Inference in Hidden Markov Models with Explicit State Duration Distributions

Mike, Chris, **Frank**

***Abstract*—Explicit-state-duration hidden Markov models (EDHMM) are HMMs that have latent states consisting of both discrete state-indicator and discrete state-duration random variables. In contrast to the implicit geometric state duration distribution possessed by the standard HMM, EDHMMs allow the direct parameterisation and estimation of per-state duration distributions. As most duration distributions have support on all positive integers, truncation or other approximations are usually required to perform EDHMM inference. In this letter we borrow from the inference techniques developed for unbounded state-cardinality (nonparametric) variants of the HMM and use them to develop a tuning-parameter free inference approach for EDHMMs. We illustrate the performance gains of this approach using synthetic data.**

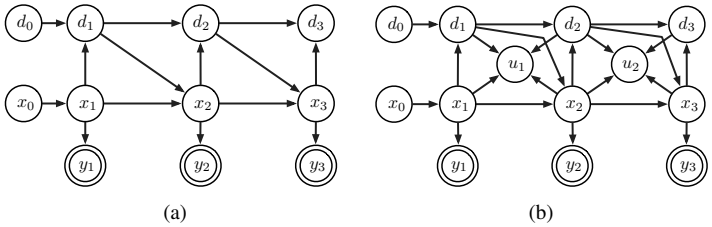


Fig. 1: a) The Explicit Duration Hidden Markov Model. The time left in the current state x_t is denoted d_t. The observation at each point in time is denoted y_t. b) The EDHMM with the additional auxiliary variable u_t used in the beam sampler.

I. Introduction

Hidden Markov models (HMMs) are a fundamental tool for data analysis and exploration. Many variants of the basic HMM have been developed in response to shortcomings in the original HMM formulation [8]. In this paper we address inference in the explicit state duration HMM. By state duration we mean the amount of time an HMM dwells in a state. In the standard HMM specification, a state's duration is implicit and, a priori, distributed geometrically.

The explicit state duration HMM (EDHMM) [8] was developed to allow explicit parameterization and direct inference of state duration distributions. Approximate EDHMM estimation and inference can, in some cases, be performed using a modified forward-backward algorithm; specifically if either the sequence is short or a tight "allowable" duration interval for each state is hard-coded a priori [11]. If the sequence is short then forward-backward can be run on a state representation that allows for all possible durations shorter than the observed sequence length. If the sequence is long then forward-backward only remains computationally tractable if only transitions between "allowed" durations (those that lie within pre-specified allowable intervals) are considered. If the true state durations lie outside those intervals then the resulting model estimates will be incorrect in the sense that the learned duration distributions will reflect only what is allowed given the pre-specified duration intervals (which may poorly describe the true duration distributions).

Our contribution is the development of a procedure for EDHMM inference that does not require any hard pre-specification of duration intervals and, as a result, is no longer approximate. The technique we use to do this is borrowed from sampling procedures developed for nonparametric Bayesian HMM variants [10]. Our key insight is simple: the machinery developed for doing inference in HMMs with a countable number of states is precisely the same as that which is needed for doing inference in an EDHMM with duration distributions over countable support. So, while the EDHMM is a distinctly parametric model, the tools from nonparametric Bayesian inference can be applied such that black-box inference becomes possible and, in fact, efficient.

In this work we show specifically that a "beam-sampling" approach [10] works for non-approximate estimation of EDHMMs, learning both the transition structure and duration distributions simultaneously.

In demonstrating our EDHMM inference technique we consider a synthetic system in which the state-cardinality is known and finite and the distributions over state durations are of scientific interest, we focus on the finite state, unknown duration distribution case. We show that the EDHMM beam sampler performs accurate tracking whilst capturing the duration distributions over the dwell time of each state as well as the probability of transitioning between states.

The remainder of the letter is organised as follows. In Section II we introduce the EDHMM; in Section III we review beam-sampling for the infinite Hidden Markov Model (iHMM) [1] and show how it relates to the EDHMM inference problem; and in Section IV we show results from using the EDHMM to model synthetic data.

II. Explicit Duration Hidden Markov Model

The EDHMM captures the relationships among state x_t, duration d_t, and observation y_t over time t. It consists of four components: the initial state distribution, the transition distributions, the observation distributions, and the duration distributions.

