Stochastic Variational Inference for Bayesian Time Series Models

Matthew James Johnson MATTJJ@CSAIL.MIT.EDU
Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA USA

Alan S. Willsky WILLSKY@MIT.EDU
Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA USA

Abstract

Bayesian models provide powerful tools for analyzing complex time series data, but performing inference with large datasets is a challenge. Stochastic variational inference (SVI) provides a new framework for approximating model posteriors with only a small number of passes through the data, enabling such models to be fit at scale. However, its application to time series models has not been studied.

In this paper we develop SVI algorithms for several common Bayesian time series models, namely the hidden Markov model (HMM), hidden semi-Markov model (HSMM), and the nonparametric HDP-HMM and HDP-HSMM. In addition, because HSMM inference can be expensive even in the minibatch setting of SVI, we develop fast approximate updates for HSMMs with durations distributions that are negative binomials or mixtures of negative binomials.

1. Introduction

Bayesian time series models can be applied to complex data in many domains, including data arising from behavior and motion (Fox et al., 2010; 2011), home energy consumption (Johnson & Willsky, 2013), physiological signals (Lehman et al., 2012), single-molecule biophysics (Lindén et al., 2013), brain-machine interfaces (Hudson, 2008), and natural language and text (Griffiths et al., 2004; Liang et al., 2007). However, scaling inference in these models to large datasets is a challenge.

Many Bayesian inference algorithms require a complete pass over the data in each iteration and thus do not scale well. In contrast, some recent Bayesian inference methods require only a small number of passes and can even operate in the single-pass or streaming settings (Broderick et al., 2013). In particular, stochastic variational inference (SVI) (Hoffman et al., 2013) provides a general framework for scalable inference based on mean field and stochastic gradient optimization. However, while SVI has been studied extensively for topic models (Hoffman et al., 2010; Wang et al., 2011; Bryant & Sudderth, 2012; Wang & Blei, 2012; Ranganath et al., 2013; Hoffman et al., 2013), it has not been applied to time series.

In this paper, we develop SVI algorithms for the core Bayesian time series models based on the hidden Markov model (HMM), namely the Bayesian HMM and hidden semi-Markov model (HSMM), as well as their nonparametric extensions based on the hierarchical Dirichlet process (HDP), the HDP-HMM and HDP-HSMM. Both the HMM and HDP-HMM are ubiquitous in time series modeling, and so the SVI algorithms developed in Sections 3 and 4 are widely applicable.

The HSMM and HDP-HSMM extend their HMM counterparts by allowing explicit modeling of state durations with arbitrary distributions. However, HSMM inference subroutines have time complexity that scales quadratically with the observation sequence length, which can be expensive even in the minibatch setting of SVI. To address this shortcoming, in Section 5 we develop a new method for Bayesian inference in (HDP-)HSMMs with negative binomial durations that allows approximate SVI updates with time complexity that scales only linearly with sequence length. The methods in this paper also provide the first batch mean field algorithm for HDP-HSMMs.

Our code is available at github.com/mattjj/pyhsmm.

2. Background

Here we review the key ingredients of SVI, namely stochastic gradient algorithms, the mean field variational inference problem, and natural gradients of the mean field objective for models with complete-data conjugacy.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

2.1. Stochastic gradient ascent

Consider the optimization problem

$$\max_{\phi} f(\phi, \bar{y}) \quad \text{where} \quad f(\phi, \bar{y}) = \sum_{k=1}^{K} g(\phi, \bar{y}^{(k)})$$

and where $\bar{y} = \{\bar{y}^{(k)}\}_{k=1}^{K}$ is fixed. Then if $\hat{k}$ is sampled uniformly over $\{1, 2, \ldots, K\}$, we have

$$\nabla_\phi f(\phi) = K \cdot \mathbb{E}_{\hat{k}} \left[\nabla_\phi g(\phi, \bar{y}^{(\hat{k})}) \right].$$

Thus we can generate approximate gradients of the objective using only one $\bar{y}^{(k)}$ at a time. A stochastic gradient algorithm for a sequence of stepsizes $\rho^{(t)}$ and positive definite matrices $G^{(t)}$ is given in Algorithm 1. From standard results (Robbins & Monro, 1951; Bottou, 1998), if $\sum_{t=1}^{\infty} \rho^{(t)} = \infty$ and $\sum_{t=1}^{\infty} (\rho^{(t)})^2 < \infty$ and $G^{(t)}$ has uniformly bounded eigenvalues, then the algorithm converges to a stationary point, i.e. $\phi^* \triangleq \lim_{t\to\infty} \phi^{(t)}$ satisfies $\nabla_\phi f(\phi^*, \bar{y}) = 0$.

Since each update in the stochastic gradient ascent algorithm only operates on one $\bar{y}^{(k)}$, or *minibatch*, at a time, it can scale to the case where $\bar{y}$ is large.

Algorithm 1 Stochastic gradient ascent

Initialize $\phi^{(0)}$
for $t = 1, 2, \ldots$ **do**
  $\hat{k}^{(t)} \leftarrow \text{Uniform}(\{1, 2, \ldots, K\})$
  $\phi^{(t)} \leftarrow \phi^{(t-1)} + \rho^{(t)} K G^{(t)} \nabla_\phi g(\phi^{(t-1)}, \bar{y}^{(\hat{k}^{(t)})})$

2.2. Stochastic variational inference

Given a probabilistic model

$$p(\phi, z, y) = p(\phi) \prod_{k=1}^{K} p(z^{(k)}|\phi) p(y^{(k)}|z^{(k)}, \phi)$$

that includes *global* latent variables ϕ, *local* latent variables $z = \{z^{(k)}\}_{k=1}^{K}$, and observations $y = \{y^{(k)}\}_{k=1}^{K}$, the mean field problem is to approximate the posterior $p(\phi, z|\bar{y})$ for fixed data $\bar{y}$ with a distribution of the form $q(\phi)q(z) = q(\phi)\prod_k q(z^{(k)})$ by finding a local minimum of the KL divergence from the approximating distribution to the posterior or, equivalently, finding a local maximum of the marginal likelihood lower bound

$$\mathcal{L} \triangleq \mathbb{E}_{q(\phi)q(z)} \left[\ln \frac{p(\phi, z, \bar{y})}{q(\phi)q(z)} \right] \leq p(\bar{y}). \tag{1}$$

SVI optimizes the objective (1) using a stochastic *natural* gradient ascent algorithm over the global factors $q(\phi)$.

