$, we have

$$\begin{aligned} q(\pi_y(h_k|y)) = Beta\Big(&\sum_t \theta_y(h_k,y)q(s_t=h_k|y)+1, \\ & \sum_t\sum_{h_j\neq h_k} \theta_y(h_k,y)q(s_t=h_j|y)+\alpha_y\Big). \end{aligned} \tag{22}$$

For $\boldsymbol{\pi}_e$, the approximate posterior distribution is given by

$$\begin{aligned} & q(\pi_e(y,h_k|h_{k'})) \\ =& Beta\Big(\sum_{t=2}^{T}\theta_e(h_k,h_{k'},y)q(s_t=h_k,y|s_{t-1}=h_{k'})+1, \\ & \sum_{t=2}^{T}\sum_{yl\neq y}\theta_e(h_k,h_{k'},yl)q(s_t=h_k,y|s_{t-1}=h_{k'}) \\ & +\sum_{h_j\neq h_k}\theta_e(h_j,h_{k'},yll)q(s_t=h_j,yl|s_{t-1}=h_{k'})+\alpha_e\Big). \end{aligned} \tag{23}$$

B. Update for $q(\mathbf{z})$

The variational distribution of $\mathbf{z}_k$ is defined as a Bernoulli distribution. Thus, the parameters of such discrete distributions can be obtained by normalizing the following quantities

$$\begin{aligned} q(z_k=1) \propto & f_1\exp\{\int q(V_k)\log(V_k)dV_k \\ & -\psi_k(r>1)\int q(T_k)\log(T_k)dT_k\}, \\ q(z_k=0) \propto & f_0\exp\{\psi_k(1)\int q(V_k)\log(1-V_k)dV_k \\ & -\int q(V_k)q(T_k)\sum_m \frac{1}{m}(V_ke^{-T_k})^m dV_k dT_k\}, \end{aligned} \tag{24}$$

where the approximation procedure

$$\begin{aligned} & \int q(V_k)q(T_k)(1-V_ke^{-T_k})dV_kdT_k \\ \simeq & \int q(V_k)q(T_k)\sum_{m=1}^{m}(V_ke^{-T_k})^m dV_kdT_k \end{aligned} \tag{25}$$

is adopted [20], and $\log f_1, \log f_2$ are given by the derivatives of $\mathcal{L}(q(\cdot))$ with respect to the Bernoulli parameters ϕ_{k1} and ϕ_{k2}, respectively. As z_k is a Bernoulli variable, the expectation in the $\mathcal{L}(q(\cdot))$ with respect to $q(z_k)$ is computed by two parts where $z_k=1$ and $z_k=0$, and so are the derivatives.

C. Update for $q(y,\mathbf{s}|X)$

The distribution $q(y,\mathbf{s}|X)$ involves the quantities $q(s_t=h_k|i)$, $q(s_t=h_k|y)$ and $q(s_t=h_k|s_{t-1}=h_{k'})$, which are all discrete distributions. Following [10], we can derive the solutions of $\log q(y,\mathbf{s}|X)$ as

$$\log q(\mathbf{s},y|\mathbf{X}) = \mathbb{E}_{q(\mathbf{z})q(\pi)}\log\mathcal{F}(y,\mathbf{s},\pi,\mathbf{z},\mathbf{X},\boldsymbol{\theta})+const. \tag{26}$$

By using the potential function defined in Eq. 14, the latent variables $\mathbf{z}$ and π can be decomposed into two parts. Thus such quantities can be computed by normalizing the following quantities.

