Appendix: Scalable Inference for Hybrid Bayesian Hidden Markov Model Using Gaussian Process Emission

Yohan Jung
Department of Industrial & Systems Engineering, KAIST
and
Jinkyoo Park
Department of Industrial & Systems Engineering, KAIST

December 17, 2021

Derivation for equations

ELBO for Eq. (13)

$$
\begin{aligned}
&\log p(Y|X) \\
&= \iiint p(Y|X,Z,A,\pi)p(Z,A,\pi|X)\; dZdAd\pi \\
&\geq \iiint \log\Big(p(Y|X,Z,A,\pi)\frac{p(Z,A,\pi)}{q(Z,A,\pi)}\Big)q(Z,A,\pi)\; dZdAd\pi \\
&= \iiint \log p(Y|X,Z,A,\pi)q(Z,A,\pi)\; dZdAd\pi - KL(q(Z,A,\pi)||p(Z,A,\pi)) \\
&= \iiint \log p(Y,Z|X,A,\pi)q(Z,A,\pi)\; dZdAd\pi + H(q(Z),p(Z)) - KL(q(Z,A,\pi)||p(Z,A,\pi)) \\
&= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] + H(q(Z)) - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - KL(q(A,\pi)||p(A,\pi)) \coloneqq \mathcal{L}
\end{aligned}
$$

where cross entropy $H(q(Z),p(Z)) = H(q(Z)) + KL(q(Z)||p(Z))$.

Batch factors for Eq. (26)

Given the sampled $i \in \{1, .., T-L+1\}$ uniformly, let $Y_L^s = \{y_i, .., y_{i+L-1}\}$ be sampled observation with the length L and X_L^s be corresponding inputs and Z_L^s corresponding hidden states. The expected log joint likelihood of $\{Z_L^s, Y_L^s\}$ given X_L^s is approximated as

$$
\begin{aligned}
&\mathbb{E}_s\Big[\mathbb{E}_q[\log p(Y_L^s, Z_L^s|X_L^s)]\Big] \\
&= \sum_{i=0}^{T-L} \frac{1}{T-L+1} \mathbb{E}_q[\log p(Y_L^{s_i}, Z_L^{s_i}|X_L^{s_i})] \\
&= \frac{1}{T-L+1} \sum_{i=0}^{T-L} \mathbb{E}_q \left[\log p(z_i) + \underbrace{\sum_{t=1}^{L} \log p(A_{z_{i+t-1}, z_{i+t}})}_{\text{transition term}} + \underbrace{\sum_{t=1}^{L} \log p(y_{i+t}|z_{i+t}, x_{i+t})}_{\text{observation term}} \right] \cdots (*) \\
&\approx \frac{1}{T-L+1} \mathbb{E}_q \left[\sum_{t=1}^{T-L+1} \log p(z_{t-1}) + L \sum_{t=1}^{T} \log p(A_{z_{t-1}, z_t}) + L \sum_{t=1}^{T} \log p(y_t|z_t, x_t) \right]
\end{aligned}
$$

This implies that transition and observation term of $\mathbb{E}_q[\log p(Y_L^s, Z_L^s|X_L^s)]$ for the sampled $i \in \{1, .., T-L+1\}$ can be approximated as

$$
\begin{aligned}
\mathbb{E}_q \left[\sum_{t=1}^{L} \log p(A_{z_{i+t-1}, z_{i+t}}) \right] &\approx \frac{L}{T-L+1} \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(A_{z_{t-1}, z_t}) \right] \\
\mathbb{E}_q \left[\sum_{t=1}^{L} \log p(y_{i+t}|z_{i+t}, x_{i+t}) \right] &\approx \frac{L}{T-L+1} \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(y_t|z_t, x_t) \right]
\end{aligned}
$$

Thus, the batch factors, C_s^A and C_s^θ, to calibrate the approximated ELBO are obtained as

$$
C_s^A = \frac{T-L+1}{L} \;\; , \;\; C_s^\theta = \frac{T-L+1}{L}
$$

The transition term in expectation in $(*)$ can be approximated as

$$
\begin{aligned}
&\sum_{i=0}^{T-L} \mathbb{E}_q \left[\sum_{t=1}^{L} \log p(A_{z_{i+t-1}, z_{i+t}}) \right] \\
&= \mathbb{E}_q \left[\sum_{j=1}^{L} \log p(A_{z_{j-1}, z_j}) + \sum_{j=2}^{L+1} \log p(A_{z_{j-1}, z_j}) + \cdots + \sum_{j=T-L+1}^{T} \log p(A_{z_{j-1}, z_j}) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(A_{z_{t-1}, z_t}) + \underbrace{\sum_{t=1}^{L-1} t \left(\log p(A_{z_{t-1}, z_t}) + \log p(A_{z_{T-t}, z_{T-t+1}}) \right)}_{\text{approximated term}} \right] \\
&\approx \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(A_{z_{t-1}, z_t}) + L \sum_{t=1}^{L-1} \left(\log p(A_{z_{t-1}, z_t}) + \log p(A_{z_{T-t}, z_{T-t+1}}) \right) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=1}^{T} \log p(A_{z_{t-1}, z_t}) \right]
\end{aligned}
$$

