Chapter 5

Stochastic Variational Inference for HMMs, HSMMs, and Nonparametric Extensions

Hierarchical Bayesian time series models can be applied to complex data in many domains, including data arising from behavior and motion [32, 33], home energy consumption [60], physiological signals [69], single-molecule biophysics [71], brain-machine interfaces [54], and natural language and text [44, 70]. However, for many of these applications there are very large and growing datasets, and scaling Bayesian inference in rich hierarchical models to these large datasets is a fundamental challenge.

Many Bayesian inference algorithms, including standard Gibbs sampling and mean field 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 [52] and can even operate in the single-pass or streaming settings [15]. In particular, stochastic variational inference (SVI) [52] provides a general framework for scalable inference based on mean field and stochastic gradient descent. However, while SVI has been studied extensively for topic models [53, 115, 17, 114, 92, 52], it has not been applied to time series.

In this chapter, we develop SVI algorithms for the core Bayesian time series models of this thesis, namely the hidden Markov model (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 here are widely applicable. However, as discussed in the previous chapter, general HSMM inference subroutines have time complexity that scales quadratically with observation sequence length, and such quadratic scaling can be impractical even in the setting of SVI. To address this shortcoming, we use the methods developed in Chapter 4 for Bayesian inference in (HDP-)HSMMs with negative binomial durations to provide approximate

Algorithm 5.1 Stochastic gradient ascent

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

SVI updates with time complexity that scales only linearly with sequence length.

In Section 5.1 we briefly review the basic ingredients of SVI. In Section 5.2, we derive SVI updates for (finite) HMMs and HSMMs, and in Section 5.3 we apply the methods derived in Chapter 4 to derive faster SVI updates for HSMMs with negative binomial durations. Finally, in Section 5.4 we extend these algorithms to the nonparametric HDP-HMM and HDP-HSMM.

5.1 Stochastic variational inference

In this section we summarize the general stochastic variational inference (SVI) framework developed in Hoffman et al. [52]. SVI involves performing stochastic gradient optimization on a mean field variational objective, so we first review basic results on stochastic gradient optimization and next provide a derivation of the form of the natural gradient of mean field objectives for complete-data conjugate models. We use the notation defined in Sections 2.3.2 and 2.4.2 throughout.

5.1.1 Stochastic gradient optimization

Consider the optimization problem

$$\arg\max_\phi f(\phi, \bar{y}) \qquad \text{where} \qquad f(\phi, \bar{y}) = \sum_{k=1}^{K} g(\phi, \bar{y}^{(k)}) \tag{5.1.1}$$

and where $\bar{y} = \{\bar{y}^{(k)}\}_{k=1}^K$ is a fixed dataset. Using the decomposition of the objective function f, if $\hat{k}$ is sampled uniformly over $\{1, 2, \ldots, K\}$, we have

$$\nabla_\phi f(\phi) = K \sum_{k=1}^{K} \frac{1}{K} \nabla_\phi g(\phi, \bar{y}^{(k)}) = K \cdot \mathbb{E}_{\hat{k}} \left[\nabla_\phi g(\phi, \bar{y}^{(\hat{k})}) \right]. \tag{5.1.2}$$

Thus we can generate approximate gradients of the objective f using only one $\bar{y}^{(k)}$ at a time. A stochastic gradient ascent algorithm for a sequence of *stepsizes* $\rho^{(t)}$ and a sequence of positive definite matrices $G^{(t)}$ is given in Algorithm 5.1.

From classical results in stochastic optimization [93, 14], if the sequence of stepsizes

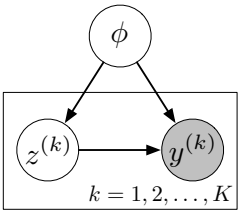

Figure 5.1: Prototypical graphical model for stochastic variational inference (SVI). The global latent variables are represented by ϕ and the local latent variables by $z^{(k)}$.

satisfies $\sum_{t=1}^{\infty} \rho^{(t)} = \infty$ and $\sum_{t=1}^{\infty} (\rho^{(t)})^2 < \infty$ and each $G^{(t)}$ has uniformly bounded eigenvalues, then the algorithm converges to a local optimum, i.e. $\phi^* \triangleq \lim_{t\to\infty} \phi^{(t)}$ satisfies $\nabla_\phi f(\phi^*, \bar{y}) = 0$ with probability 1. If $\bar{y}$ is a large dataset, then each update in a stochastic gradient algorithm only operates on one $\bar{y}^{(k)}$, or *minibatch*, at a time; therefore, stochastic gradient algorithms can scale to the large-data setting. To make a single-pass algorithm, the minibatches can be sampled without replacement. The choice of stepsize sequence can significantly affect the performance of a stochastic gradient optimization algorithm. There are automatic methods to tune or adapt the sequence of stepsizes [104, 92], though we do not discuss them here.

SVI uses a particular stochastic gradient ascent algorithm to optimize a mean field variational Bayesian objective over large datasets $\bar{y}$, as we review next.

■ 5.1.2 Stochastic variational inference

Using the notation of Section 2.3.2, given a probabilistic model of the form

$$p(\phi, z, y) = p(\phi) \prod_{k=1}^{K} p(z^{(k)}|\phi) p(y^{(k)}|z^{(k)}, \phi) \tag{5.1.3}$$

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

$$\mathcal{L} \triangleq \mathbb{E}_{q(\phi)q(z)} \left[\ln \frac{p(\phi, z, \bar{y})}{q(\phi)q(z)} \right] \leq p(\bar{y}). \tag{5.1.4}$$

SVI optimizes the objective (5.1.4) using a stochastic *natural* gradient ascent algorithm over the global factors $q(\phi)$. See Figure 5.1 for a graphical model.

Gradients of $\mathcal{L}$ with respect to the parameters of $q(\phi)$ have a convenient form if we

assume the prior $p(\phi)$ and each complete-data likelihood $p(z^{(k)}, y^{(k)}|\phi)$ are a conjugate pair of exponential family densities. That is, if we have

$$\ln p(\phi) = \langle \eta_\phi, t_\phi(\phi) \rangle - Z_\phi(\eta_\phi) \tag{5.1.5}$$

$$\ln p(z^{(k)}, y^{(k)}|\phi) = \langle \eta_{zy}(\phi), t_{zy}(z^{(k)}, y^{(k)}) \rangle - Z_{zy}(\eta_{zy}(\phi)) \tag{5.1.6}$$

then conjugacy identifies the statistic of the prior with the natural parameter and log partition function of the likelihood via $t_\phi(\phi) = (\eta_{zy}(\phi), -Z_{zy}(\eta_{zy}(\phi)))$, so that

$$p(\phi|z^{(k)}, \bar{y}^{(k)}) \propto \exp\{\langle \eta_\phi + (t_{zy}(z^{(k)}, \bar{y}^{(k)}), 1), t_\phi(\phi) \rangle\}. \tag{5.1.7}$$

Conjugacy implies the optimal $q(\phi)$ has the same form as the prior; that is, without loss of generality we have $q(\phi) = \exp\{\langle \widetilde{\eta}_\phi, t_\phi(\phi) \rangle - Z_\phi(\widetilde{\eta}_\phi)\}$ for some variational parameter $\widetilde{\eta}_\phi$.