Natural gradients of $\mathcal{L}$ with respect to the parameters of $q(\phi)$ have a convenient form if the prior $p(\phi)$ and each complete-data likelihood $p(z^{(k)}, y^{(k)}|\phi)$ are a conjugate pair of exponential family distributions. That is, if

$$\ln p(\phi) = \langle \eta_\phi, t_\phi(\phi) \rangle - A_\phi(\eta_\phi)$$
$$\ln p(z^{(k)}, y^{(k)}|\phi) = \langle \eta_{zy}(\phi), t_{zy}(z^{(k)}, y^{(k)}) \rangle - A_{zy}(\eta_{zy}(\phi))$$

then conjugacy (Bernardo & Smith, 2009, Proposition 5.4) implies that $t_\phi(\phi) = (\eta_{zy}(\phi), -A_{zy}(\eta_{zy}(\phi)))$, so that

$$p(\phi|z^{(k)}, \bar{y}^{(k)}) \propto \exp\{\langle \eta_\phi + (t_{zy}(z^{(k)}, \bar{y}^{(k)}), 1), t_\phi(\phi) \rangle\}.$$

Conjugacy also implies the optimal $q(\phi)$ is in the same family, i.e. $q(\phi) = \exp\{\langle \widetilde{\eta}_\phi, t_\phi(\phi) \rangle - A_\phi(\widetilde{\eta}_\phi)\}$ for some parameter $\widetilde{\eta}_\phi$ (Bishop, 2006, Section 10.4.1).

With this structure, there is a simple expression for the gradient of $\mathcal{L}$ with respect to $\widetilde{\eta}_\phi$. To simplify notation, we write $t(z, \bar{y}) \triangleq \sum_{k=1}^{K} (t_{zy}(z^{(k)}, \bar{y}^{(k)}), 1)$, $\widetilde{\eta} \triangleq \widetilde{\eta}_\phi$, $\eta \triangleq \eta_\phi$, and $A \triangleq A_\phi$. Then dropping terms constant over $\widetilde{\eta}$ we have

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\phi)q(z)} \left[\ln p(\phi|z, \bar{y}) - \ln q(\phi) \right] \\ &= \langle \eta + \mathbb{E}_{q(z)}[t(z, \bar{y})], \nabla A(\widetilde{\eta}) \rangle - (\langle \widetilde{\eta}, \nabla A(\widetilde{\eta}) \rangle - A(\widetilde{\eta}))\end{aligned}$$

where we have used the exponential family identity $\mathbb{E}_{q(\phi)}[t_\phi(\phi)] = \nabla A(\widetilde{\eta})$. Differentiating over $\widetilde{\eta}$, we have

$$\nabla_{\widetilde{\eta}} \mathcal{L} = \left(\nabla^2 A(\widetilde{\eta}) \right) \left(\eta + \mathbb{E}_{q(z)}[t(z, \bar{y})] - \widetilde{\eta} \right).$$

The natural gradient $\widetilde{\nabla}_{\widetilde{\eta}}$ is defined (Hoffman et al., 2013) as $\widetilde{\nabla}_{\widetilde{\eta}} \triangleq \left(\nabla^2 A(\widetilde{\eta}) \right)^{-1} \nabla_{\widetilde{\eta}}$, and so expanding we have

$$\widetilde{\nabla}_{\widetilde{\eta}} \mathcal{L} = \eta + \sum_{k=1}^{K} \mathbb{E}_{q(z^{(k)})}[(t_{zy}(z^{(k)}, \bar{y}^{(k)}), 1)] - \widetilde{\eta}.$$

Therefore a stochastic natural gradient ascent algorithm on the global variational parameter $\widetilde{\eta}_\phi$ proceeds at iteration t by sampling a minibatch $\bar{y}^{(k)}$ and taking a step of some size $\rho^{(t)}$ in an approximate natural gradient direction via

$$\widetilde{\eta}_\phi \leftarrow (1 - \rho^{(t)}) \widetilde{\eta}_\phi + \rho^{(t)} (\eta_\phi + s \cdot \mathbb{E}_{q^*(z^{(k)})}[t(z^{(k)}, \bar{y}^{(k)})])$$

where $s \triangleq |\bar{y}|/|\bar{y}^{(k)}|$ scales the minibatch statistics to represent the full dataset. In each step we find the optimal local factor $q^*(z^{(k)})$ with standard mean field updates and the current value of $q(\phi)$. There are automatic methods to tune the sequence of stepsizes (Snoek et al., 2012; Ranganath et al., 2013), though we do not explore them here.

2.3. Hidden Markov Models

A Bayesian Hidden Markov Model (HMM) on N states includes priors on the model parameters, namely the initial state distribution and transition matrix rows $\pi = \{\pi^{(i)}\}_{i=0}^{N}$

and the observation parameters $\theta = \{\theta^{(i)}\}_{i=1}^N$. The full generative model over the parameters, a state sequence $x_{1:T}$ of length T, and an observation sequence $y_{1:T}$ is

$$\theta^{(i)} \overset{\text{iid}}{\sim} p(\theta), \quad \pi^{(i)} \sim \text{Dir}(\alpha^{(i)}), \quad A \triangleq \begin{pmatrix} - & \pi^{(1)} & - \\ & \vdots & \\ - & \pi^{(N)} & - \end{pmatrix}$$
$$x_1 \sim \pi^{(0)}, \quad x_{t+1} \sim \pi^{(x_t)}, \qquad y_t \sim p(y_t|\theta^{(x_t)})$$

where we abuse notation slightly here and use $p(\theta)$ and $p(y_t|\theta)$ to denote the prior distribution over θ and the conditional observation distribution, respectively. When convenient, we collect the transition rows $\{\pi^{(i)}\}_{i=1}^N$ into the transition matrix A and write $q(A) \triangleq \prod_{i=1}^N q(\pi^{(i)})$.

Conditioned on the model parameters (π, θ) and a fixed observation sequence $\bar{y}_{1:T}$, the distribution of $x_{1:T}$ is Markov on a chain graph. Defining likelihood potentials L by $L_{t,i} \triangleq p(\bar{y}_t|\theta^{(i)})$, the density $p(x_{1:T}|\bar{y}_{1:T}, \pi, \theta)$ is

$$\exp\left\{\ln \pi^{(0)}_{x_1} + \sum_{t=1}^{T-1} \ln A_{x_t,x_{t+1}} + \sum_{t=1}^{T} \ln L_{t,x_t} - Z\right\}. \tag{2}$$

where Z is the normalizing constant for the distribution. We say $p(x_{1:T}|\bar{y}_{1:T}, \pi, \theta) = \text{HMM}(A, \pi^{(0)}, L)$.

In mean field inference for HMMs (Beal, 2003), we approximate the full posterior $p(\pi, \theta, x_{1:T}|\bar{y}_{1:T})$ with a mean field variational family $q(\pi)q(\theta)q(x_{1:T})$ and update each variational factor in turn while fixing the others. When updating $q(x_{1:T})$, by taking expectations of the log of (2) with respect to the variational distribution over parameters, we see the update sets $q(x_{1:T}) = \text{HMM}(\widetilde{A}, \widetilde{\pi}^{(0)}, \widetilde{L})$ with

$$\widetilde{A}_{i,j} \triangleq \exp\left\{\mathbb{E}_{q(\pi)} \ln A_{i,j}\right\} \quad \widetilde{\pi}^{(0)}_i \triangleq \exp\left\{\ln \mathbb{E}_{q(\pi^{(0)})} \pi^{(0)}_i\right\}$$

$$\widetilde{L}_{t,i} \triangleq \exp\left\{\mathbb{E}_{q(\theta_i)} \ln p(\bar{y}_t|x_t = i)\right\}. \tag{3}$$

We can compute the expectations with respect to $q(x_{1:T})$ necessary for the other factors' updates using the standard HMM message passing recursions for forward messages F and backward messages B using these HMM parameters:

$$F_{t,i} \triangleq \sum_{j=1}^{N} F_{t-1,j} \widetilde{A}_{ji} \widetilde{L}_{t,i} \qquad F_{1,i} \triangleq \widetilde{\pi}^{(0)}_i \tag{4}$$

$$B_{t,i} \triangleq \sum_{j=1}^{N} \widetilde{A}_{ij} \widetilde{L}_{t+1,j} B_{t+1,j} \qquad B_{T,i} \triangleq 1. \tag{5}$$

The messages can be computed in $\mathcal{O}(TN^2)$ time.