Here, the observation term in expectation in $(*)$ can be approximated as

$$
\begin{aligned}
&\sum_{i=0}^{T-L} \mathbb{E}_q \left[\sum_{t=1}^{L} \log p(y_{i+t}|z_{i+t}, x_{i+t}) \right] \\
&= \mathbb{E}_q \left[\sum_{j=1}^{L} \log p(y_j|z_j, x_j) + \cdots + \sum_{j=T-L+1}^{T} \log p(y_j|z_j, x_j) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(y_t|z_t, x_t) + \underbrace{\sum_{t=1}^{L-1} t \left(\log p(y_t|z_t, x_t) + \log p(y_{T-t+1}|z_{T-t+1}, x_{T-t+1}) \right)}_{\text{approximated term}} \right] \\
&\approx \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(y_t|z_t, x_t) + L \sum_{t=1}^{L-1} \left(\log p(y_t|z_t, x_t) + \log p(y_{T-t+1}|z_{T-t+1}, x_{T-t+1}) \right) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=1}^{T} \log p(y_t|z_t, x_t) \right]
\end{aligned}
$$

SM kernel Approximation for Eq. (31)

Given the parameters of SM kernel $\{w_q, \mu_q, \sigma_q\}_{q=1}^Q$, we sample spectral points $\boldsymbol{s}_q = \{s_{q,i}\}_{i=1}^m$ from Gaussian distribution $N(S; \mu_q, \sigma_q)$ by reparametrization trick as

$$s_{q,i} = \mu_q + \sigma_q \circ \epsilon_i,$$

where $\epsilon_i \sim N(\epsilon; 0, I)$ for $i = 1, .., m$. If we define the feature map $\phi_{\boldsymbol{s}_q}(x)$ as

$$\phi_{\boldsymbol{s}_q}(x) = \frac{1}{\sqrt{m}} [\cos 2\pi s_{q,1}, \sin 2\pi s_{q,1}, .., \cos 2\pi s_{q,m}, \sin 2\pi s_{q,m}] \in R^{1\times 2m},$$

then $\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T$ can approximate $k_q(x-y)$ which is the inducted kernel from Gaussian Spectral density $N(S; \mu_q, \sigma_q)$ by Bochner's theorem.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \left(\cos 2\pi {s_{q,i}}^T x\right)\left(\cos 2\pi {s_{q,i}}^T y\right) + \left(\sin 2\pi {s_{q,i}}^T x\right)\left(\sin 2\pi {s_{q,i}}^T y\right)\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \cos 2\pi {s_{q,i}}^T (x-y)\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \frac{e^{i2\pi {s_{q,i}}^T (x-y)} + e^{-i2\pi {s_{q,i}}^T (x-y)}}{2}\right] \\
&= \frac{1}{2}\left(k_q(x-y) + k_q(y-x)\right) = k_q(x-y).
\end{aligned}$$

Using the above derivation, if we define sampled spectral points $\boldsymbol{s} = \cup_{q=1}^Q \{s_{q,i}\}_{i=1}^m$ with $s_{q,i} \sim N(\mu_q, \sigma_q^2)$ and the feature map $\phi^{SM}(x) = \left[\sqrt{w_1}\phi_{\{s_{1,i}\}_{i=1}^m}(x), .., \sqrt{w_Q}\phi_{\{s_{Q,i}\}_{i=1}^m}(x)\right]$, then $\phi^{SM}(x)\phi^{SM}(y)^T$ is an unbiased estimator of $k_{SM}(x,y)$ as

$$\mathbb{E}_{\boldsymbol{s}}\left[\phi^{SM}(x)\phi^{SM}(y)^T\right] = \sum_{q=1}^Q w_q \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)}\left[\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T\right] = \sum_{q=1}^Q w_q k_q(x-y) = k_{SM}(x-y).$$

Regularized Lower bound for Eq. (32)

