Supplemental Materials for: Stochastic Gradient MCMC Methods for Hidden Markov Models

Yi-An Ma, Nicholas J. Foti, Emily B. Fox

1 Gradient of the Posterior

For the hidden Markov model (HMM), the posterior distribution of all hyperparameters θ can be calculated by the Bayes rule, where

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta).$$

Since

$$p(\mathbf{y}, \mathbf{x}|\theta) = \pi_0(x_0)\prod_{t=1}^{T} A_{x_t,x_{t-1}} \cdot \prod_{t=1}^{T} p(y_t|x_t),$$

where $\mathbf{y} = (y_1, \cdots y_T)$ denotes the data as real valued vector, and $\mathbf{x} = (x_1, \cdots x_T)$ as discrete valued vector with $x_t \in \{1, \cdots K\}, \forall t$. We can directly marginalize out the hidden variables, $\mathbf{x}$, with matrix multiplication as

$$p(\mathbf{y}|\theta) = \mathbf{1}_T^{\mathrm{T}}\, P(y_T)A \cdots P(y_1)A\, \boldsymbol{\pi}_0,$$

where $P(y_T)$ is a diagonal matrix and $P_{i,j}(y_t) = p(y_t|x_t = i)\delta_{i,j}$; $\mathbf{1}_T^{\mathrm{T}} = (1, \cdots, 1)$ is a row vector of k ones ($^{\mathrm{T}}$ denotes transpose); $(\boldsymbol{\pi}_0)_i = \pi_0(x_0 = i)$. Hence the same as Eq. (2), (3) and (8) of the main paper, the posterior distribution is:

$$p(\theta|\mathbf{y}) = \mathbf{1}_T^{\mathrm{T}}\, P(y_T)A \cdots P(y_1)A\, \boldsymbol{\pi}_0 \cdot p(\theta).$$

When we divide the whole sequence into subsequences of

$$\mathbf{y}_{\tau,L} = (y_{\tau-L}, \ldots, y_\tau, \ldots, y_{\tau+L}),$$

the posterior can be rewritten as:

$$p(\theta|\mathbf{y}) \propto \mathbf{1}^{\mathrm{T}} \prod_{\mathbf{y}_{\tau,L}\in\mathcal{S}} P(\mathbf{y}_{\tau,L})\boldsymbol{\pi}_0 \cdot p(\theta), \tag{1}$$

where $\mathcal{S}$ is the minimum set of $\mathbf{y}_{\tau,L}$ covering $\mathbf{y}$.

We can then use gradient information of the posterior distribution to construct MCMC algorithms. The gradient of the log-posterior distribution is:

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta_i} = \sum_{\tau=1}^{|\mathcal{S}|} \frac{\mathbf{1}^{\mathrm{T}} P(\mathbf{y}_{|\mathcal{S}|,L})A \cdots \dfrac{\partial\left(P(\mathbf{y}_{\tau,L})A\right)}{\partial\theta_i} \cdots P(\mathbf{y}_{1,L})A\boldsymbol{\pi}_0}{\mathbf{1}^{\mathrm{T}} P(\mathbf{y}_{|\mathcal{S}|,L})A \cdots P(\mathbf{y}_{\tau,L})A \cdots P(\mathbf{y}_{1,L})A\boldsymbol{\pi}_0} + \frac{\partial \ln p(\theta)}{\partial\theta_i}.$$

Denote $\mathbf{q}^{\mathrm{T}}_{\tau+L+1} = \mathbf{1}^{\mathrm{T}}_{T} P(y_T)A \cdots P(y_{t+1})A$ and $\boldsymbol{\pi}_{\tau-L-1} = P(y_{t-1})A \cdots P(y_1)A\boldsymbol{\pi}_0$. Then

$$\begin{aligned} \frac{\partial U(\theta)}{\partial\theta_i} &= -\frac{\partial \ln p(\mathbf{y}|\theta)}{\partial\theta_i} - \frac{\partial p(\theta)}{\partial\theta_i} \\ &= -\sum_{\mathbf{y}_\tau\in\widetilde{\mathcal{S}}} \frac{\mathbf{q}^{\mathrm{T}}_{\tau+L+1} \dfrac{\partial P(\mathbf{y}_\tau)}{\partial\theta_i} \boldsymbol{\pi}_{\tau-L-1}}{\mathbf{q}^{\mathrm{T}}_{\tau+L+1} P(\mathbf{y}_\tau)\boldsymbol{\pi}_{\tau-L-1}} - \frac{\partial \ln p(\theta)}{\partial\theta_i}, \end{aligned} \tag{2}$$

as shown in Eq. (11) of the main paper.