2.4. Hidden semi-Markov Models

The Hidden semi-Markov Model (HSMM) (Murphy, 2002; Hudson, 2008; Johnson & Willsky, 2013) augments the generative process of the HMM by sampling a duration from a state-specific duration distribution each time a state is entered. That is, if state i is entered at time t, a duration d is sampled $d \sim p(d|\vartheta^{(i)})$ for some parameter $\vartheta^{(i)}$ and the state stays fixed until x_{t+d-1}, when a Markov transition step selects a new state for x_{t+d}. For identifiability, self-transitions are often ruled out; the transition matrix A is constrained via $A_{i,i} = 0$ and the Dirichlet prior on each row is placed on the off-diagonal entries. The parameters $\pi^{(0)}$ and θ are treated as in the HMM.

Analogous to the HMM case, for a fixed observation sequence $\bar{y}_{1:T}$, we define likelihood potentials L by $L_{t,i} \triangleq p(\bar{y}_t|\theta^{(i)})$ and now define duration potentials D via $D_{d,i} \triangleq p(d|\vartheta^{(i)})$ and say $p(x_{1:T}|\bar{y}_{1:T}, \pi, \theta, \vartheta) = \text{HSMM}(A, \pi^{(0)}, L, D)$. In mean field inference for HSMMs, as developed in (Hudson, 2008), we approximate the posterior $p(\theta, \vartheta, \pi, x_{1:T}|\bar{y}_{1:T})$ with a variational family $q(\theta)q(\vartheta)q(\pi)q(x_{1:T})$. When updating the factor $q(x_{1:T})$ we have $q(x_{1:T}) = \text{HSMM}(\widetilde{A}, \widetilde{\pi}^{(0)}, \widetilde{L}, \widetilde{D})$ using the definitions in (3) and

$$\widetilde{D}_{d,i} \triangleq \exp\left\{\mathbb{E}_{q(\vartheta^{(i)})} \ln p(d|\vartheta^{(i)})\right\}.$$

Expectations with respect to $q(x_{1:T})$ can be computed in terms of the standard HSMM forward messages (F, F^*) and backward messages (B, B^*) via the recursions (Murphy, 2002):

$$F_{t,i} \triangleq \sum_{d=1}^{t-1} F^*_{t-d,i} \widetilde{D}_{d,i} \widetilde{L}_{t-d+1:t,i}, \quad F^*_{t,i} \triangleq \sum_{j=1}^{N} \widetilde{A}_{j,i} F_{t,j} \tag{6}$$

$$B^*_{t,i} \triangleq \sum_{d=1}^{T-t} B_{t+d,i} \widetilde{D}_{d,i} \widetilde{L}_{t+1:t+d,i} \quad B_{t,i} \triangleq \sum_{j=1}^{N} B^*_{t,j} \widetilde{A}_{i,j} \tag{7}$$

with $F^*_{1,i} \triangleq \pi^{(0)}_i$ and $B_{T,i} \triangleq 1$. These messages require $\mathcal{O}(T^2N + TN^2)$ time to compute for general duration distributions.

3. SVI for HMMs and HSMMs

In this section we apply SVI to both HMMs and HSMMs and express the SVI updates in terms of HMM and HSMM messages. For notational simplicity, we consider a dataset of K sequences each of length T, written $\bar{y} = \{\bar{y}^{(k)}_{1:T}\}_{k=1}^K$, and take each minibatch to be a single sequence, which we write without the superscript index as $\bar{y}_{1:T}$.

3.1. SVI update for HMMs

In terms of the notation in Section 2.2, the global variables are the HMM parameters and the local variables are the hidden states; that is, $\phi = (A, \pi^{(0)}, \theta)$ and $z = x_{1:T}$. To make the updates explicit, we assume the observation parameter priors $p(\theta^{(i)})$ and likelihoods $p(y_t|\theta^{(i)})$ are conjugate pairs of exponential family distributions for each i so that the

conditionals have the form

$$p(\theta^{(i)}|y) \propto \exp\left\{\langle \eta_\theta^{(i)} + (t_y^{(i)}(y), 1), t_\theta^{(i)}(\theta^{(i)})\rangle\right\}.$$

At each iteration of the SVI algorithm we sample a sequence $\bar{y}_{1:T}$ from the dataset and perform a stochastic gradient step on $q(A)q(\pi^{(0)})q(\theta)$ of some size ρ. To compute the gradient, we need to collect expected sufficient statistics with respect to the optimal factor for $q(x_{1:T})$, which in turn depends on the current value of $q(A)q(\pi^{(0)})q(\theta)$.

Writing the priors and mean field factors as

$$p(\pi^{(i)}) = \mathrm{Dir}(\alpha), \quad p(\theta^{(i)}) \propto \exp\left\{\langle \eta_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle\right\},$$
$$q(\pi^{(i)}) = \mathrm{Dir}(\widetilde{\alpha}^{(i)}), \quad q(\theta^{(i)}) \propto \exp\left\{\langle \widetilde{\eta}_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle\right\}$$

and using the messages F and B as in (4) and (5), we define

$$\begin{aligned}
\hat{t}_y^{(i)} &\triangleq \mathbb{E}_{q(x_{1:T})}\textstyle\sum_{t=1}^T \mathbb{I}[x_t = i] t_y^{(i)}(\bar{y}_t) \\
&= \textstyle\sum_{t=1}^T F_{t,i} B_{t,i} \cdot (t_y^{(i)}(\bar{y}_t), 1)/Z \qquad (8) \\
(\hat{t}_{\text{trans}}^{(i)})_j &\triangleq \mathbb{E}_{q(x_{1:T})}\textstyle\sum_{t=1}^{T-1} \mathbb{I}[x_t = i, x_{t+1} = j] \\
&= \textstyle\sum_{t=1}^{T-1} F_{t,i} \widetilde{A}_{i,j} \widetilde{L}_{t+1,j} B_{t+1,j}/Z \qquad (9) \\
(\hat{t}_{\text{init}})_i &\triangleq \mathbb{E}_{q(x_{1:T})} \mathbb{I}[x_1 = i] = \widetilde{\pi}_0 B_{1,i}/Z
\end{aligned}$$

where $\mathbb{I}[\cdot]$ is 1 if its argument is true and 0 otherwise and Z is the normalizing constant $Z \triangleq \sum_{i=1}^N F_{T,i}$.