Given this structure, we can find a simple expression for the gradient of $\mathcal{L}$ with respect to the global variational parameter $\widetilde{\eta}_\phi$. 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 $Z \triangleq Z_\phi$. Then we have

$$\mathcal{L} = \mathbb{E}_{q(\phi)q(z)} [\ln p(\phi|z, \bar{y}) - \ln q(\phi)] + \text{const.} \tag{5.1.8}$$

$$= \langle \eta + \mathbb{E}_{q(z)}[t(z, \bar{y})], \nabla Z(\widetilde{\eta}) \rangle - (\langle \widetilde{\eta}, \nabla Z(\widetilde{\eta}) \rangle - Z(\widetilde{\eta})) + \text{const.} \tag{5.1.9}$$

where the constant term does not depend on $\widetilde{\eta}$ and where we have used the exponential family identity $\mathbb{E}_{q(\phi)} [t_\phi(\phi)] = \nabla Z(\widetilde{\eta})$ from Proposition 2.2.2. Differentiating over $\widetilde{\eta}$, we have

$$\nabla_{\widetilde{\eta}} \mathcal{L} = \left(\nabla^2 Z(\widetilde{\eta})\right) \left(\eta + \mathbb{E}_{q(z)}[t(z, \bar{y})] - \widetilde{\eta}\right). \tag{5.1.10}$$

The factor $\nabla^2 Z(\widetilde{\eta})$ is the Fisher information of the prior $p(\phi)$ and, because the prior and variational factor are in the same exponential family, it is also the Fisher information of the global variational factor $q(\phi)$. The natural gradient $\widetilde{\nabla}_{\widetilde{\eta}}$ can be defined in terms of the gradient [52] via $\widetilde{\nabla}_{\widetilde{\eta}} \triangleq \left(\nabla^2 Z(\widetilde{\eta})\right)^{-1} \nabla_{\widetilde{\eta}}$, and so we have

$$\widetilde{\nabla}_{\widetilde{\eta}} \mathcal{L} = \left(\eta + \mathbb{E}_{q(z)}[t(z, \bar{y})] - \widetilde{\eta}\right). \tag{5.1.11}$$

Expanding $q(z) = \prod_{i=1}^K q(z^{(k)})$ and $t(z, \bar{y}) \triangleq \sum_{k=1}^K (t_{zy}(z^{(k)}, \bar{y}^{(k)}), 1)$ we can write

$$\widetilde{\nabla}_{\widetilde{\eta}} \mathcal{L} = \left(\eta + \sum_{k=1}^K \mathbb{E}_{q(z^{(k)})}[t(z^{(k)}, \bar{y}^{(k)})] - \widetilde{\eta}\right) \tag{5.1.12}$$

and so the natural gradient decomposes into local terms as required for stochastic

Algorithm 5.2 Stochastic Variational Inference (SVI)

Initialize global variational parameter $\widetilde{\eta}_\phi^{(1)}$
for $t = 1, 2, \ldots$ **do**
  $\hat{k} \leftarrow$ sample Uniform$(\{1, 2, \ldots, K\})$
  $q^*(z^{(\hat{k})}) \leftarrow \textsc{LocalMeanField}(\widetilde{\eta}^{(t)}, \bar{y}^{(\hat{k})})$, e.g. Eq. (5.1.14)
  $\widetilde{\eta}_\phi^{(t+1)} \leftarrow (1-\rho^{(t)})\widetilde{\eta}_\phi^{(t)} + \rho^{(t)}\left(\eta_\phi + s \cdot \mathbb{E}_{q^*(z^{(\hat{k})})}\left[t(z^{(\hat{k})}, \bar{y}^{(\hat{k})})\right]\right)$

gradient optimization in (5.1.2).

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

$$\widetilde{\eta}_\phi \leftarrow (1-\rho^{(t)})\widetilde{\eta}_\phi + \rho^{(t)}\left(\eta_\phi + s \cdot \mathbb{E}_{q^*(z^{(k)})}[t(z^{(k)}, \bar{y}^{(k)})]\right) \tag{5.1.13}$$

where $q^*(x_{1:T})$ is defined below and where s scales the stochastic gradient update on the minibatch to represent the full size of the dataset; that is, if k is sampled uniformly and we use $|y|$ and $|y^{(k)}|$ to denote the sizes of the dataset and minibatch, respectively, we have $s = |y|/|y^{(k)}|$. In each step we find the optimal local factor $q^*(z^{(k)})$ using the standard mean field update from Proposition 2.3.3 and the current value of $q(\phi)$, i.e. we compute:

$$q^*(z^{(k)}) \propto \exp\left\{\mathbb{E}_{q(\phi)}[\ln p(z^{(k)}|\phi)p(\bar{y}^{(k)}|z^{(k)}, \phi)]\right\}. \tag{5.1.14}$$

We summarize the general SVI algorithm in Algorithm 5.2.

■ 5.2 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 $\bar{y} = \{\bar{y}_{1:T}^{(k)}\}_{k=1}^K$, and take each minibatch to be a single sequence written simply $\bar{y}_{1:T}$, suppressing the minibatch index k for simplicity. We also assume all sequences have the same initial state distribution $\pi^{(0)}$.