$$\begin{aligned} & q(s_t=h_k|i) \\ \propto & \exp\{f_t(i)\theta_x(k,i) \\ & \{\mathbb{E}_{q(\boldsymbol{\pi}_x)}[\log\pi_x(s_t=h_k|i)+\sum_{j\neq k}\log(1-\pi_x(s_t=h_k|i))] \\ & +\mathbb{E}_{q(\mathbf{z})}[\log z_k+\sum_{j\neq k}\log(1-z_k))]\}\}, \end{aligned} \tag{27}$$

which means the distribution of state variable s_t being h_k given the ith dimension of $\mathbf{x}_t$. Similarly,

$$\begin{aligned} & q(s_t=h_k|y) \\ \propto & \exp\{\theta_y(k,y) \\ & \{\mathbb{E}_{q(\boldsymbol{\pi}_y)}[\log\pi_y(s_t=h_k|y)+\sum_{j\neq k}\log(1-\pi_y(s_t=h_k|y))] \\ & +\mathbb{E}_{q(\mathbf{z})}[\log z_k+\sum_{j\neq k}\log(1-z_k))]\}\}, \end{aligned} \tag{28}$$

and

$$\begin{aligned} & q(s_t=h_k,y|s_{t-1}=h_{k'}) \\ \propto & \exp\{\theta_y(k,y) \\ & \{\mathbb{E}_{q(\pi_y)}[\log\pi_y(s_t=h_k|y)+\sum_{j\neq k}\log(1-\pi_y(s_t=h_k|y))] \\ & +\mathbb{E}_{q(\mathbf{z})}[\log z_k+\sum_{j\neq k}\log(1-z_k))]\} \\ & +\theta_e(k,k',y)\{\mathbb{E}_{q(\boldsymbol{\pi}_e)}[\log\pi_y(s_t=h_k,y|s_{t-1}=h_{k'}) \\ & +\sum_{j\neq k}\log(1-\pi_y(s_t=h_k,y|s_{t-1}=h_{k'}))] \\ & +\mathbb{E}_{q(\mathbf{z})}[\log z_k+\sum_{j\neq k}\log(1-z_k))]+\mathbb{E}_{q(\mathbf{z})}[\log z_{k'}]\}\}. \end{aligned} \tag{29}$$

The above quantities can be normalized to form discrete distributions which are the necessary quantities in the update for $\boldsymbol{\pi}$ and $\mathbf{z}$.

For the rest of the latent variables, $\{T_k, V_k, d_k\}_{k=1}^K$, α and γ, the updates of such variables are independent of observations which are similar to the results in [20].

V. Optimization for Model Parameters

We have obtained the variational distribution of latent variables $\boldsymbol{\pi}$ and $\mathbf{z}$. The model parameters $\boldsymbol{\theta}$ can be optimized through the variational EM algorithm where the latent variables are integrated out by taking expectation with respect to their approximate posterior distributions. Further, the model parameters, including $\boldsymbol{\theta}_x$, $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_e$, can be optimized through gradient-based algorithms, such as limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm. The optimization objective function is the lower bound of the log likelihood,

$$\begin{aligned}&\mathcal{L}(\log p(y|X,\boldsymbol{\theta}))\\ &=\log\sum_{\mathbf{s}}\mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\pi})}\mathcal{F}\left(y,\mathbf{s},\pi,\mathbf{z},\mathbf{X},\boldsymbol{\theta}\right)\\ &\quad-\log\sum_{\mathbf{s},y'}\mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\pi})}\mathcal{F}\left(y',\mathbf{s},\pi,\mathbf{z},\mathbf{X},\boldsymbol{\theta}\right).\end{aligned}\tag{30}$$

The gradients of the objective function with respect to the model parameters are as follows. For node parameters $\boldsymbol{\theta}_x$, we have

$$\begin{aligned}&\frac{\partial\mathcal{L}(\log p(y|X,\boldsymbol{\theta}))}{\partial\theta_x(k,i)}\\ &=\sum_{t=1}^{T}p(s_t=h_k|y,\mathbf{X},\boldsymbol{\theta})f_t\left(i\right)\mathbb{E}\log\pi_x\left(k|i\right)z_k\\ &\quad-\sum_{y',t=1}^{T}p(s_t=h_k,y'|\mathbf{X},\boldsymbol{\theta})f_t\left(i\right)\mathbb{E}\log\pi_x\left(k|i\right)z_k.\end{aligned}\tag{31}$$