With these expected statistics, taking a natural gradient step in the parameters of $q(A)$, $q(\pi_0)$, and $q(\theta)$ of size ρ is

$$\widetilde{\eta_\theta}^{(i)} \leftarrow (1-\rho)\widetilde{\eta_\theta}^{(i)} + \rho(\eta_\theta^{(i)} + s \cdot \hat{t}_y^{(i)}) \qquad (10)$$
$$\widetilde{\alpha}^{(i)} \leftarrow (1-\rho)\widetilde{\alpha}^{(i)} + \rho(\alpha^{(i)} + s \cdot \hat{t}_{\text{trans}}^{(i)}) \qquad (11)$$
$$\widetilde{\alpha}^{(0)} \leftarrow (1-\rho)\widetilde{\alpha}^{(0)} + \rho(\alpha^{(0)} + s \cdot \hat{t}_{\text{init}}^{(i)}) \qquad (12)$$

where $s \triangleq |\bar{y}|/|\bar{y}_{1:T}|$ as in Section 2.2.

3.2. SVI update for HSMMs

The SVI updates for the HSMM are very similar to those for the HMM with the addition of a duration update, but writing the expectations in terms of the HSMM messages is substantially different. The form of these expected statistics follow from the standard HSMM E-step (Murphy, 2002; Hudson, 2008).

Using the HSMM messages (F, F^*) and (B, B^*) defined in (6)-(7), we can write

$$\begin{aligned}
(\hat{t}_{\text{trans}}^{(i)})_j &\triangleq \mathbb{E}_{q(x_{1:T})}\textstyle\sum_{t=1}^{T-1} \mathbb{I}[x_t = i, x_{t+1} = j, x_t \neq x_{t+1}] \\
&= \textstyle\sum_{t=1}^{T-1} F_{t,i} B^*_{t,j} \widetilde{A}_{i,j}/Z
\end{aligned}$$

where Z is the normalizer $Z \triangleq \sum_{i=1}^N B^*_{1,i} \widetilde{\pi}_i^{(0)}$.

To be written in terms of the HSMM messages the expected state indicators $\mathbb{I}[x_t = i]$ must be expanded to

$$\mathbb{I}[x_t = i] = \sum_{\tau < t} \mathbb{I}[x_{\tau+1} = i \neq x_\tau] - \mathbb{I}[x_\tau = i \neq x_{\tau+1}]$$

Intuitively, this expansion expresses that a state is occupied after a transition into it occurs and until a transition out occurs while it is occupied. Then we have

$$\begin{aligned}
\mathbb{E}_{q(x_{1:T})}\mathbb{I}[x_{t+1} = i, x_t \neq x_{t+1}] &= F^*_{t,i} B^*_{t,i}/Z \\
\mathbb{E}_{q(x_{1:T})}\mathbb{I}[x_t = i, x_t \neq x_{t+1}] &= F_{t,i} B_{t,i}/Z.
\end{aligned}$$

from which we can compute $\mathbb{E}_{q(x_{1:T})}\mathbb{I}[x_t = i]$, which we use in the definition of $\hat{t}_y^{(i)}$ given in (8).

Finally, we compute the expected duration statistics as indicators on every possible duration $d = 1, 2, \ldots, T$ via

$$\begin{aligned}
(\hat{t}_{\text{dur}}^{(i)})_d &\triangleq \mathbb{E}\textstyle\sum_t \mathbb{I}[x_t \neq x_{t+1}, x_{t+1:t+d} = i, x_{t+d+1} \neq i] \\
&= \sum_{t=1}^{T-d+1} \widetilde{D}_{d,i} F^*_{t,i} B_{t+d,i} (\textstyle\prod_{t'=t}^{t+d} \widetilde{L}_{t',i})/Z. \qquad (13)
\end{aligned}$$

Note that this step alone requires $\mathcal{O}(T^2 N)$ time even after the messages have been computed.

If the priors and mean field factors over duration parameters are $p(\vartheta^{(i)}) \propto \exp\{\langle \eta_\vartheta^{(i)}, t_\vartheta^{(i)}(\vartheta^{(i)})\rangle\}$ and $q(\vartheta^{(i)}) \propto \exp\{\langle \widetilde{\eta}_\vartheta^{(i)}, t_\vartheta^{(i)}(\vartheta^{(i)})\rangle\}$, and the duration likelihood is $p(d|\vartheta^{(i)}) = \exp\{\langle t_\vartheta^{(i)}(\vartheta^{(i)}), (t_d(d), 1)\rangle\}$ then the duration factor update is

$$\widetilde{\eta}_\vartheta^{(i)} \leftarrow (1-\rho)\widetilde{\eta}_\vartheta^{(i)} + \rho(\eta_\vartheta^{(i)} + s(\sum_{d=1}^T (\hat{t}_{\text{dur}}^{(i)})_d \cdot (t_d(d), 1))).$$

4. SVI for HDP-HMMs and HDP-HSMMs

In this section we extend our methods to the Bayesian nonparametric versions of these models, the HDP-HMM and the HDP-HSMM. The generative model for the HDP-HMM with scalar concentration parameters $\alpha, \gamma > 0$ is

$$\begin{aligned}
&\beta \sim \mathrm{GEM}(\gamma), \quad \pi^{(i)} \sim \mathrm{DP}(\alpha\beta), \quad \theta^{(i)} \overset{\text{iid}}{\sim} p(\theta^{(i)}) \\
&x_1 \sim \pi^{(0)}, \qquad x_{t+1} \sim \pi^{(x_t)}, \qquad y_t \sim p(y_t|\theta^{(x_t)})
\end{aligned}$$

where $\beta \sim \mathrm{GEM}(\gamma)$ denotes sampling from a stick breaking distribution defined by

$$v_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \gamma), \qquad \beta_k = \prod_{j<k}(1 - v_j)v_k$$

and $\pi^{(i)} \sim \mathrm{DP}(\alpha\beta)$ denotes sampling a Dirichlet process

$$w \sim \mathrm{GEM}(\alpha), \qquad z_k \overset{\text{iid}}{\sim} \beta, \qquad \pi^{(i)} = \sum_{k=1}^\infty w_k \delta_{z_k}.$$

To perform mean field inference in HDP models, we approximate the posterior with a truncated variational distribution. While a common truncation is to limit the two stick-breaking distributions in the definition of the HDP (Hoffman et al., 2013), a more convenient truncation for our models is the "direct assignment" truncation, used in (Liang et al., 2007) for batch mean field with the HDP-HMM and in (Bryant & Sudderth, 2012) in an SVI algorithm for LDA. The direct assignment truncation limits the support of $q(x_{1:T})$ to the finite set $\{1, 2, \ldots, K\}^T$ for a truncation parameter K, i.e. fixing $q(x_{1:T}) = 0$ when any $x_t > K$. Thus the other factors, namely $q(\pi)$, $q(\beta)$, and $q(\theta)$, only differ from their priors in their distribution over the first K components. As opposed to standard truncation, this family of approximations is nested over K, enabling a search procedure over the truncation parameter as developed in (Bryant & Sudderth, 2012). A similar search procedure can be used with the HDP-HMM and HDP-HSMM algorithms in this paper, though we do not explore it here.