■ 5.2.1 SVI update for HMMs

Recall from Section 2.4 that a Bayesian HMM with N states defines a joint distribution over an initial state distribution $\pi^{(0)}$, a row-stochastic transition matrix A, observation parameters $\theta = \{\theta_i\}_{i=1}^N$, and K hidden state sequences $x_{1:T}^{(k)}$ and observation sequences $y_{1:T}^{(k)}$ for $k = 1, 2, \ldots, K$. We use $\pi^{(i)}$ to denote the ith row of A ($i = 1, 2, \ldots, N$) and $\pi = \{\pi_i\}_{i=0}^N$ to collect the transition rows and the initial state distribution. When

convenient, we use the alternative notations $p(\pi) = p(\pi^{(0)})p(A) = \prod_{i=0}^{N} p(\pi^{(i)})$ to denote the distribution over the initial state distribution and transition matrix and $p(\theta) = \prod_{i=1}^{N} p(\theta^{(i)})$ to denote the distribution over the observation parameters. The joint density for a Bayesian HMM is then

$$p(\pi^{(0)})p(A)p(\theta)\prod_{k=1}^{K} p(x_{1:T}^{(k)}, y_{1:T}^{(k)}|\pi^{(0)}, A, \theta). \tag{5.2.1}$$

In terms of the notation in Section 5.1.2, 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 derive explicit conjugate updates, we assume the observation model is conjugate in that $(p(\theta^{(i)}), p(y|\theta^{(i)}))$ is a conjugate pair of exponential family densities for each $i = 1, 2, \ldots, N$ and write

$$p(\pi^{(i)}) = p(\pi^{(i)}|\alpha^{(i)}) = \mathrm{Dir}(\alpha^{(i)}) \qquad i = 0, 1, \ldots, N \tag{5.2.2}$$

$$p(\theta^{(i)}) = p(\theta^{(i)}|\eta_\theta^{(i)}) = \exp\{\langle \eta_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle - Z_\theta^{(i)}(\eta_\theta^{(i)})\} \qquad i = 1, 2, \ldots, N \tag{5.2.3}$$

$$p(y_t|\theta^{(i)}) = \exp\{\langle t_\theta^{(i)}(\theta^{(i)}), (t_y^{(i)}(y_t), 1)\rangle\} \qquad i = 1, 2, \ldots, N. \tag{5.2.4}$$

Correspondingly the variational family is $q(\pi)q(A)q(\theta)\prod_{k=1}^{K} q(x_{1:T}^{(k)})$ with

$$q(\pi^{(i)}) = q(\pi^{(i)}|\widetilde{\alpha}^{(i)}) = \mathrm{Dir}(\widetilde{\alpha}^{(i)}) \qquad i = 0, 1, \ldots, N \tag{5.2.5}$$

$$q(\theta^{(i)}) = q(\theta^{(i)}|\widetilde{\eta}_\theta^{(i)}) = \exp\{\langle \widetilde{\eta}_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle - Z_\theta^{(i)}(\widetilde{\eta}_\theta^{(i)})\} \qquad i = 1, 2, \ldots, N. \tag{5.2.6}$$

That is, each variational factor is in the same (conjugate) prior family as the corresponding factor in the joint distribution p. Therefore we wish to optimize over the variational parameters for the initial state distribution $\widetilde{\alpha}^{(0)}$, the variational parameters for the transition distribution $\widetilde{\alpha}^{(i)}$ $(i = 1, 2, \ldots, N)$, and the variational parameters for the observation parameter distributions $\widetilde{\eta}_\theta$.

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 collect expected sufficient statistics with respect to the optimal factor for $q(x_{1:T})$, which in turn depends on the current value of $q(A)q(\pi^{(0)})q(\theta)$. Recall from Section 2.4.2 that we define

$$\widetilde{\pi}^{(i)} \triangleq \mathbb{E}_{q(\pi)}\left[\ln \pi^{(i)}\right] \qquad \widetilde{L}_{ij} \triangleq \mathbb{E}_{q(\theta)}\left[\ln p(\bar{y}_t|\theta^{(i)})\right] \tag{5.2.7}$$

Algorithm 5.3 HMM SVI

Initialize global variational parameters $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
for $t = 1, 2, \ldots$ **do**
  Sample minibatch index $\hat{k}$ uniformly from $\{1, 2, \ldots, K\}$
  Using minibatch $\bar{y}^{(\hat{k})}$, compute each $\hat{t}_y^{(i)}$, $\hat{t}_{\text{trans}}^{(i)}$, and $\hat{t}_{\text{init}}^{(i)}$
    with Eqs. (5.2.8)-(5.2.10)
  Update each $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
    with Eqs. (5.2.11)-(5.2.13)

and collect the $\widetilde{\pi}^{(i)}$ into a matrix $\widetilde{A}$, where the ith row of $\widetilde{A}$ is $\widetilde{\pi}^{(i)}$. Then using the HMM messages F and B defined in Section 2.4 we write the expected statistics as

$$\hat{t}_y^{(i)} \triangleq \mathbb{E}_{q(x_{1:T})} \sum_{t=1}^{T} \mathbb{I}[x_t = i] t_y^{(i)}(\bar{y}_t) = \sum_{t=1}^{T} F_{t,i} B_{t,i} \cdot (t_y^{(i)}(\bar{y}_t), 1)/Z \tag{5.2.8}$$

$$(\hat{t}_{\text{trans}}^{(i)})_j \triangleq \mathbb{E}_{q(x_{1:T})} \sum_{t=1}^{T-1} \mathbb{I}[x_t = i, x_{t+1} = j] = \sum_{t=1}^{T-1} F_{t,i} \widetilde{A}_{i,j} \widetilde{L}_{t+1,j} B_{t+1,j}/Z \tag{5.2.9}$$

$$(\hat{t}_{\text{init}})_i \triangleq \mathbb{E}_{q(x_{1:T})} \mathbb{I}[x_1 = i] = \widetilde{\pi}_0 B_{1,i}/Z \tag{5.2.10}$$

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

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

$$\widetilde{\eta}_\theta^{(i)} \leftarrow (1-\rho)\widetilde{\eta}_\theta^{(i)} + \rho(\eta_\theta^{(i)} + s \cdot \hat{t}_y^{(i)}) \tag{5.2.11}$$

$$\widetilde{\alpha}^{(i)} \leftarrow (1-\rho)\widetilde{\alpha}^{(i)} + \rho(\alpha^{(i)} + s \cdot \hat{t}_{\text{trans}}^{(i)}) \tag{5.2.12}$$

$$\widetilde{\alpha}^{(0)} \leftarrow (1-\rho)\widetilde{\alpha}^{(0)} + \rho(\alpha^{(0)} + s \cdot \hat{t}_{\text{init}}^{(i)}) \tag{5.2.13}$$

where $s = |\bar{y}|/|\bar{y}^{(k)}|$ scales the minibatch gradient to represent the full dataset, as in Section 5.1. When the dataset comprises K sequences where the length of sequence k is $T^{(k)}$, we have $s = (\sum_{k'=1}^{K} T^{(k')})/T^{(k)}$.

We summarize the overall algorithm in 5.3.

■ 5.2.2 SVI update for HSMMs

The SVI updates for the HSMM are similar to those for the HMM with the addition of a duration update, though expressing the expected sufficient statistics in terms of the

HSMM messages is substantially different. The form of these expected statistics follows from the HSMM E-step [78, 54].