We define the observation sequence $\mathcal{Y} = \{y_1, y_2, \ldots, y_T\}$; the latent state sequence $\mathcal{X} = \{x_0, x_1, x_2, \ldots, x_T\}$; and the remaining time in each segment $\mathcal{D} = \{d_1, d_2, \ldots, d_T\}$, where $x_t \in \{1, 2, \ldots, K\}$ with K the maximum number of

states, $d_t \in \{1, 2, \ldots\}$, and $y_t \in \mathbb{R}^n$. We assume that the Markov chain on the latent states is homogenous, i.e., that $p(x_t = i | x_{t-1} = j, A) = a_{i,j} \forall t$ where A is a $K \times K$ matrix with element $a_{i,j}$ at row i and column j. The prior on A is row-wise Dirichlet with zero prior mass on self-transitions, i.e. $p(a_{i,:}) = \mathrm{Dir}(1/K - 1, \ldots, 0, \ldots 1/K - 1)$ where $a_{i,:}$ is a row vector and the ith Dirichlet parameter is 0. Each state is imbued with its own duration distribution $p(d_t | x_t = k) = p(d_t | \lambda_k)$ with parameter λ_k. Each duration distribution parameter is drawn from a prior $p(\lambda_k)$ which can be chosen in an application specific way. The collection of all duration distribution parameters is $\lambda = \{\lambda_1, \ldots, \lambda_K\}$. Each state is also imbued with an observation generating distribution $p(y_t | x_t = k) = p(y_t | \theta_k)$ with parameter θ_k. Each observation distribution parameter is drawn from a prior $p(\theta_k)$ also to be chosen according to the application. The set of all observation distribution parameters is θ. In the following exposition, explicit conditional dependencies on component distribution parameters are omitted to focus on the particulars unique to the EDHMM.

In an EDHMM the transitions between states are only allowed at the end of a segment:

$$p(x_t | x_{t-1}, d_{t-1}) = \begin{cases} \delta(x_t, x_{t-1}) & \text{if } d_{t-1} > 1 \\ p(x_t | x_{t-1}) & \text{otherwise} \end{cases} \tag{1}$$

where the Kronecker delta $\delta(a, b) = 1$ if $a = b$ and zero otherwise. The duration distribution generates segment lengths at every state switch:

$$p(d_t | x_t, d_{t-1}) = \begin{cases} \delta(d_t, d_{t-1} - 1) & \text{if } d_{t-1} > 1 \\ p(d_t | x_t) & \text{otherwise.} \end{cases} \tag{2}$$

The joint distribution of the EDHMM is

$$p(\mathcal{X}, \mathcal{D}, \mathcal{Y}) = p(x_0) p(d_0) \prod_{t=1}^{T} p(y_t | x_t, \theta) p(x_t | x_{t-1}, d_{t-1}, A) p(d_t | x_t, d_{t-1}, \lambda) \tag{3}$$

corresponding to the graphical model in Figure 1a. Alternative choices to define the duration variable d_t exist; see [3] for details. Algorithm 1 illustrates the EDHMM as a generative model.

III. EDHMM Inference

Our aim is to estimate the conditional posterior distribution of the latent states ($\mathcal{X}$ and $\mathcal{D}$) and parameters (θ, λ and A) given observations $\mathcal{Y}$ by samples drawn via Markov chain Monte Carlo. Sampling θ and A given $\mathcal{X}$ proceeds per usual textbook approaches [2]. Sampling λ given $\mathcal{D}$ is straightforward in most situations. Gibbs sampling $\mathcal{X}$ is possible but, for reasons similar to those in the case of graphical models with chain dependency structure, one would not expect Gibbs sampling to work well for this model. The main contribution of this paper is to show how to generate exact conditional samples of $\mathcal{X}$ and $\mathcal{D}$ efficiently given all other random variables.

A. Forward Filtering, Backward Sampling

We can, in theory, use the forward messages from the forward backward algorithm [8] to sample the conditional posterior distribution of $\mathcal{X}$ and $\mathcal{D}$. To do this we treat each state-duration tuple as a single random variable (introducing the notation $z_t = \{x_t, d_t\}$). Doing so recovers the standard hidden Markov model structure and hence standard forward messages can be used directly. A forward filtering, backward sampler for $\mathcal{Z} = \{z_1, \ldots, x_T\}$ conditioned on all other random variables requires the classical forward messages:

$$\alpha_t(z_t) = \sum_{z_{t-1}} p(z_t | z_{t-1}) p(y_t | z_t) \alpha_t(z_{t-1}) \tag{4}$$

where the transition probability can be factorised according to our modelling assumptions:

$$p(z_t | z_{t-1}) = p(x_t | x_{t-1}, d_{t-1}) p(d_t | d_{t-1}, x_t) p(y_t | x_t). \tag{5}$$

Unfortunately the sum in (4) has an infinite number of terms in the usual case of duration distributions with countably infinite support. The standard approach to EDHMM inference involves truncating these sums in order to make them tractable. A reasonable computational artifice is to truncate all durations at some maximum $d_{\max}$; however, this leads to tremendous unneccesary computation for those states of duration much less than the chosen maximum. In general, setting allowable duration intervals $d_{\min} < d < d_{\max}$ is highly problematic; moreover, poor choices can lead to poor inference.