A disadvantage to the direct assignment truncation is that the update to $q(\beta)$ is not conjugate given the other factors as in Hoffman et al. (2013). Following Liang et al. (2007), to simplify the update we use a point estimate by writing $q(\beta) = \delta_{\beta^*}(\beta)$. Since the main effect of β is to enforce shared sparsity among the $\pi^{(i)}$, it is reasonable to expect that a point approximation for $q(\beta)$ will suffice.

The updates to the factors $q(\theta)$ and $q(x_{1:T})$ are identical to those derived in the previous sections. To derive the SVI update for $q(\pi)$, we write the relevant part of the untruncated model and truncated variational factors as $p((\pi^{(i)}_{1:K}, \pi^{(i)}_{\text{rest}})) = \text{Dir}(\alpha \cdot (\beta_{1:K}, \beta_{\text{rest}}))$ and $q((\pi^{(i)}_{1:K}, \pi^{(i)}_{\text{rest}})) = \text{Dir}(\widetilde{\alpha}^{(i)})$, respectively, where $\pi^{(i)}_{\text{rest}} \triangleq 1 - \sum_{k=1}^K \pi^{(i)}_k$ and $\beta_{\text{rest}} \triangleq 1 - \sum_{k=1}^K \beta_k$ for $i = 1, \ldots, K$. Therefore the updates to $q(\pi^{(i)})$ are identical to those in (11) except the number of variational parameters is $K + 1$ and the prior hyperparameters are replaced with $\alpha \cdot (\beta_{1:K}, \beta_{\text{rest}})$.

The gradient of the variational objective with respect to β^* is given by

$$\begin{aligned}\nabla_{\beta^*} \mathcal{L} &= \nabla_{\beta^*} \left\{ \mathbb{E}_{q(\pi)} \left[\ln \frac{p(\beta, \pi)}{q(\beta) q(\pi)} \right] \right\} \\ &= \nabla_{\beta^*} \left\{ \ln p(\beta^*) + \sum_{i=1}^K \mathbb{E}_{q(\pi^{(i)})} \ln p(\pi^{(i)} | \beta^*) \right\}\end{aligned}$$

$$\frac{\partial}{\partial \beta^*_k} \ln p(\beta^*) = 2 \sum_{i \geq k} \frac{1}{1 - \sum_{j<i} \beta^*_j} - (\gamma - 1) \sum_{i \geq k} \frac{1}{1 - \sum_{j \leq i} \beta^*_j}$$

$$\begin{aligned}&\frac{\partial}{\partial \beta^*_k} \mathbb{E}_{q(\pi)} [\ln p(\pi^{(i)} | \beta^*)] \\ &\quad = \gamma \psi(\widetilde{\alpha}^{(i)}_k) - \gamma \psi(\widetilde{\alpha}^{(i)}_{K+1}) + \gamma \psi(\gamma \sum_{j=1}^{K+1} \beta^*_j) - \gamma \psi(\beta^*_k).\end{aligned}$$

We use this gradient expression to take a truncated gradient step on β^* during each SVI update, using a backtracking line search to ensure the updated value satisfies $\beta^* \geq 0$.

The updates for $q(\pi)$ and $q(\beta)$ in the HDP-HSMM differ only in that the variational lower bound expression changes slightly because the support of each $q(\pi^{(i)})$ is restricted to the off-diagonal (and renormalized). We can adapt $q(\pi^{(i)})$ by simply dropping the ith component from the representation and writing

$$q((\pi^{(i)}_{1:K \setminus i}, \pi^{(i)}_{\text{rest}})) = \text{Dir}(\widetilde{\alpha}^{(i)}_{\setminus i}),$$

and we change the second term in the gradient for β^* to

$$\begin{aligned}&\frac{\partial}{\partial \beta^*_k} \mathbb{E}_{q(\pi)} [\ln p(\pi^{(i)} | \beta^*)] \\ &\quad = \gamma \psi(\widetilde{\alpha}^{(i)}_k) - \gamma \psi(\widetilde{\alpha}^{(i)}_{K+1}) + \gamma \psi(\gamma \sum_{j \neq i} \beta^*_j) - \gamma \psi(\beta^*_k)\end{aligned}$$

when $k \neq i$, otherwise the partial derivative is 0.

Using these gradient expressions for β^* and a suitable gradient-based batch optimization procedure we can also perform batch mean field updates for the HDP-HSMM.

5. Fast updates for negative binomial HSMMs

General HSMM inference is much more expensive than HMM inference, having runtime $\mathcal{O}(T^2 N + T N^2)$ compared to just $\mathcal{O}(T N^2)$ on N states and a sequence of length T. The quadratic dependence on T can be severely limiting even in the minibatch setting of SVI, since minibatches often must be sufficiently large for good performance (Hoffman et al., 2013; Broderick et al., 2013).

A common approach to mitigate HSMM computational complexity (Hudson, 2008; Johnson & Willsky, 2013) is to limit the support of the duration distributions, either in the model or as an approximation in the message passing computation, and thus limit the terms in the sums of (6) and (7). This truncation approach can be readily applied to the algorithms presented in this paper. However, truncation can be undesirable or ineffective if states have long durations. In the this section, we develop approximate updates for a particular class of duration distributions with unbounded support for which the computational complexity is only linear in T.

5.1. HMM embeddings of negative binomial HSMMs

The negative binomial distribution we use has two parameters (r, p), where $0 < p < 1$ and r is a positive integer. Its probability mass function (PMF) for $k = 1, 2, \ldots$ is

$$p(k | r, p) = \binom{k + r - 2}{k - 1} \exp\{(k - 1) \ln p + r \ln(1 - p)\}. \tag{14}$$

Fixing r, the family of distributions parameterized by p is an exponential family and admits a conjugate Beta prior. However, as a family over (r, p) it is not exponential because the of the binomial coefficient base measure term which depends on r. When $r = 1$ the distribution is geometric, and so the class of HSMMs with negative binomial durations include HMMs. By varying r and p, the mean and variance can be controlled separately, making the negative binomial a popular choice for duration distributions (Bulla & Bulla, 2006; Fearnhead, 2006).