2 Lyapunov Exponent

The question of buffer length is equivalent to: for two random vectors $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$, what's the expected length of LB such that after the application of $P(\mathbf{y}_{LB})$, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ will synchronize? This is a question of random dynamical systems and can be answered through defining the *Lyapunov exponent*.

We first transform $\boldsymbol{\pi}$ through stereographic projection into $K-1$ dimensions and denote as: $\mathbf{r}$. Then operator $P(y_t)A[\,\cdot\,]$ is projected to new space and the equivalent dynamics over $\mathbf{r}$ becomes: F_{y_t}. We define the Lyapunov exponent $\mathfrak{L}$ through the projected random dynamics F_{y_t} as

$$\mathfrak{L} = \int_{\Omega\times\mathbb{R}^{K-1}} \ln ||\nabla_{\mathbf{r}} F_y(\mathbf{r})|| \mathrm{d}\mu_y \mathrm{d}\mu_{\mathbf{r}}, \tag{3}$$

where $y \in \Omega$. Measure μ_y corresponds to the distribution of the data y_t, and $\mu_{\mathbf{r}}$ is the invariant measure of $\mathbf{r}$ under the dynamics of $P(y_t)A$, which will be estimated through sampling.

Once the Lyapunov exponent $\mathfrak{L}$ is calculated, we can set the buffer length:

$$B = \frac{1}{\mathfrak{L}} \ln \left(\frac{\delta}{\delta_0} \right), \tag{4}$$

where $\delta = 10^{-3}$ is the error tolerance and $\delta_0 = 2$ is the maximum initial error for probability vectors.

3 Subsequence Sampling Procedure

We use the following sampling procedure to obtain the subsequences used to compute stochastic gradient estimates. In order to enforce the non-overlapping mixing-time constraint between adjacent subsequences, we sample them sequentially. This results in the following form for the probability of the minibatch $\widetilde{\mathcal{S}}$: $p(\widetilde{\mathcal{S}}) = \prod_{n=0}^{R-1} L/|\mathcal{S}_n|$, where $|\mathcal{S}_0| = T$, $|\mathcal{S}_n| = |\mathcal{S}_{n-1}| - (\nu + 2B + 2L) - L_{\text{overlap}}$. The quantity L_{overlap} is calculated as follows:

$$L^0_{\text{overlap}} = |\tau_n|,$$

$$L^T_{\text{overlap}} = |T - \tau_n|,$$

$$L^L_{\text{overlap}} = \min_{n'=1, \tau_{n'} < \tau_n}^{n-1} \{|\tau_n - \tau_{n'}|\} - L - B,$$

$$L^R_{\text{overlap}} = \min_{n'=1, \tau_{n'} > \tau_n}^{n-1} \{|\tau_n - \tau_{n'}|\} - L - B.$$

If $\min\{L^0_{\text{overlap}}, L^T_{\text{overlap}}, L^L_{\text{overlap}}, L^R_{\text{overlap}}\} \geq 2\nu + 3L + 3B$, the minimum number of observations required to fit an entire subsequence while respecting minimum gap ν, $L_{\text{overlap}} = 0$. Otherwise, L_{overlap} equals to the sum of all the above terms that are less than $2\nu + 3L + 3B$.

Since $T \gg L, B, \nu$, then $p(\widetilde{\mathcal{S}})$ provides the correct probability of the minibatch $\widetilde{\mathcal{S}}$.