To derive explicit updates, we assume the duration prior and likelihood are a conjugate pair of exponential families. Writing the duration parameters as $\vartheta = \{\vartheta^{(i)}\}_{i=1}^N$, we can write the prior, variational factor, and likelihood up to proportionality as

$$p(\vartheta^{(i)}) \propto \exp\{\langle \eta_\vartheta^{(i)}, t_\vartheta^{(i)}(\vartheta^{(i)})\rangle\}, \tag{5.2.14}$$

$$p(d|\vartheta^{(i)}) = \exp\{\langle t_\vartheta^{(i)}(\vartheta^{(i)}), (t_d(d), 1)\rangle\}, \tag{5.2.15}$$

$$q(\vartheta^{(i)}) \propto \exp\{\langle \widetilde{\eta}_\vartheta^{(i)}, t_\vartheta^{(i)}(\vartheta^{(i)})\rangle\}. \tag{5.2.16}$$

Using the HSMM messages (F, F^*) and (B, B^*) with $\widetilde{L}$ and $\widetilde{A}$ from the previous section, we can write

$$(\hat{t}_{\text{trans}}^{(i)})_j \triangleq \mathbb{E}_{q(x_{1:T})} \sum_{t=1}^{T-1} \mathbb{1}[x_t = i, x_{t+1} = j, x_t \neq x_{t+1}] \tag{5.2.17}$$

$$= \sum_{t=1}^{T-1} F_{t,i} B^*_{t,j} \widetilde{A}_{i,j} / Z \tag{5.2.18}$$

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

To be written in terms of the HSMM messages the expected label sequence indicators $\mathbb{1}[x_t = i]$ must be expanded to

$$\mathbb{1}[x_t = i] = \sum_{\tau < t} \mathbb{1}[x_{\tau+1} = i, x_\tau \neq x_{\tau+1}] - \mathbb{1}[x_\tau = i, x_\tau \neq x_{\tau+1}]. \tag{5.2.19}$$

Intuitively, this expansion expresses that a state is occupied after a transition into it occurs and until the first transition occurs out of that state and to another. Then we have

$$\mathbb{E}_{q(x_{1:T})} \mathbb{1}[x_{t+1} = i, x_t \neq x_{t+1}] = F^*_{t,i} B^*_{t,i} / Z \tag{5.2.20}$$

$$\mathbb{E}_{q(x_{1:T})} \mathbb{1}[x_t = i, x_t \neq x_{t+1}] = F_{t,i} B_{t,i} / Z. \tag{5.2.21}$$

from which we can compute $\mathbb{E}_{q(x_{1:T})} \mathbb{1}[x_t = i]$, which we use in the definition of $\hat{t}_y^{(i)}$ given in (5.2.8).

Finally, defining $\widetilde{D}_{di} \triangleq \mathbb{E}_{q(\vartheta)} \left[p(d|\vartheta^{(i)})\right]$, we compute the expected duration statistics as indicators on every possible duration d via

Algorithm 5.4 HSMM SVI

Initialize global variational parameters $\widetilde{\eta}_\vartheta^{(i)}$, $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
for $t = 1, 2, \ldots$ **do**
  Sample minibatch index $\hat{k}$ uniformly from $\{1, 2, \ldots, K\}$
  Using minibatch $\bar{y}^{(\hat{k})}$, compute each $\hat{t}_{\text{dur}}^{(i)}$, $\hat{t}_y^{(i)}$, $\hat{t}_{\text{trans}}^{(i)}$, and $\hat{t}_{\text{init}}^{(i)}$
    with Eqs. (5.2.8),(5.2.10), (5.2.18), and (5.2.23)
  Update each $\widetilde{\eta}_\vartheta^{(i)}$, $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
    with Eqs. (5.2.11)-(5.2.13) and (5.2.24)

$$(\hat{t}_{\text{dur}}^{(i)})_d \triangleq \mathbb{E}_{q(x_{1:T})}\left[\sum_t \mathbb{1}[x_t \neq x_{t+1}, x_{t+1:t+d} = i, x_{t+d+1} \neq i]\right] \tag{5.2.22}$$

$$= \sum_{t=1}^{T-d+1} \widetilde{D}_{d,i} F^*_{t,i} B_{t+d,i} \Big(\prod_{t'=t}^{t+d} \widetilde{L}_{t',i}\Big)/Z. \tag{5.2.23}$$

Note that this step alone requires $\mathcal{O}(T^2 N)$ time.

With these expected statistics, the updates to the observation, transition, and initial state factors are (5.2.11), (5.2.12), and (5.2.13). The duration factor update is

$$\widetilde{\eta}_\vartheta^{(i)} \leftarrow (1-\rho)\widetilde{\eta}_\vartheta^{(i)} + \rho\Big(\eta_\vartheta^{(i)} + s\Big(\sum_{d=1}^{T} (\hat{t}_{\text{dur}}^{(i)})_d \cdot (t_d(d), 1)\Big)\Big). \tag{5.2.24}$$

We summarize the overall algorithm in 5.4.

While these updates can be used for any family of duration models, they can be computationally expensive: as described in Chapter 4, both computing the HSMM messages and computing the expected statistics (5.2.22) require time that scales quadratically with the sequence length T, which can be severely limiting even in the minibatch setting. In the next section, we apply the techniques developed in Chapter 4 to the SVI algorithm to derive updates for which the computational complexity scales only linearly with T.

■ 5.3 Linear-time updates for negative binomial HSMMs

General HSMM inference is much more expensive than HMM inference, having runtime $\mathcal{O}(T^2N + TN^2)$ compared to just $\mathcal{O}(TN^2)$ 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 [52, 15]. In this section, we develop approximate SVI updates for a particular class of duration

distributions with unbounded support for which the computational complexity is only linear in T.