B. EDHMM Beam Sampling

A recent contribution to inference in the infinite Hidden Markov Model (iHMM) [1] suggests a way around truncation [10]. The iHMM is an HMM with a countable number of states. Computing the forward message for a forward filtering, backward sampler for the latent states in an iHMM also requires a sum over a countable number of elements. The "beam sampling" approach [10], which we can apply largely without modification, is to truncate this sum by introducing a "slice" (see [6]) auxiliary variable $\mathcal{U} = \{u_1, u_2, \ldots, u_T\}$ at each time step. The auxiliary variables are chosen in such a way as to automatically limit each sum in the forward pass to a finite number of terms while allowing inference to remain exact.

The particular choice of auxiliary variable u_t is important. We follow [10] in choosing u_t to be conditionally distributed given the current and previous state and duration in the following way (see the graphical model in Figure 1b):

$$p(u_t | z_t, z_{t-1}) = \frac{\mathbb{I}(0 < u_t < p(z_t | z_{t-1}))}{p(z_t | z_{t-1})}. \tag{6}$$

Given $\mathcal{U}$ it possible to sample the state $\mathcal{X}$ and duration $\mathcal{D}$ conditional posterior. Using notation $\mathcal{Y}_{t_1}^{t_2} = \{y_{t_1}, y_{t_1+1}, \ldots, y_{t_2}\}$ to indicate sub-ranges of a sequence, the new forward mes-

Algorithm 1 Generate Data

sample $x_0 \sim p(x_0)$
sample $d_0 \sim p(d_0)$
for $t = 1, 2, \ldots, T$ **do**
  if $d_{t-1} = 1$ **then**
    a new segment starts:
    sample $x_t \sim p(x_t|x_{t-1})$
    sample $d_t \sim p(d_t|x_t)$
  else
    the segment continues:
    $x_t = x_{t-1}$
    $d_t = d_{t-1} - 1$
  end if
  sample $y_t \sim p(y_t|x_t)$
end for

Algorithm 2 Sample the EDHMM

Initialise parameters A, λ, θ. Initialize u_t small $\forall\, T$
for sweep $\in \{1, 2, 3, \ldots\}$ **do**
  Forward: run (7) to get $\hat{\alpha}_t$ given $\mathcal{U}$ and $\mathcal{Y}\ \forall\, T$
  Backward: sample $z_T \sim \hat{\alpha}_T$
  for $t \in \{T, T-1, \ldots, 1\}$ **do**
    sample $z_{t-1} \sim \mathbb{I}(u_{t+1} < p(z_t|z_{t-1}))\hat{\alpha}_{t-1}$
  end for
  Slice:
  for $t \in \{1, 2, \ldots, T\}$ **do**
    evaluate $l = p(d_t|x_t, d_{t-1})p(x_t|x_{t-1}, d_{t-1})$
    sample $u_{t+1} \sim \text{Uniform}(0, l)$
  end for
  sample parameters A, λ, θ
end for

sages we compute are:

$$\begin{aligned}\hat{\alpha}_t(z_t) &= p(z_t, \mathcal{Y}_1^t, \mathcal{U}_1^{t-1}) \qquad (7)\\ &= \sum_{z_{t-1}} p(z_t, z_{t-1}, \mathcal{Y}_1^t, \mathcal{U}_1^{t-1})\\ &\propto \sum_{z_{t-1}} p(u_{t-1}|z_t, z_{t-1}) p(z_t, z_{t-1}, \mathcal{Y}_1^t, \mathcal{U}_1^{t-1})\\ &= \sum_{z_{t-1}} \mathbb{I}(0 < u_{t-1} < p(z_t|z_{t-1})) p(y_t|z_t) \hat{\alpha}_{t-1}(z_{t-1}).\end{aligned}$$

The indicator function results in non-zero probabilities in the forward message for only those states z_t whose likelihood given z_{t-1} is greater than u_t. The set of z_t's for which this is true will always be finite.