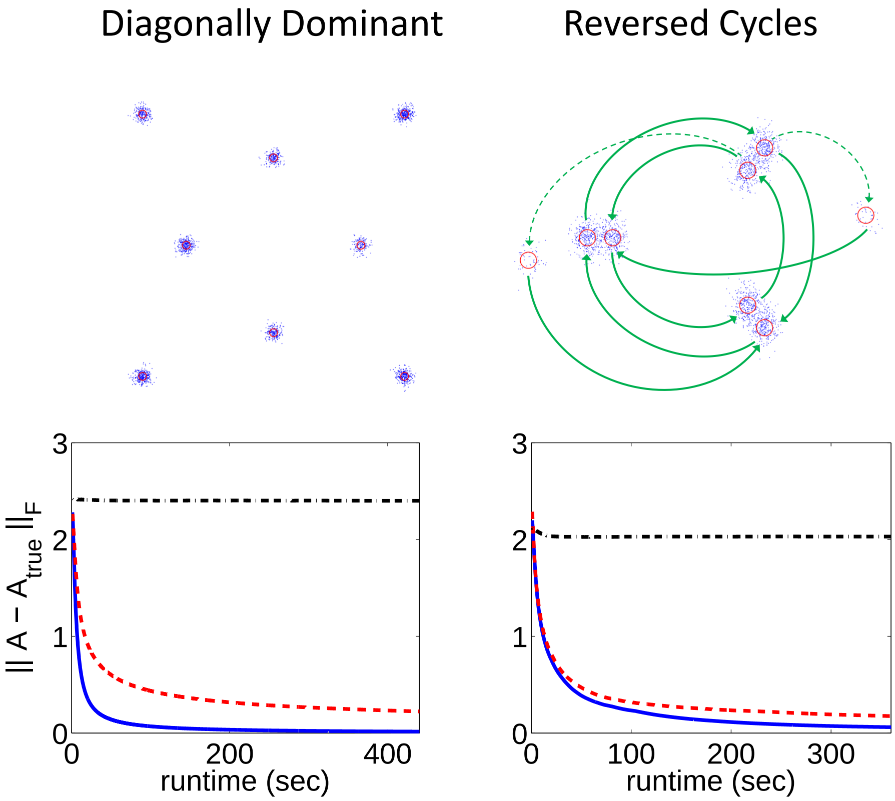

Figure 1: Synthetic experiments with hard-to-capture dynamics. Diagonally dominant (DD) (*left*) and reversed cycles (RC) (*right*) experiments. *First Row:* The emission distributions corresponding to 8 different states. Arrows in the RC case indicate the Markov transition structure with transition between bridge states as dashed arrows. *Second Row:* Decrease of error in transition matrix estimation versus runtime. Comparisons are made for SG-RLD algorithms with estimated buffer, without buffer, and treating data as i.i.d. All of the experiments use a constant computation budget by varying the number of subchains, $|\tilde{S}|$, with the length of the subchains, L.

4 Detailed Descriptions of Experiments

4.1 Evaluating Buffer Effectiveness

The first data set, *diagonally dominant* (DD) consists of a Markov chain that heavily self-transitions. Most subchains in a minibatch thus contain redundant information with observations generated from the same latent state. Although transitions are rarely observed, the emission means are set to be distinct so that this example is likelihood-dominated and highly identifiable. See Fig. 1 (top left). For this data we choose $L = 2$ and $|\widetilde{S}| = 10$ subsequences in order to incorporate observations from distant parts of the observation sequence. This corresponds to an extreme setting where each gradient is based only on 5 observations. The transition matrix and emission parameters used for this experiment were:

$$A_{DD} = \begin{pmatrix} .999 & .001 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .999 & .001 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .999 & .001 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .999 & .001 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .999 & .001 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .999 & .001 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .999 & .001 \\ .001 & 0 & 0 & 0 & 0 & 0 & 0 & .999 \end{pmatrix}.$$

$$\boldsymbol{\mu}_{DD} = \{(0, 20); (20, 0); (-30, -30); (30, -30); (-20, 0); (0, -20); (30, 30); (-30, 30); \}$$

and $\Sigma_{DD} = I$ for all states.

The second dataset we consider contains two *reversed cycles* (RC): the Markov chain strongly transitions from states $1 \to 2 \to 3 \to 1$ and $5 \to 7 \to 6 \to 5$ with a small probability of transiting between cycles via bridge states 4 and 8. See Fig. 1 (top right). The emission means for the two cycles are very similar but occur in reverse order with respect to the transitions. The emission variance is larger, making states 1 and 5, 2 and 6, 3 and 7 indiscernible by themselves. Transition information in observing long enough dynamics is thus crucial to identify between states $1, 2, 3$ and $5, 6, 7$. Therefore, we set $L = 5$ and $|\widetilde{S}| = 4$. Note that same amount of data are used in the calculation of the gradient. The transition matrix and emission parameters were:

$$A_{RC} = \begin{pmatrix} .01 & 0 & .85 & 0 & 0 & 0 & 0 & 1 \\ .99 & .01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .99 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .15 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & .01 & 0 & .85 & 0 \\ 0 & 0 & 0 & 0 & .99 & .01 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .99 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .15 & 0 \end{pmatrix}.$$

$$\boldsymbol{\mu} = \{(-50,0);(30,-30);(30,30);(-100,-10);(40,-40);(-65,0);(40,40);(100,10)\},$$

and $\Sigma_{RC} = 20 * I$ for all states.

We use a non-conjugate flat prior to demonstrate the flexibility of our algorithm. We initialize with a short run of k-means clustering to ensure that different states have different emission parameters.

4.2 Non-conjugate Emission Distribution

For the non-conjugate experiment, we used the following transition matrix:

$$\begin{pmatrix} .1 & .9 \\ .9 & .1 \end{pmatrix}.$$

For emission probability, we use a log-normal distribution: $p_k(y) \propto e^{-\dfrac{\ln(y-\mu_k)^2}{2\sigma_k^2}}$ with parameters: $\mu_1 = 0$, $\mu_2 = 4$; $\sigma_1 = \sigma_2 = 2$.

In the non-conjugate model, we use the following priors on the emission parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2 \sim \mathcal{N}(0,1)$.