Following the development in Chapter 4, we consider HSMMs with negative binomial duration distributions. Each duration likelihood has parameters r and p with the form

$$p(k|r,p) = \binom{k+r-2}{k-1} \exp\{(k-1)\ln p + r\ln(1-p)\} \tag{5.3.1}$$

for $k = 1, 2, \ldots$. The negative binomial likelihood is not an exponential family of densities over (r,p), and it has no simple conjugate prior. We use priors of the form $p(r,p) = p(r)p(p)$ with $p(r)$ a finite categorical distribution with support $\{1,2,\ldots,r_{\max}\}$ and $p(p)$ an independent Beta distribution, i.e.

$$p(r) \propto \exp\{\langle \nu, \mathbb{1}_r \rangle\}, \quad p(p) = \mathrm{Beta}(a,b) \propto \exp\{(a-1)\ln(p) + (b-1)\ln(1-p)\}. \tag{5.3.2}$$

Similarly, we define a corresponding mean field factor $q(r,p) = q(r)q(p|r)$ as

$$q(r) \propto \exp\{\langle \widetilde{\nu}, \mathbb{1}_r \rangle\}, \qquad q(p|r) = \mathrm{Beta}(\widetilde{a}^{(r)}, \widetilde{b}^{(r)}). \tag{5.3.3}$$

Thus for N states we have prior hyperparameters $\{(\nu^{(i)}, a^{(i)}, b^{(i)})\}_{i=1}^N$ and variational parameters $\{(\nu^{(i)}, \{a^{(r,i)}, b^{(r,i)}\}_{r=1}^{r_{\max}})\}_{i=1}^N$. To simplify notation, we suppress the indices r and i when possible.

We write $d^{(i)}(x_{1:T})$ to denote the set of durations for state i in the state sequence $x_{1:T}$. Dropping indices for simplicity, the part of the variational lower bound objective that depends on $q(r,p)$ is

$$\mathcal{L} \triangleq \mathbb{E}_{q(r,p)q(x_{1:T})}\left[\ln \frac{p(r,p,d(x_{1:T}))}{q(r,p)}\right] \tag{5.3.4}$$

$$= \mathbb{E}_{q(r)} \ln \frac{p(r)}{q(r)} + \mathbb{E}_{q(r)q(x_{1:T})} h(r, d(x_{1:T})) + \mathbb{E}_{q(r)}\left\{\mathbb{E}_{q(p|r)} \ln \frac{p(p)\bar{p}(d(x_{1:T})|r,p)}{q(p|r)}\right\} \tag{5.3.5}$$

where $h(r, d(x_{1:T})) \triangleq \sum_{d' \in d(x_{1:T})} \ln \binom{r+d'-2}{d'-1}$ arises from the negative binomial base measure term and $\ln \bar{p}(d(x_{1:T})|r,p) \triangleq \sum_{d' \in d(x_{1:T})} (d' \ln p + r \ln(1-p))$ collects the negative binomial PMF terms excluding the base measure.

First, we show that the SVI updates to each $q(p|r)$ can be considered independent of each other and of $q(r)$ by taking the natural gradient of $\mathcal{L}$. The only terms in (5.3.4) that depend on $q(p|r)$ are in the final term. Since the expectation over $q(r)$ in the

final term is simply a weighted finite sum, taking the gradient of $\mathcal{L}$ with respect to the parameters $(\widetilde{a}^{(r)}, \widetilde{b}^{(r)})$ for $r = 1, 2, \ldots, r_{\max}$ yields a sum of gradients weighted by each $q(r)$. Each gradient in the sum is that of a variational lower bound with fixed r, and because $q(p|r)$ is conjugate to the negative binomial likelihood with fixed r, each gradient has a simple conjugate form. As a result of this decomposition, if we collect the variational parameters of $q(r, p)$ into $\widetilde{\eta}_\vartheta \triangleq (\widetilde{\nu}, (\widetilde{a}^{(1)}, \widetilde{b}^{(1)}), \ldots, (\widetilde{a}^{(r_{\max})}, \widetilde{b}^{(r_{\max})}))$, then the Fisher information matrix

$$J(\widetilde{\eta}_\vartheta) \triangleq \mathbb{E}_{(r,p)\sim q(r,p)} \left[(\nabla_{\widetilde{\eta}_\vartheta} \ln q(r,p))(\nabla_{\widetilde{\eta}_\vartheta} \ln q(r,p))^{\mathsf{T}} \right] \tag{5.3.6}$$

is block diagonal with the same partition structure as $\widetilde{\eta}_\vartheta$. If we denote the Fisher information of $q(p|r)$ as $J(\widetilde{a}^{(r)}, \widetilde{b}^{(r)})$, then the $(r+1)$th diagonal block of $J(\widetilde{\eta}_\vartheta)$ can be written as $q(r)J(\widetilde{a}^{(r)}, \widetilde{b}^{(r)})$, and so the $q(r)$ factors cancel in the natural gradient. Therefore the natural gradient updates to each $(\widetilde{a}^{(r)}, \widetilde{b}^{(r)})$ are independent and can be computed using simple conjugate Beta updates.

Next, we derive updates to $q(r)$. Since $q(r)$ is a discrete distribution with finite support, we write its complete-data conditional in an exponential family form trivially:

$$p(r|p, d(x_{1:T})) \propto \exp\{\langle \nu + t_r(p, d(x_{1:T})), \mathbb{1}_r \rangle\} \tag{5.3.7}$$

$$(t_r(p, d(x_{1:T})))_r \triangleq \textstyle\sum_{d' \in d(x_{1:T})} \ln p(p|d', r) + \ln h(r, d'). \tag{5.3.8}$$

From the results in Section 5.1.2 the jth component of the natural gradient of (5.3.4) with respect to the parameters of $q(r)$ is

$$\left(\widetilde{\nabla}_{\widetilde{\nu}} \mathcal{L}\right)_j = \nu_j + \mathbb{E}_{q(p|r=j)q(x_{1:T})} t_r(p, d(x_{1:T})) - \widetilde{\nu}_j \tag{5.3.9}$$

Due to the log base measure term $\ln h(r, d')$ in (5.3.8), these expected statistics require $\mathcal{O}(T^2 N)$ time to compute exactly even after computing the HSMM messages using (5.2.23). The HSMM SVI algorithm developed in Section 5.2.2 provides an exact algorithm using this update. However, we can use the efficient sampling-based algorithms developed in Chapter 4 to compute an approximate update more efficiently.