The backwards sampling step recursively samples a state sequence from the distribution $p(z_t|z_{t+1}, \mathcal{Y}, \mathcal{U})$ which can expressed in terms of the forward variable:

$$\begin{aligned}p(z_t|z_{t+1}, \mathcal{Y}, \mathcal{U}) &\propto p(z_{t+1}, z_t, \mathcal{Y}, \mathcal{U})\\ &\propto p(u_t|z_t, z_{t+1}) p(z_{t+1}|z_t) \hat{\alpha}_t(z_t)\\ &\propto \mathbb{I}(0 < u_t < p(z_{t+1}|z_t)) \hat{\alpha}_t(z_t). \quad (8)\end{aligned}$$

The full EDHMM beam sampler is given in Algorithm 2, which makes use of the forward recursion in (7), the slice sampler in (6), and the backwards sampler in (8).

C. Related Work

The need to accommodate explicit state duration distributions in HMMs has long been recognised. Rabiner [8] details the basic approach which expands the state space to include dwell time before applying a slightly modified Baum-Welch algorithm. This approach specifies a maximum state duration, limiting practical application to cases with short sequences and dwell times. This approach, generalised under the name "segmental hidden Markov models", includes more general transitions than those Rabiner considered, allowing the next state and duration to be conditioned on the previous state and duration [5]. Efficient approximate inference procedures were developed in the context of speech recognition [7] and evolved into symmetric approaches suitable for implementation *in-silico* [11].

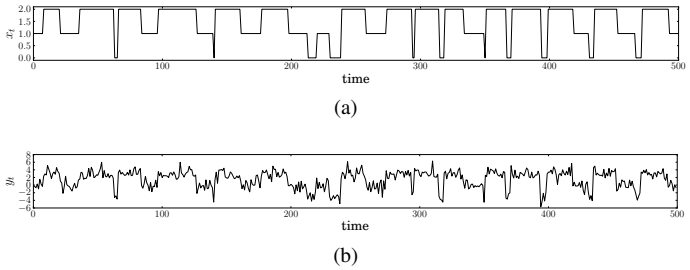
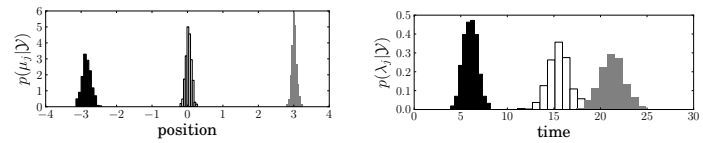


Fig. 2: Example a) state and b) observation sequence generated by the explicit duration HMM. Here K = 3; $p(y_t|x_t = j) = \text{N}(\mu_j, 1)$ with $\mu_1 = -3$, $\mu_2 = 0$, and $\mu_3 = 3$; and $p(d_t|x_t = j) = \text{Poisson}(\lambda_j)$ with $\lambda_1 = 5$, $\lambda_2 = 15$, and $\lambda_3 = 20$.



(a) (b)

Fig. 3: Samples from the posterior distributions of a) the observation distribution means and b) the duration distribution rate parameters.

Recently, a "sticky" variant of the hierarchical Dirichlet process HMM (HDP-HMM) has been developed [4]. The HDP-HMM has countable state-cardinality [9] allowing estimation of the number of states in the HMM; the sticky aspect addresses long dwell times by introducing a parameter in the prior that favours self-transition.

IV. Experiments

A. Synthetic Data

The first experiment uses the 500 data points shown in Figure 2 generated from an EDHMM with three states. To generate the data the duration distributions were chosen to

be Poisson with rates $\lambda_1 = 5$, $\lambda_2 = 15$, $\lambda_3 = 20$; each observation distribution was Gaussian with means of $\mu_1 = -3$, $\mu_2 = 0$, and $\mu_3 = 3$, each with a variance of 1. The transition distributions A were set to

$$\begin{bmatrix} 0 & 0.3 & 0.7 \\ 0.6 & 0 & 0.4 \\ 0.3 & 0.7 & 0 \end{bmatrix}.$$

For inference, broad, uninformative priors were chosen for the parameters of the duration and observation distributions. The parameters of the observation distribution for all states were given a normal-inverse-Wishart (N-IW) prior with parameters $\nu_0 = 2$, $\Lambda_0 = 1$, $\kappa = 0.1$ and $\mu_0 = 0$. The rate parameters for all states were given $\mathrm{Gamma}(1, 10^5)$ priors.