A negative binomial random variable can be represented as a sum of r geometric random variables: if $x \sim \text{NB}(r, p)$ and $y = 1 + \sum_{i=1}^{r} z_i$ with $p(z_i = k) = p^k(1-p)$, then $x \sim y$. Therefore given an HSMM in which the durations of state i are distributed as $\text{NB}(r^{(i)}, p^{(i)})$ we can construct an HMM on $\sum_{i=1}^{N} r^{(i)}$ states that encodes the same process, where HSMM state i corresponds to r_i states in the HMM. We call this construction an *HMM embedding* of the HSMM, and the resulting HMM transition matrix $\bar{A}$ is

$$\bar{A} \triangleq \begin{pmatrix} C_1 & \bar{A}_{12} & \cdots \\ \bar{A}_{21} & C_2 & \\ \vdots & & \end{pmatrix}$$

$$C_i \triangleq \begin{pmatrix} p^{(i)} & 1-p^{(i)} & \\ & \ddots & \\ & p^{(i)} & 1-p^{(i)} \end{pmatrix}, \; \bar{A}_{ij} \triangleq \begin{pmatrix} & & \\ & & \\ A_{ij}\bar{p}_1^{(ij)} & \cdots & A_{ij}\bar{p}_{r^{(j)}}^{(ij)} \end{pmatrix}$$

where $\bar{p}^{(ij)}$ is defined in the supplementary materials. We write the HMM embedding state sequence as $\bar{x}_{1:T}$, where each $\bar{x}_t$ decomposes as $\bar{x}_t = (x_t, k)$ for $k = 1, 2, \ldots, r^{(x_t)}$ according to the block structure of $\bar{A}$. If we define $R \triangleq \frac{1}{N}\sum_{i=1}^{N} r^{(i)}$, then passing messages in this structured HMM embedding can be done in $\mathcal{O}(TNR + TN^2)$ time.

This construction draws on ideas from HSMMs with "parametric macro-states" (Guédon, 2005, Section 3) and on expanded-state HMMs (ESHMMs) (Russell & Moore, 1985; Russell & Cook, 1987; Johnson, 2005). However, this precise construction for negative binomial durations does not appear in those works. Furthermore, we extend these ideas by applying Bayesian inference as well as methods to fit (a posterior over) the $r^{(i)}$ parameter, as we discuss in the next section.

If every $r^{(i)}$ is fixed and the $p^{(i)}$ are the only duration parameters, we can use the HMM embedding to perform efficient conjugate SVI (or batch mean field) updates to the duration factors $q(p^{(i)})$. We write the duration prior and mean field factors as

$$p(p^{(i)}) = \text{Beta}(a^{(i)}, b^{(i)}) \qquad q(p^{(i)}) = \text{Beta}(\widetilde{a}^{(i)}, \widetilde{b}^{(i)}).$$

The embedding allows us to write the variational lower bound for the HSMM as an equivalent HMM variational lower bound with effective transition matrix $\mathbb{E}_{q(\pi)q(p)q(\theta)} \ln \bar{A}$. Defining $q(\bar{x}_{1:T})$ as the corresponding distribution over HMM states, we can write expected sufficient statistics $\hat{t}_d^{(i)} \triangleq (\hat{t}_1^{(i)}, \hat{t}_0^{(i)})$ for the duration factors in terms of the expected transition statistics in the HMM embedding:

$$\hat{t}_{d,1}^{(i)} \triangleq \mathbb{E}_{q(\bar{x}_{1:T})} \sum_{t=1}^{T-1} \sum_{k=1}^{r^{(i)}} \mathbb{I}[\bar{x}_t = \bar{x}_{t+1} = (i,k)]$$

$$\hat{t}_{d,0}^{(i)} \triangleq \mathbb{E}_{q(\bar{x}_{1:T})} \sum_{t=1}^{T-1} \sum_{k=1}^{r^{(i)}} \mathbb{I}[\bar{x}_t = (i,k), \bar{x}_{t+1} \neq \bar{x}_{t+1}].$$

We can compute these expected transition statistics efficiently from the HMM messages using (9). The SVI update to the duration factors is then of the form

$$\widetilde{a}^{(i)} \leftarrow (1-\rho)\widetilde{a}^{(i)} + \rho(a^{(i)} + s \cdot \hat{t}_{d,1}^{(i)}) \tag{15}$$

$$\widetilde{b}^{(i)} \leftarrow (1-\rho)\widetilde{b}^{(i)} + \rho(b^{(i)} + s \cdot \hat{t}_{d,0}^{(i)}) \tag{16}$$

for some stepsize ρ and minibatch scaling s. We can similarly write the transition, initial state, and observation statistics for the other HSMM mean field factors in terms of its embedding:

$$\hat{t}_y^{(i)} \triangleq \mathbb{E}_{q(\bar{x}_{1:T})} \textstyle\sum_{t=1}^{T} \sum_{k=1}^{r^{(i)}} \mathbb{I}[\bar{x}_t = (i,k)] t_y^{(i)}(\bar{y}_t)$$

$$(\hat{t}_{\text{trans}}^{(i)})_j \triangleq \mathbb{E}_{q(\bar{x}_{1:T})} \textstyle\sum_{t=1}^{T-1} \mathbb{I}[\bar{x}_t = (i, r^{(i)}) \neq \bar{x}_{t+1}]$$

$$(\hat{t}_{\text{init}})_i \triangleq \mathbb{E}_{q(\bar{x}_{1:T})} \mathbb{I}[\bar{x}_1 = (x_t, 1)].$$

Using these statistics we perform SVI updates to the corresponding the mean field factors as in (10)-(12).

We have shown that by working with an efficient HMM embedding representation we can compute updates to the HSMM mean field factors in time $\mathcal{O}(TNR + TN^2)$ when durations are negative binomially distributed with fixed r. In the next subsection we extend these fast updates to include a variational representation to the posterior of r.

5.2. Approximate updates for fitting $q(r, p)$

By learning r as well as p, negative binomial HSMMs can learn to be HMMs when appropriate and generally provide a much more flexible class of duration distributions. In this subsection, we derive an exact SVI update step for fitting both r and p, explain its computational difficulties, and propose a fast approximate alternative based on sampling.

We define mean field factors $q(r^{(i)})q(p^{(i)}|r^{(i)})$, where each $q(r^{(i)})$ is a categorical distribution with finite support taken to be $\{1, 2, \ldots, r_{\text{max}}^{(i)}\}$ and each $q(p^{(i)}|r^{(i)})$ is Beta, i.e.

$$q(r^{(i)}) \propto \exp\{\langle \widetilde{\nu}^{(i)}, \mathbb{I}_{r^{(i)}} \rangle\}$$

$$q(p^{(i)}|r^{(i)}) = \text{Beta}(\widetilde{a}^{(i,r)}, \widetilde{b}^{(i,r)}).$$

where $\mathbb{I}_{r^{(i)}}$ is an indicator vector with the $r^{(i)}$th entry set to 1 and the others to zero. The priors are defined similarly. To simplify notation, in this section we often drop the superscript i from the notation.

Two challenges arise in computing updates to $q(r,p)$. First, the optimal variational factor on the HSMM states is

$$q(x_{1:T}) \propto \exp \mathbb{E}_{q(\pi)q(r,p)q(\theta)} \ln p(x_{1:T}|\bar{y}_{1:T}, \pi, r, p, \theta).$$

Due to the expectation over $q(r)$, this factor does not have the form required to use the efficient HMM embedding of Section 5.1, and so the corresponding general HSMM messages require $\mathcal{O}(T^2N + TN^2)$ time to compute. Second, as we show next, due to the base measure term in (14), to compute an exact update to $q(r,p)$ requires computing the expected duration indicator statistics of (13), a computation which itself requires $\mathcal{O}(T^2N)$ time even after computing the HSMM messages.