To achieve an update runtime that is linear in T, we use a sampling method inspired by the sampling-based SVI update used in Wang and Blei [114]. For some sample count S, we collect S model parameter samples $\{(\hat{\pi}^{(\ell)}, \hat{\theta}^{(\ell)}, \hat{r}^{(\ell)}, \hat{p}^{(\ell)})\}_{\ell=1}^{S}$ using the current global mean field factors according to

$$\hat{\pi}^{(\ell)} \sim q(\pi) \qquad \hat{\theta}^{(\ell)} \sim q(\theta) \qquad (\hat{r}^{(\ell)}, \hat{p}^{(\ell)}) \sim q(r, p). \tag{5.3.10}$$

and for each set of parameters we sample a state sequence

$$\hat{x}_{1:T}^{(\ell)} \sim p(x_{1:T}|\bar{y}_{1:T}, \hat{\pi}^{(\ell)}, \hat{\theta}^{(\ell)}, \hat{r}^{(\ell)}, \hat{p}^{(\ell)}). \tag{5.3.11}$$

Using the methods developed in Chapter 4, each such sample can be drawn in time $\mathcal{O}(TNR + TN^2)$. We denote the set of state sequence samples as $\mathcal{S} = \{\hat{x}_{1:T}^{(\ell)}\}_{\ell=1}^{S}$ and we set $\hat{q}(x_{1:T}) = \frac{1}{S}\sum_{\hat{x}\in\mathcal{S}} \delta_{\hat{x}}(x_{1:T})$. As the number of samples S grows, the distribution $\hat{q}(x_{1:T})$ approximates $\mathbb{E}_{q(\pi)q(\theta)q(r,p)}\left[p(x_{1:T}|\bar{y}_{1:T}, \pi, \theta, r, p)\right]$, while the optimal mean field update sets $q(x_{1:T}) \propto \exp\left\{\mathbb{E}_{q(\pi)q(\theta)q(r,p)} \ln p(x_{1:T}|\bar{y}_{1:T}, \pi, \theta, r, p)\right\}$. As discussed in Wang and Blei [114], since this sampling approximation does not optimize the variational lower bound directly, it should yield an inferior objective value. However, Wang and Blei [114] found this approximate SVI update yielded better predictive performance in some topic models, and provided an interpretation as an approximate expectation propagation (EP) update. As we show in Section 5.5, this update can be very effective for fitting HSMMs as well.

Given the sample-based representation $\hat{q}(x_{1:T})$, it is easy to compute the expectation over states in (5.3.9) by plugging in the sampled durations. The update to the parameters of $q(r^{(i)}, p^{(i)})$ becomes

$$\widetilde{\nu}^{(i)} \leftarrow (1-\rho)\widetilde{\nu}^{(i)} + \rho\left(\nu^{(i)} + s \cdot \hat{t}_r^{(i)}\right) \tag{5.3.12}$$

$$\widetilde{a}^{(i,r)} \leftarrow (1-\rho)\widetilde{a}^{(i,r)} + \rho\left(a^{(i)} + s \cdot \hat{t}_a^{(i,r)}\right) \tag{5.3.13}$$

$$\widetilde{b}^{(i,r)} \leftarrow (1-\rho)\widetilde{b}^{(i,r)} + \rho\left(b^{(i)} + s \cdot \hat{t}_b^{(i,r)}\right) \tag{5.3.14}$$

for $i = 1, 2, \ldots, N$ and $r = 1, 2, \ldots, r_{\max}$, where

$$\hat{t}_a^{(i,r)} \triangleq \frac{1}{S}\sum_{\hat{x}\in\mathcal{S}} \sum_{d\in d^{(i)}(x)} (d-1) \tag{5.3.15}$$

$$\hat{t}_b^{(i,r)} \triangleq \frac{1}{S}\sum_{\hat{x}\in\mathcal{S}} \sum_{d\in d^{(i)}(\hat{x})} r \tag{5.3.16}$$

$$(\hat{t}_r^{(i)})_r \triangleq \mathbb{E}_{q(p|r)\hat{q}(x_{1:T})}\left[t_r(p, d^{(i)}(\hat{x}_{1:T}))\right] \tag{5.3.17}$$

$$\begin{aligned} &= \left(\widetilde{a}^{(i,r)} + \hat{t}_a^{(i,r)} - 1\right)\mathbb{E}_{q(p|r)}\left[\ln(p^{(i,r)})\right] + \left(\widetilde{b}^{(i,r)} + \hat{t}_b^{(i,r)} - 1\right)\mathbb{E}_{q(p|r)}\left[\ln(1-p^{(i,r)})\right] \\ &\quad + \sum_{\hat{x}\in\mathcal{S}} \sum_{d\in d^{(i)}(\hat{x})} \ln\binom{d+r-2}{d-1}. \end{aligned} \tag{5.3.18}$$

Similarly, we revise Eqs. (5.2.8)-(5.2.10) to compute the other expected sufficient statis-

Algorithm 5.5 Negative Binomial HSMM SVI

Initialize global variational parameters $\widetilde{\eta}_\vartheta^{(i)}$, $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
for $t = 1, 2, \ldots$ **do**
  Sample minibatch index $\hat{k}$ uniformly from $\{1, 2, \ldots, K\}$
  Using minibatch $\bar{y}^{(\hat{k})}$, generate state sequence samples
    according to Eqs. (5.3.10) and (5.3.11) and form $\hat{q}(x_{1:T})$
  Using $\hat{q}(x_{1:T})$, compute each $\hat{t}_{\text{dur}}^{(i)}$, $\hat{t}_y^{(i)}$, $\hat{t}_{\text{trans}}^{(i)}$, and $\hat{t}_{\text{init}}^{(i)}$
    with Eqs. (5.3.19)-(5.3.21) and (5.3.15)-(5.3.18)
  Update each $\widetilde{\eta}_\vartheta^{(i)}$, $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\alpha}^{(0)}$
    with Eqs. (5.2.11)-(5.2.13) and (5.3.12)-(5.3.14)

tics using $\hat{q}(x_{1:T})$:

$$\hat{t}_y^{(i)} \triangleq \mathbb{E}_{\hat{q}(x_{1:T})} \sum_{t=1}^{T} \mathbb{I}[x_t = i] t_y^{(i)}(\bar{y}_t) \tag{5.3.19}$$

$$(\hat{t}_{\text{trans}}^{(i)})_j \triangleq \mathbb{E}_{\hat{q}(x_{1:T})} \sum_{t=1}^{T-1} \mathbb{I}[x_t = i, x_{t+1} = j] \tag{5.3.20}$$

$$(\hat{t}_{\text{init}})_i \triangleq \mathbb{E}_{\hat{q}(x_{1:T})} \mathbb{I}[x_1 = i] \tag{5.3.21}$$

We summarize the overall algorithm in 5.5.

■ 5.4 Extending to the HDP-HMM and HDP-HSMM

In this section we extend our methods to the Bayesian nonparametric versions of these models, the HDP-HMM and the HDP-HSMM. These updates essentially replace the transition updates in the previous algorithms.