One thousand samples were collected from the EDHMM beam sampler after a burn-in of 500 samples. The learned posterior distribution of the state duration parameters and means of the observation distributions are shown in Figure 3. The EDHMM achieves high accuracy in the estimated posterior distribution of the observation means, despite the overlap in observation distributions. The rate parameter distributions are reasonably estimated given the small number of observed segments.

A second experiment was performed using similar synthetic data to demonstrate the ability of the EDHMM to distinguish between states having differing duration distributions but the same observation distribution. Here the same model was used as above except data were generated using $\mu_1 = 0$, $\mu_2 = 0$, and $\mu_3 = 3$. The same priors were used as above, and again the sampler was allowed to run for 1000 samples with a burn-in of 500 samples.

Figure 4 shows that the sampler clearly separates the high state associated with μ_3 from the other two states and clearly reveals the presence of two low states with differing duration distributions. That it is possible for the EDHMM to recover two states with differing duration distributions requires a third state. Figure 4b shows posterior samples that indicate that the model is mixing over ambiguities about states 0 and 1 as it should.

Finally, Figure 5 shows that the number of allowed transitions decays with the number of iterations, significantly speeding up the runtime of the later samples after the burn-in period.

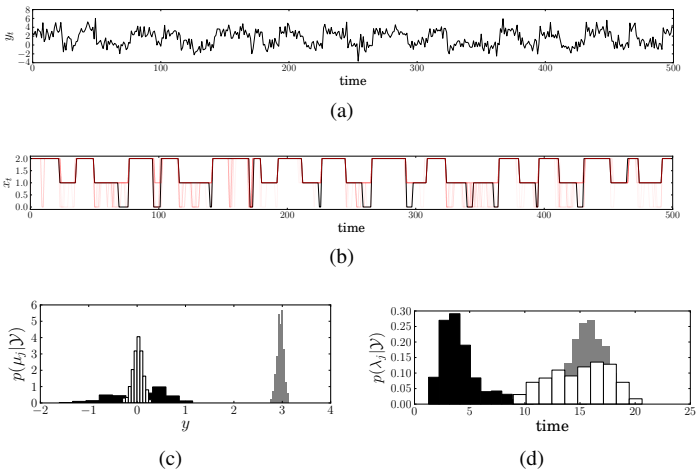


Fig. 4: Results of the beam sampler applied to a system that generates states with identical observation distribution but differing durations. The observations are shown in a), and the true states are shown in b) overlaid with 20 randomly selected state traces produced by the sampler after burn-in. Samples from the posterior observation distribution mean are shown in c), and samples from the posterior duration distribution rates are shown in d).

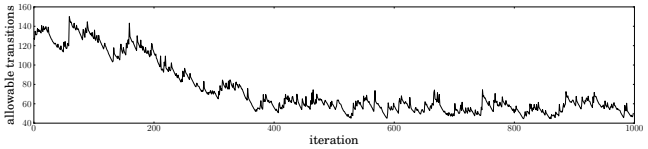


Fig. 5: Mean number of transitions visited over the first 1000 iterations of the beam sampler.

V. Discussion

We presented a beam sampler for the explicit state duration HMM. This sampler draws state sequences from the true posterior distribution without any need to make truncation approximations. It remains future work to combine the explicit state duration HMM and the iHMM. Python code associated with the EDHMM is available online.[1]

References

[1] Matthew J Beal, Z Ghahramani, and C E Rasmussen. The Infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, 1:577–584, 2002.

[2] C M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Silvia Chiappa. Unified Treatment of Hidden Markov Switching Models. April 2011.

[4] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for systems with state persistence. *Proceedings of the 25th International Conference on Machine Learning (2008)*, 25:312–319, 2008.

[5] M J F Gales and S J Young. The Theory of Segmental Hidden Markov Models. Technical report, Cambride University Engineering Department, 1993.

[6] Radford M Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

[7] M Ostendorf, V V Digalakis, and O A Kimball. From {HMM}s to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1996.

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

[9] Y W Teh, M I Jordan, M J Beal, and D M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[10] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. *Proceedings of the 25th International Conference on Machine Learning (2008)*, 25:1088–1095, 2008.

[11] Hisahi Yu, Shun-Zeng, Kobayashi. Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Transactions on Signal Processing*, 54(5):1947–1951, 2006.

[1] http://github.com/mikedewar/EDHMM