Let $q(s)$ be variational distribution defined in Eq. (30). We can derive the lower bound $\mathcal{L}$ as follows:

$$\begin{aligned}
\log p(Y|X) &= \log \int p(Y, s|X) ds = \log \int p(Y|X,s)\frac{p(s)}{q(s)} q(s) ds \\
&\geq \int \log\left(p(Y|X,s)\frac{p(s)}{q(s)}\right) q(s) ds \\
&= \int \log p(Y|X,s) q(s) ds - KL(q(s)||p(s)) \\
&\approx \frac{1}{J}\sum_{j=1}^J \log p(Y|X, s^{(j)}) - KL(q(s)||p(s)),
\end{aligned}$$

where $s^{(j)}$ is j-th sampled spectral points from $q(s)$.

ELBO for Eq. (33)

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - KL(q(A,\pi)||p(A,\pi)) \\
&= \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\log p(y_t|x_t,z_t)\right] - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\frac{1}{J}\sum_{j=1}^{J}\log p(y_t|z_t,x_t,s^{(j)})\right] \\
&\qquad - \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}KL(q(S|z_t)||p(S|z_t))\right] - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\frac{1}{J}\sum_{j=1}^{J}\log p(y_t|z_t,x_t,s^{(j)})\right] \\
&\qquad - T\sum_{k=1}^{K}KL(q(S|z_t=k)||p(S|z_t=k)) - KL(q(A,\pi)||p(A,\pi)).
\end{aligned}
$$

The first inequality holds by replacing the derived lower bound of $\log p(y_t|x_t,z_t)$ in Eq. (32) with $\log p(y_t|x_t,z_t)$ in Eq. (14). The second inequality is derived using $q(z_t=k)\leq 1$ for all t,k.

As we express the above terms via the expected joint likelihood $\mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right]$ represented as

$$
\mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\frac{1}{J}\sum_{j=1}^{J}\log p(y_t|z_t,x_t,s^{(j)})\right],
$$

that uses the approximate GP emission with its likelihood $\frac{1}{J}\sum_{j=1}^{J}\log p(y_t|z_t,x_t,s^{(j)})$ instead of the exact GP emission model with its likelihood $\log p(y_t|z_t,x_t)$, we obtain the following equivalent terms:

$$
= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - T\sum_{k=1}^{K}KL(q(s|z_t=k)||p(s|z_t=k)) - KL(q(A,\pi)||p(A,\pi)).
$$

Since the additional KL divergence term $KL(q(s|z_t=k)||p(s|z_t=k))$ of the hidden state $z_t=k$ with the trainable parameters $\{\mu_q,\sigma_q\}_{q=1}^{Q}$ and fixed parameters $\bigcup_{i=1}^{m}\{\mu_{q,i},\sigma_{q,i}\}_{q=1}^{Q}$ for prior distribution, is reduced as

$$
KL(q(s|z_t=k)||p(s|z_t=k)) = \sum_{q=1}^{Q}\sum_{i=1}^{m}\mathrm{KL}(N(u_q,\sigma_q^2)||N(\widetilde{\mu}_{q,i},\widetilde{\sigma}_{q,i}^2)) = \sum_{q=1}^{Q}\mathrm{KL}(N(u_q,\sigma_q^2)||N(\widetilde{\mu}_{q,1},\widetilde{\sigma}_{q,1}^2)),
$$

where the last equality holds if we set $\{\widetilde{\mu}_{q.i},\widetilde{\sigma}_{q,i}^2\}_{i=2}^{M}=\{\mu_q,\sigma_q\}$, we obtain the following form $\mathcal{L}_{asm}$ by using the re-expressed KL terms above

$$
= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - T\sum_{k=1}^{K}\sum_{q=1}^{Q}\mathrm{KL}(N(u_q,\sigma_q^2)\,||\,N(\widetilde{\mu}_{q,1},\widetilde{\sigma}_{q,1}^2)) - KL(q(A,\pi)||p(A,\pi)) \coloneqq \mathcal{L}_{asm}.
$$

Approximate GP Emission for Eq. (34)

For the corresponding approximate GP emission, we can obtain $\prod_{j=1}^{J}p(y_t|z_t,x_t,s^{(j)})^{\frac{1}{j}}$ because of

$$
\log p(y_t|z_t,x_t) \approx \frac{1}{J}\sum_{j=1}^{J}\log p(y_t|z_t,x_t,s^{(j)}) = \log\prod_{j=1}^{J}p(y_t|z_t,x_t,s^{(j)})^{\frac{1}{j}}.
$$