Using the notation of Section 2.5 the generative model for the HDP-HMM with scalar concentration parameters $\alpha, \gamma > 0$ is

$$\beta \sim \text{GEM}(\gamma), \qquad \pi^{(i)} \sim \text{DP}(\alpha\beta), \qquad \theta^{(i)} \overset{\text{iid}}{\sim} p(\theta^{(i)}) \tag{5.4.1}$$

$$x_1 \sim \pi^{(0)}, \qquad x_{t+1} \sim \pi^{(x_t)}, \qquad y_t \sim p(y_t|\theta^{(x_t)}) \tag{5.4.2}$$

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

$$v_j \overset{\text{iid}}{\sim} \text{Beta}(1, \gamma), \qquad \beta_k = \prod_{j<k} (1 - v_j) v_k \tag{5.4.3}$$

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

$$w \sim \mathrm{GEM}(\alpha) \qquad z_k \overset{\text{iid}}{\sim} \beta \qquad \pi^{(i)} = \sum_{k=1}^{\infty} w_k \delta_{z_k}. \tag{5.4.4}$$

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 [52], a more convenient truncation for our models is the "direct assignment" truncation, used in [70] for batch mean field with the HDP-HMM and in [17] in an SVI algorithm for LDA. The direct assignment truncation limits the support of $q(x_{1:T})$ to the finite set $\{1, 2, \ldots, M\}^T$ for a truncation parameter M, i.e. fixing $q(x_{1:T}) = 0$ when any $x_t > M$. Thus the other factors, namely $q(\pi)$, $q(\beta)$, and $q(\theta)$, only differ from their priors in their distribution over the first M components. As opposed to standard truncation, this family of approximations is nested over M, enabling a search procedure over the truncation parameter as developed in [17]. A similar search procedure can be used with the HDP-HMM and HDP-HSMM algorithms developed 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. [52]. Following Liang et al. [70], 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

$$p((\pi^{(i)}_{1:M}, \pi^{(i)}_{\text{rest}})) = \mathrm{Dir}(\alpha \cdot (\beta_{1:M}, \beta_{\text{rest}})) \tag{5.4.5}$$

$$q((\pi^{(i)}_{1:M}, \pi^{(i)}_{\text{rest}})) = \mathrm{Dir}(\widetilde{\alpha}^{(i)}) \tag{5.4.6}$$

where $i = 1, 2, \ldots, M$ and where $\pi^{(i)}_{\text{rest}} \triangleq 1 - \sum_{k=1}^{M} \pi^{(i)}_k$ and $\beta_{\text{rest}} \triangleq 1 - \sum_{k=1}^{M} \beta_k$. Therefore the updates to $q(\pi^{(i)})$ are identical to those in (5.2.12) except the number of variational parameters is $M + 1$ and the prior hyperparameters are replaced with $\alpha \cdot (\beta_{1:M}, \beta_{\text{rest}})$.

To derive a gradient of the variational objective with respect to β^*, we write

$$\nabla_{\beta^*}\mathcal{L} = \nabla_{\beta^*}\left\{\mathbb{E}_{q(\pi)}\left[\ln\frac{p(\beta,\pi)}{q(\beta)q(\pi)}\right]\right\} \tag{5.4.7}$$

$$= \nabla_{\beta^*}\left\{\ln p(\beta^*) + \sum_{i=1}^{M}\mathbb{E}_{q(\pi^{(i)})}\ln p(\pi^{(i)}|\beta^*)\right\} \tag{5.4.8}$$

where $\ln p(\beta^*) = \ln p_v(v(\beta^*)) + \ln\det\left.\frac{\partial v}{\partial\beta}\right|_{\beta^*}$, $\ln p_v(v) = (\gamma-1)\sum_j \ln(1-v_j)$, and $v_i(\beta) = \frac{\beta_i}{1-\sum_{j<i}\beta_j}$. The Jacobian $\frac{\partial v}{\partial\beta}$ is lower-triangular, and is given by

$$\left(\frac{\partial v}{\partial\beta}\right)_{ij} = \begin{cases} 0 & i<j \\ \frac{1}{1-\sum_{k<i}\beta_k} & i=j \\ \frac{-\beta_i}{(1-\sum_{k<i}\beta_k)^2} & i>j \end{cases} \tag{5.4.9}$$

and so taking partial derivatives we have

$$\frac{\partial}{\partial\beta_k^*}\ln p(\beta^*) = 2\sum_{i\geq k}\ln\frac{1}{1-\sum_{j<i}\beta_j^*} - (\gamma-1)\sum_{i\geq k}\ln\frac{1}{1-\sum_{j\leq i}\beta_j^*} \tag{5.4.10}$$

$$\frac{\partial}{\partial\beta_k^*}\mathbb{E}_{q(\pi)}[\ln p(\pi^{(i)}|\beta^*)] = \gamma\psi(\widetilde{\alpha}_k^{(i)}) - \gamma\psi(\widetilde{\alpha}_{M+1}^{(i)}) + \gamma\psi(\gamma\textstyle\sum_{j=1}^{M+1}\beta_j^*) - \gamma\psi(\beta_k^*). \tag{5.4.11}$$

We use this gradient expression to take a truncated gradient step on β^* during each SVI update, where we use a backtracking line search[1] to ensure the updated value satisfies the constraint $\beta^* \geq 0$.

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

$$q((\pi_{1:M\setminus i}^{(i)}, \pi_{\text{rest}}^{(i)})) = \text{Dir}(\widetilde{\alpha}_{\setminus i}^{(i)}), \tag{5.4.12}$$

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

$$\frac{\partial}{\partial\beta_k^*}\mathbb{E}_{q(\pi)}[\ln p(\pi^{(i)}|\beta^*)] = \begin{cases} \gamma\psi(\widetilde{\alpha}_k^{(i)}) - \gamma\psi(\widetilde{\alpha}_{M+1}^{(i)}) + \gamma\psi(\gamma\sum_{j\neq i}\beta_j^*) - \gamma\psi(\beta_k^*) & k\neq i \\ 0 & k=i \end{cases}. \tag{5.4.13}$$

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

[1] In a backtracking line search, for some fixed parameter $\kappa \in (0,1)$, given an initial point x and an increment Δ, while $x+\Delta$ is infeasible we set $\Delta \leftarrow \kappa\Delta$.

5.5 Experiments

We conclude this chapter with a numerical study to validate the proposed algorithms.