First, we show that the update to $q(p|r)$ is straightforward. To derive an update for $q(p|r)$, we write the relevant part of the variational lower bound as

$$\begin{aligned}\mathcal{L} &\triangleq \mathbb{E}_{q(r,p)q(x_{1:T})}\left[\ln \frac{p(r,p,D)}{q(r,p)}\right] \qquad (17)\\ &= \mathbb{E}_{q(r)} \ln \frac{p(r)}{q(r)} + \mathbb{E}_{q(r)q(x_{1:T})} h(r,D) \\ &\qquad + \mathbb{E}_{q(r)}\left\{\mathbb{E}_{q(p|r)} \ln \frac{p(p)\bar{p}(D|r,p)}{q(p|r)}\right\}\end{aligned}$$

where D is the set of relevant durations in $x_{1:T}$, $h(r,D) \triangleq \sum_{k\in D} \ln \binom{r+k-2}{k-1}$ arises from the negative binomial base measure term, and $\ln \bar{p}(D|r,p) \triangleq \sum_{k\in D} k \ln p + r \ln(1-p)$ collects the negative binomial PMF terms excluding the base measure. The only terms in (17) that depend on $q(p|r)$ are in the final bracketed term. Furthermore, each of these terms corresponds to the variational lower bound for the fixed-r case described in Section 5.1, and so each $q(p|r)$ factor can be updated with Eqs. (15)-(16).

To compute an update for $q(r)$, we note that since it is a distribution with finite support we can write its complete-data conditional in exponential family form trivially via

$$\begin{aligned}p(r|p,D) &\propto \exp\{\langle \nu + t_r(p,D), \mathbb{I}_r\rangle\}\\ t_r(p,D)_j &\triangleq \textstyle\sum_{k\in D} \ln p(p|k, r=j) + \ln h(j,k)\end{aligned}$$

and so from the results in Section 2.2 the natural gradient of (17) with respect to the parameters of $q(r)$ is

$$\widetilde{\nabla}_{\widetilde{\nu}} \mathcal{L} = \nu + \mathbb{E}_{q(p|r)q(x_{1:T})} t_r(p,D) - \widetilde{\nu}. \qquad (18)$$

Due to the base measure term $h(j,k)$, to compute this update requires evaluating the expected statistics of Eq. (13), which require $\mathcal{O}(T^2N)$ time.

Therefore computing an exact SVI update on the $q(r,p)$ factors is expensive both because the HSMM messages cannot be computed using the methods of Section 5.1 and because given the messages the required statistics are expensive to compute. To achieve an update running time that is linear in T, we propose to use a sample approximation to $q(x_{1:T})$ inspired by the method developed in (Wang & Blei, 2012). That is, we sample negative binomial HSMM models from the distribution $q(\pi)q(\theta)q(r,p)$ and use the embedding to generate a sample of $x_{1:T}$ under each model. Using the message passing methods of Section 5.1, the first state sequence sample for each model can be collected in time $\mathcal{O}(TNR + TN^2)$, and additional state sequence samples from the same model can be collected in time $\mathcal{O}(TN)$.

As discussed in Wang & Blei (2012), this sampling approximation does not optimize the variational lower bound over $q(x_{1:T})$ and so it should yield an inferior objective value. Indeed, while the optimal mean field update sets $q(x_{1:T}) \propto \exp\{\mathbb{E}_q[\ln p(\pi, \theta, \vartheta, x_{1:T}, \bar{y}_{1:T})]\}$, this update approximates $q(x_{1:T}) \propto \mathbb{E}_q[p(\pi, \theta, \vartheta, x_{1:T}, \bar{y}_{1:T})]$. However, Wang & Blei (2012) found this approximate update yielded better predictive performance in some topic models, and provided an interpretation as an approximate expectation propagation (EP) update.

With the collected samples $\mathcal{S} \triangleq \{x_{1:T}^{(k)}\}_{k=1}^{S}$, we set the factor $\hat{q}(x_{1:T}) = \frac{1}{S}\sum_{\hat{x}_{1:T}\in\mathcal{S}} \delta_{\hat{x}_{1:T}}(x_{1:T})$, where we use the notation $\hat{q}(x_{1:T})$ to emphasize that it is a sample-based representation. It is straightforward to compute the expectation over states in (18) by plugging in the sampled durations. The update to the parameters of $q(r^{(i)}, p^{(i)})$ is

$$\begin{aligned}\widetilde{\nu}^{(i)} &\leftarrow (1-\rho)\widetilde{\nu}^{(i)} + \rho(\nu^{(i)} + s\cdot \hat{t}_r^{(i)})\\ \widetilde{a}^{(i,r)} &\leftarrow (1-\rho)\widetilde{a}^{(i,r)} + \rho(a^{(i)} + s\cdot \hat{t}_a^{(i,r)})\\ \widetilde{b}^{(i,r)} &\leftarrow (1-\rho)\widetilde{b}^{(i,r)} + \rho(b^{(i)} + s\cdot \hat{t}_b^{(i,r)})\end{aligned}$$

$$\begin{aligned}\hat{t}_a^{(i,r)} &\triangleq \frac{1}{S}\sum_{\hat{x}\in\mathcal{S}}\sum_{d\in D^{(i)}(\hat{x})} (d-1)\\ \hat{t}_b^{(i,r)} &\triangleq \frac{1}{S}\sum_{\hat{x}\in\mathcal{S}}\sum_{d\in D^{(i)}(\hat{x})} r\\ (\hat{t}_r^{(i)})_r &\triangleq (\widetilde{a}^{(i,r)} + \hat{t}_a^{(i,r)} - 1)\mathbb{E}_{q(p|r)}[\ln(p^{(i,r)})]\\ &\quad + (\widetilde{b}^{(i,r)} + \hat{t}_b^{(i,r)} - 1)\mathbb{E}_{q(p|r)}[\ln(1-p^{(i,r)})]\\ &\quad + \sum_{\hat{x}\in\mathcal{S}}\sum_{d\in D^{(i)}(\hat{x})} \ln\binom{d+r-2}{d-1}\end{aligned}$$

and where $D^{(i)}(\hat{x}_{1:T})$ denotes the set of durations of state i in the sequence $\hat{x}_{1:T}$. The updates to the other factors are as before but with expectations taken over $\hat{q}(x_{1:T})$.

The methods presented in this section for HSMMs with negative binomial durations can be extended in several

ways. In particular, one can use similar methods to perform efficient updates when state durations are modeled as mixtures of negative binomial distributions. Since each negative binomial can separately parameterize mean and variance, in this way one can generate a flexible and convenient family of duration distributions analogous to the Gaussian mixture model pervasive in density modeling.