For the label parameters $\boldsymbol{\theta}_y$, the gradients are given by

$$\begin{aligned}&\frac{\partial\mathcal{L}(\log p(y|X,\boldsymbol{\theta}))}{\partial\theta_y(k,y)}\\ &=\sum_{t=1}^{T}p(s_t=h_k|y,\mathbf{X},\boldsymbol{\theta})\mathbb{E}\log\pi_y\left(k|y\right)z_k\\ &\quad-\sum_{y',t=1}^{T}p(s_t=h_k,y'|\mathbf{X},\boldsymbol{\theta})\mathbb{E}\log\pi_y\left(k|y'\right)z_k.\end{aligned}\tag{32}$$

For the edge parameters $\boldsymbol{\theta}_e$, we have

$$\begin{aligned}&\frac{\partial\mathcal{L}(\log p(y|X,\boldsymbol{\theta}))}{\partial\theta_e(k,k',y)}\\ &=\sum_{t=2}^{T}p(s_t=h_k,s_{t-1}=h_{k'}|y,\mathbf{X},\boldsymbol{\theta})\mathbb{E}\log\pi_e\left(k,y|k'\right)z_kz_{k'}\\ &-\sum_{y',t=2}^{T}p(s_t=h_k,s_{t-1}=h_{k'},y'|\mathbf{X},\boldsymbol{\theta})\mathbb{E}\log\pi_e\left(k,y'|k'\right)z_kz_{k'}.\end{aligned}\tag{33}$$

The marginal distributions $p(s_t = h_k|y,\mathbf{X},\boldsymbol{\theta})$ and $p(s_t = h_k, s_{t-1} = h_{k'}|y,\mathbf{X},\boldsymbol{\theta})$ can be obtained by the forward-backward algorithm along the linear chain. The distributions $p(s_t = h_k, y'|\mathbf{X},\boldsymbol{\theta})$ and $p(s_t = h_k, s_{t-1} = h_{k'}, y'|\mathbf{X},\boldsymbol{\theta})$ are computed by the above two marginal distributions and the class distribution $p(y|\mathbf{X},\boldsymbol{\theta})$ which is computed by $Z(y)/\sum_{y'}Z(y')$ where $Z(y') = \mathbb{E}\sum_{\mathbf{s}}\mathcal{F}\left(y',\mathbf{s},\pi,\mathbf{z},\mathbf{X},\boldsymbol{\theta}\right)$.

TABLE I: Emission Distributions for All Hidden States.

Hidden State	Mean	Variance
h1	21	0.4
h2	23	0.8
h3	36	0.6
h4	26	0.1
h5	11	0.8
h6	8	0.8
h7	46	0.2
h8	1	0.2

TABLE II: Transition Matrix of Class 1.

	h1	h2	h3	h4
h1	0.40	0.40	0.10	0.10
h2	0.10	0.40	0.10	0.40
h3	0.40	0.10	0.40	0.10
h4	0.10	0.10	0.40	0.40

TABLE III: Transition Matrix of Class 2.

	h1	h2	h5	h6
h1	0.10	0.70	0.10	0.10
h2	0.10	0.10	0.70	0.40
h5	0.10	0.10	0.10	0.70
h6	0.70	0.10	0.10	0.10

TABLE IV: Transition Matrix of Class 3.

	h1	h2	h7	h8
h1	0.25	0.60	0.10	0.05
h2	0.25	0.20	0.30	0.25
h7	0.10	0.30	0.30	0.30
h8	0.35	0.05	0.10	0.50

VI. Experiment

We evaluate the performance of the BP-HCRF on a synthetic data set and compare it with the HCRF. The data set is generated by three HMMs with different transition matrixes and emission distributions, which leads to a classification problem with three classes. In total, eight hidden states are defined and two of them are shared in all classes. The corresponding Gaussian emission distributions are shown in Table I. The hidden states h_1 and h_2 are shared in all classes. The transition matrixes of three classes are given in the Table II, Table III and Table IV respectively.