As a performance metric, we approximate a variational posterior predictive density on held-out data; that is, for the HMM models we estimate

$$p(\bar{y}_{\text{test}}|\bar{y}_{\text{train}}) = \int\int p(\bar{y}_{\text{test}}|\pi,\theta)p(\pi,\theta|\bar{y}_{\text{train}})\mathrm{d}\pi\mathrm{d}\theta \tag{5.5.1}$$

$$\approx \mathbb{E}_{q(\pi)q(\theta)}p(\bar{y}_{\text{test}}|\pi,\theta) \tag{5.5.2}$$

by sampling models from the fit variational distribution. Similarly, for HSMM models we estimate $p(\bar{y}_{\text{test}}|\bar{y}_{\text{train}}) \approx \mathbb{E}_{q(\pi)q(\theta)q(\vartheta)}p(\bar{y}_{\text{test}}|\pi,\theta,\vartheta)$. In each experiment, we chose $\rho^{(t)} = (t+\tau)^{-\kappa}$ with $\tau = 0$ and $\kappa = 0.6$. Gaussian emission parameters were generated from Normal-Inverse-Wishart (NIW) distributions with $\mu_0 = 0$, $\Sigma_0 = I$, $\kappa_0 = 0.1$, and $\nu_0 = 7$. For the HDP models, we set the truncation parameters to be twice the true number of modes. Every SVI algorithm examined uses only a single pass through the training data.

First, we compare the performance of SVI and batch mean field algorithms for the HDP-HMM on synthetic data with fully conjugate priors. We sampled a 10-state HMM with 2-dimensional Gaussian emissions and generated a dataset of 250 sequences of length 4000 for a total of 10^6 frames. We chose a random subset of 95% of the generated sequences to be training sequences and held out 5% as test sequences. We repeated the fitting procedures on the training set 5 times with initializations drawn from the prior, and we report the average performance with standard deviation error bars. In Figure 5.2, the SVI procedure (in blue) produces fits that are on par with those from the batch algorithm (in green) but orders of magnitude faster. In particular, note that the SVI algorithm consistently converges to a local optimum of the mean field objective in a single pass through the training set, requiring roughly the amount of time needed for a single iteration of the batch mean field algorithm. This relative speedup grows linearly with the size of the dataset, making the SVI algorithm especially useful when the batch algorithm is infeasible.

Similarly, we compare the SVI and batch mean field algorithms for the HDP-HSMM. We sampled a 6-state HSMM with 2-dimensional Gaussian emissions and negative binomial durations, where each of the negative binomial parameters were sampled as $p \sim \text{Beta}(1,1)$ and $r \sim \text{Uniform}(\{1,2,\ldots,10\})$. From the model we generated a dataset of 50 sequences of length 2000 and generated an additional test set of 5 sequences with the same length. Figure 5.3 shows again that the SVI procedure (in blue) fits the data orders of magnitude faster than the batch update (in green), and again it requires only a single pass through the training set.

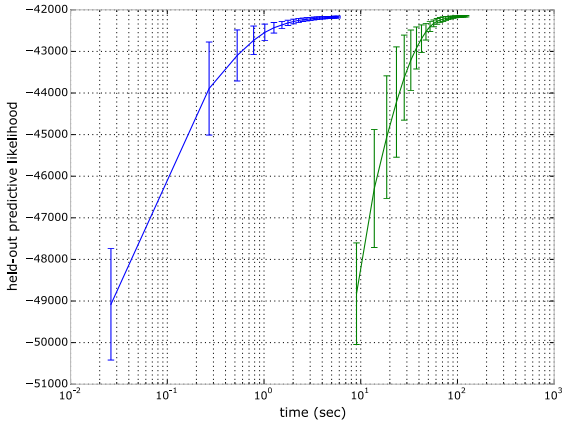

Figure 5.2: A comparison of the HMM SVI algorithm with batch mean field. Algoritm 5.3 is shown in blue and the batch mean field algorithm is shown in green.

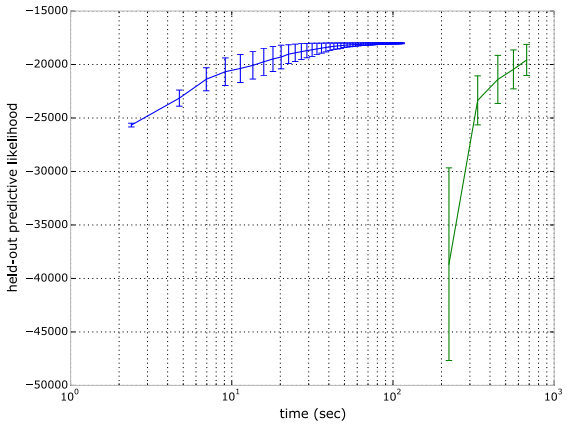

Figure 5.3: A comparison of the HSMM SVI algorithm with batch mean field. Algorithm 5.4 is shown in blue and the batch mean field algorithm is shown in green.

Finally, we compare the performance of the exact SVI update for the HSMM with that of the approximate update proposed in Section 5.3. We sampled a 6-state HSMM with 2-dimensional Gaussian emissions and Poisson durations, where each of the Poisson duration parameters is sampled as $\lambda \sim \text{Gamma}(40, 2)$. From the model we generated a dataset of 50 sequences of length 3000 and generated an additional test set of 5 sequences with the same length. We fit the data with negative binomial HDP-HSMMs where the priors on the negative binomial parameters were again $p \sim \text{Beta}(1, 1)$ and $r \sim \text{Uniform}(\{1, 2, \ldots, 10\})$. We set the number of state sequence samples generated in

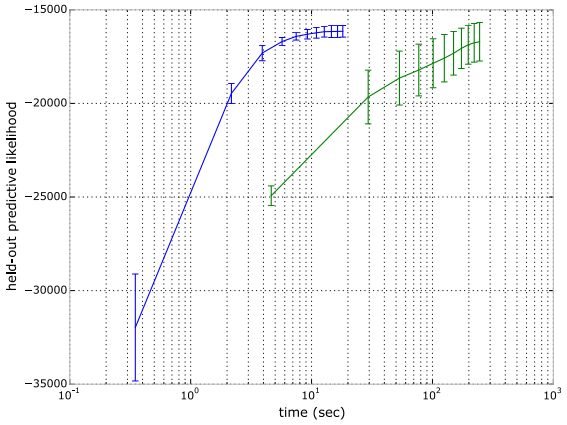

Figure 5.4: A comparison of HSMM SVI algorithms. The approximate update scheme of Algorithm 5.5 is shown in blue and the exact update scheme of Algorithm 5.4 is shown in green.

the sampling-based approximate update to $S = 10$. Figure 5.4 shows that the sampling-based updates (in blue) are effective and that they provide a significant speedup over the exact SVI update (in green). Note that, since the figure compares two SVI algorithms, both algorithms scale identically with the size of the dataset. However, the time required for the exact update scales quadratically with the minibatch sequence length T, while the sampling-based update scales only linearly with T. Therefore this approximate SVI update is most useful when minibatch sequences are long enough so that the exact update is infeasible.