6. Experiments

We conclude with a numerical study to validate these algorithms and in particular measure the effectiveness of the approximate updates proposed in Section 5.2. As a performance metric, we evaluate an approximate posterior predictive density on held-out data, writing

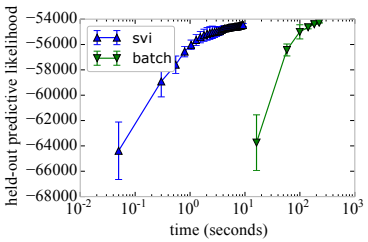
$$p(\bar{y}_{\text{test}}|\bar{y}_{\text{train}}) = \int\int p(\bar{y}_{\text{test}}|\pi\theta)p(\pi,\theta|\bar{y}_{\text{train}})\mathrm{d}\pi\mathrm{d}\theta$$
$$\approx \mathbb{E}_{q(\pi)q(\theta)}p(\bar{y}_{\text{test}}|\pi,\theta)$$

and approximating the expectation by sampling models from the variational distribution. To reproduce these figures, see the code in the supplementary materials.

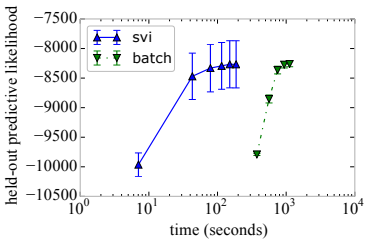
First, we compare the performance of SVI and batch mean field algorithms for the HDP-HMM. We sampled a 10-state HMM with 2-dimensional Gaussian emissions and generated a dataset of 100 observation sequences of length 3000 each. We chose a random subset of 95% of the sequences as training sequences and held out 5% as test sequences. We repeated the fitting procedures 5 times with identical initializations drawn from the prior, and we report the median performance with standard deviation error bars. The SVI procedure made only one pass through the training set. Figure 1(a) shows that the SVI algorithm produces fits that are comparable in performance in the time it takes the batch algorithm to complete a single iteration.

Similarly, we compare the SVI and batch mean field algorithms for the HDP-HSMM with Poisson durations. Due to the much greater computational complexity of HSMM inference, we generated a set of 30 sequences of length 2000 each and used 90% of the sequences in the training set. Figure 1(b) again demonstrates that the SVI algorithm can fit such models in the time it takes the batch algorithm to complete a single iteration.

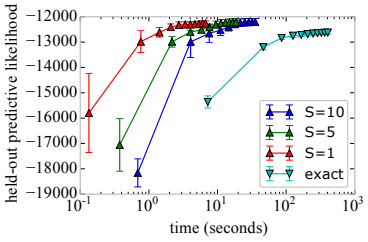
Finally, we compare the performance of the exact SVI update for the HSMM with that of the fast approximate update proposed in Section 5. We again generated data using Poisson duration distributions, but we train models using negative binomial durations where $p \sim \text{Beta}(1,1)$ and $r \sim \text{Uniform}(\{1,2,\ldots,10\})$. We generated 55 observation sequences of length 3000 and used 90% of the sequences in the training set. We compare the sampling algorithm's performance for several numbers of samples S.



(a) SVI vs Batch HDP-HMM



(b) SVI vs Batch HDP-HSMM



(c) Exact vs Approx. HSMM SVI

Figure 1. Synthetic numerical experiments.

Figure 1(c) shows that the approximate update from Section 5 results in higher predictive performance than that of the model trained with the exact update even using a single sample. This performance is likely dataset-dependent, but the experiment demonstrates that the approximate update may be very effective in some cases.

7. Conclusion

This paper develops scalable SVI-based inference for HMMs, HSMMs, HDP-HMMs, and HDP-HSMMs, and provides a technique to make Bayesian inference in negative binomial HSMMs much more practical. These models are widely applicable to time series inference, so these algorithms and our code may be immediately useful to the community.

Acknowledgements

This research was supported in part under AFOSR Grant FA9550-12-1-0287.

References

Beal, Matthew James. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

Bernardo, José M and Smith, Adrian FM. *Bayesian theory*, volume 405. Wiley. com, 2009.

Bishop, Christopher M. *Pattern recognition and machine learning*, volume 1. Springer, 2006.

Bottou, Léon. Online learning and stochastic approximations. *On-line learning in neural networks*, 17:9, 1998.

Broderick, Tamara, Boyd, Nicholas, Wibisono, Andre, Wilson, Ashia C, and Jordan, Michael. Streaming variational bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735, 2013.

Bryant, Michael and Sudderth, Erik B. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, pp. 2708–2716, 2012.

Bulla, Jan and Bulla, Ingo. Stylized facts of financial time series and hidden semi-markov models. *Computational Statistics & Data Analysis*, 51(4):2192–2209, 2006.

Fearnhead, Paul. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2): 203–213, 2006.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. Sharing features among dynamical systems with beta processes. In *Neural Information Processing Systems 22*. MIT Press, 2010.

Fox, Emily, Sudderth, Erik B, Jordan, Michael I, and Willsky, Alan S. Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585, 2011.

Griffiths, Thomas L, Steyvers, Mark, Blei, David M, and Tenenbaum, Joshua B. Integrating topics and syntax. In *Advances in neural information processing systems*, pp. 537–544, 2004.

Guédon, Yann. Hidden hybrid markov/semi-markov chains. *Computational statistics & Data analysis*, 49(3):663–688, 2005.

Hoffman, M, Blei, D, Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Hoffman, Matthew, Bach, Francis R, and Blei, David M. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

Hudson, Nicolas H. *Inference in hybrid systems with applications in neural prosthetics*. PhD thesis, California Institute of Technology, 2008.

Johnson, Matthew J. and Willsky, Alan S. Bayesian nonparametric hidden semi-markov models. *Journal of Machine Learning Research*, 14:673–701, February 2013.

Johnson, Michael T. Capacity and complexity of hmm duration modeling techniques. *Signal Processing Letters, IEEE*, 12(5): 407–410, 2005.

Lehman, Li-wei H, Nemati, Shamim, Adams, Ryan P, and Mark, Roger G. Discovering shared dynamics in physiological signals: Application to patient monitoring in icu. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 5939–5942. IEEE, 2012.

Liang, Percy, Petrov, Slav, Jordan, Michael I, and Klein, Dan. The infinite pcfg using hierarchical dirichlet processes. In *EMNLP-CoNLL*, pp. 688–697, 2007.

Lindén, Martin, Johnson, Stephanie, van de Meent, Jan-Willem, Phillips, Rob, and Wiggins, Chris H. Analysis of dna looping kinetics in tethered particle motion experiments using hidden markov models. *Biophysical Journal*, 104(2):418A–418A, 2013.

Murphy, K. Hidden semi-markov models (segment models). *Technical Report*, November 2002. URL http://www.cs.ubc.ca/~murphyk/Papers/segment.pdf.

Ranganath, Rajesh, Wang, Chong, David, Blei, and Xing, Eric. An adaptive learning rate for stochastic variational inference. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 298–306, 2013.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Russell, Martin and Moore, Roger. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pp. 5–8. IEEE, 1985.

Russell, Martin J and Cook, A. Experimental evaluation of duration modelling techniques for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, volume 12, pp. 2376–2379. IEEE, 1987.

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*, 2012.

Wang, Chong and Blei, David. Truncation-free online variational inference for bayesian nonparametric models. In *Advances in Neural Information Processing Systems 25*, pp. 422–430, 2012.

Wang, Chong, Paisley, John W, and Blei, David M. Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.