For each class, 100 sequences with length $T = 100$ are generated as the test set. Additional five sequences with the same length for every class are generated as the training set. Since the variational inference may fall into the local optimal solution, we randomly initialize the parameters of variational distribution for times to obtain the final results. The truncation level K is set to 10 and R is set to 4 in our experiments. The experiments are run on randomly split data sets for five times. The performances of the BP-HCRF and the HCRF in terms of average accuracies and corresponding standard deviations are reported in Table V. It is shown that both the BP-HCRF and the HCRF achieve high accuracies

TABLE V: Performances of BP-HCRF and HCRF.

Model	BP-HCRF	HCRF
Accuracy %	$\mathbf{99.93 \pm 0.15}$	99.80 ± 0.30

and our model outperforms the HCRF which demonstrates the effectiveness and superiority of our model.

VII. Conclusion

In this paper, we have proposed a novel Bayesian nonparametric model BP-HCRF with its corresponding variational approach for model inference. The stick-breaking construction of BPs is employed in the proposed model, which enables to include countable infinite hidden states, and provides the feasibility of the variational inference. With the property of BPs, the number of hidden states can be inferred from data, which reduces the cost of model selection. We conducted experiments on synthetic data set, and the experimental results have shown the effectiveness of the proposed BP-HCRF.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Projects 61673179 and 61370175, Shanghai Knowledge Service Platform Project (No. ZF1213), and the Fundamental Research Funds for the Central Universities.

References

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 8th International Conference on Machine Learning*, 2001, pp. 282–289.

[2] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1848–1852, 2007.

[3] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[4] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," *Advances in Neural Information Processing Systems*, vol. 14, pp. 577–584, 2002.

[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 312–319.

[6] E. B. Fox, E. B. Sudderth, M. Jordan, and A. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, pp. 1020–1056, 2011.

[7] N. Foti, J. Xu, D. Laird, and E. Fox, "Stochastic variational inference for hidden Markov models," *Neural Information Processing Systems*, vol. 27, pp. 1–9, 2015.

[8] A. Zhang, S. Gultekin, and J. Paisley, "Stochastic variational inference for the HDP-HMM," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016, pp. 800–808.

[9] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *IEEE Transactions in Neural Networks and Learning Systems*, pp. 170–177, 2013.

[10] K. Bousmalis, S. Zafeiriou, L.-P. Morency, M. Pantic, and Z. Ghahramani, "Variational infinite hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1917–1929, 2015.

[11] N. L. Hjort, "Nonparametric Bayes estimators based on beta processes in models for life history data," *The Annals of Statistics*, vol. 18, pp. 1259–1294, 1990.

[12] J. Paisley and M. I. Jordan, "A constructive definition of the beta process," *ArXiv e-prints*, pp. 1–19, 2016.

[13] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Sharing features among dynamical systems with beta processes," *Neural Information Processing Systems*, vol. 22, pp. 549–557, 2010.

[14] E. B. Fox, M. C. Hughes, E. B. Sudderth, and M. I. Jordan, "Joint modeling of multiple related time series via the beta process with application to motion capture segmentation," *Annals of Applied Statistics*, vol. 8, pp. 1281–1313, 2014.

[15] S. Sun, J. Zhao, and Q. Gao, "Modeling and recognizing human trajectories with beta process hidden markov models," *Pattern Recognition*, vol. 48, pp. 2407–2417, 2015.

[16] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007, pp. 564–571.

[17] Z. Ghahramani and T. L. Griffiths, "Infinite latent feature models and the Indian buffet process," *Advances in Neural Information Processing Systems*, vol. 18, pp. 475–482, 2006.

[18] T. L. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[19] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *ArXiv e-prints*, pp. 1–42, 2016.

[20] J. W. Paisley, L. Carin, and D. M. Blei, "Variational inference for stick-breaking beta process priors," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 889–896.

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