A Variational Sequential Monte Carlo – Supplementary Material

A.1 Proof of Proposition 1

We start by noting that the distribution of all random variables generated by the VSMC algorithm is given by

$$\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}, b_T\,; \lambda) = \frac{w_T^{b_T}}{\sum_\ell w_T^\ell} \prod_{i=1}^N r(x_1^i\,; \lambda) \cdot \prod_{t=2}^T \prod_{i=1}^N \frac{w_{t-1}^{a_{t-1}^i}}{\sum_\ell w_{t-1}^\ell} r(x_t^i | x_{t-1}^{a_{t-1}^i}\,; \lambda). \tag{12}$$

We are interested in the marginal distribution $q(x_{1:T}\,;\lambda) \triangleq \widetilde{\phi}(x_{1:T}\,;\lambda) = \mathbb{E}_{b_{1:T}}[\widetilde{\phi}(x_{1:T}^{b_{1:T}}, b_{1:T}\,;\lambda)]$. A key observation is that the distribution of $b_{1:T} \mid x_{1:T}$, the conditional distribution of the ancestral path of the returned particle, is uniform on $\{1,\dots,N\}^T$. Thus we get

$$q(x_{1:T}\,;\lambda) = \frac{\widetilde{\phi}(x_{1:T}^{b_{1:T}}, b_{1:T}\,;\lambda)}{\widetilde{\phi}(b_{1:T} \mid x_{1:T}\,;\lambda)} = \frac{1}{N^{-T}} \sum_{a_{1:T-1}^{-b_{1:T-1}}} \int \widetilde{\phi}(x_{1:T}^{b_{1:T}}, x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)\, \mathrm{d}x_{1:T}^{\neg b_{1:T}}, \tag{13}$$

where

$$\begin{aligned}
&\frac{1}{N^{-T}} \widetilde{\phi}(x_{1:T}^{b_{1:T}}, x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda) \\
&= N^T \frac{w_1^{b_1}}{\sum_\ell w_1^\ell} r(x_1^{b_1}\,;\lambda) \prod_{t=2}^T \frac{w_t^{b_t}}{\sum_\ell w_t^\ell} r(x_t^{b_t} \mid x_{t-1}^{b_{t-1}}\,;\lambda) \cdot \prod_{\substack{i=1\\ i\neq b_1}}^N r(x_1^i\,;\lambda) \cdot \prod_{t=2}^T \prod_{\substack{i=1\\ i\neq b_t}}^N \frac{w_{t-1}^{a_{t-1}^i}}{\sum_\ell w_{t-1}^\ell} r(x_t^i | x_{t-1}^{a_{t-1}^i}\,;\lambda) \\
&= p(x_1^{b_1}, y_1) \prod_{t=2}^T \frac{p(x_{1:t}^{b_{1:t}}, y_{1:t})}{p(x_{1:t-1}^{b_{1:t-1}}, y_{1:t-1})} \prod_{t=1}^T \frac{1}{\frac{1}{N}\sum_\ell w_t^\ell} \cdot \prod_{\substack{i=1\\ i\neq b_1}}^N r(x_1^i\,;\lambda) \cdot \prod_{t=2}^T \prod_{\substack{i=1\\ i\neq b_t}}^N \frac{w_{t-1}^{a_{t-1}^i}}{\sum_\ell w_{t-1}^\ell} r(x_t^i | x_{t-1}^{a_{t-1}^i}\,;\lambda) \\
&= p(x_{1:T}^{b_{1:T}}, y_{1:T}) \prod_{t=1}^T \frac{1}{\frac{1}{N}\sum_\ell w_t^\ell} \cdot \widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda).
\end{aligned}$$

We insert the above expression in (13) and we get

$$\begin{aligned}
q(x_{1:T}\,;\lambda) &= p(x_{1:T}^{b_{1:T}}, y_{1:T}) \sum_{a_{1:T-1}^{\neg b_{1:T-1}}} \int \left(\prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^i \right)^{-1} \cdot \widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)\, \mathrm{d}x_{1:T}^{\neg b_{1:T}} \\
&= p(x_{1:T}^{b_{1:T}}, y_{1:T}) \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)} \left[\left(\prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^i \right)^{-1} \right].
\end{aligned} \tag{14}$$

□

A.2 Proof of Theorem 1

The evidence lower bound (ELBO), using the above result about the distribution of $q(x_{1:T}\,;\lambda)$, is given by

$$\begin{aligned}
\mathcal{L}(\lambda) &= \mathbb{E}_{q(x_{1:T}\,;\lambda)} \left[\log p(x_{1:T}, y_{1:T}) - \log q(x_{1:T}\,;\lambda)\right] \\
&= -\int \Bigg\{ p(x_{1:T}^{b_{1:T}}, y_{1:T}) \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)} \left[\frac{1}{\prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^i} \right] \cdot \\
&\qquad \cdot \log \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)} \left[\frac{1}{\prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^i} \right] \Bigg\} \mathrm{d}x_{1:T}^{b_{1:T}}.
\end{aligned} \tag{15}$$

Note that $-t\log t$ is a concave function for $t > 0$, this means by the conditional Jensen's inequality we have $-\mathbb{E}[t]\log\mathbb{E}[t] \geq -\mathbb{E}[t\log t]$. If we apply this to (15) we get

$$\begin{aligned}\mathcal{L}(\lambda) &\geq \int \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}}\,;\lambda)}\left[\frac{p(x_{1:T}^{b_{1:T}}, y_{1:T})}{\prod_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N} w_t^i}\sum_{t=1}^{T}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_t^i\right)\right]\mathrm{d}x_{1:T}^{b_{1:T}} \\ &= \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda)}\left[\sum_{t=1}^{T}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_t^i\right)\right] = \widetilde{\mathcal{L}}(\lambda),\end{aligned}$$

where the last step follows because $q(x_{1:T}\,;\lambda)$ is the marginal of $\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda)$. $\square$

A.3 Stochastic Optimization

For the control variates we use

$$\sum_{t=2}^{T} c_t \mathbb{E}_{s(\cdot)\widetilde{\phi}(\cdot|\cdot\,;\lambda)}\left[\sum_{i=1}^{N}\nabla\log w_{t-1}^{a_{t-1}^i} - \sum_{\ell=1}^{N}\frac{w_{t-1}^{\ell}}{\sum_m w_{t-1}^m}\nabla\log w_{t-1}^{\ell}\right]$$

where

$$c_t = \mathbb{E}_{s(\cdot)\widetilde{\phi}(\cdot|\cdot\,;\lambda)}\left[\sum_{t'=t}^{T}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{t'}^i\right)\right].$$

In practice we use a stochastic estimate of c_t.

For $T = 2$ we can use a leave-one-out estimator of the ancestor variable score function gradient

$$\sum_{i=1}^{N}\mathbb{E}_{s(\cdot)\widetilde{\phi}(\cdot|\cdot\,;\lambda)}\left[\log\left(\frac{N-1}{N}\frac{\sum_{\ell=1}^{N} w_2^{\ell}}{\sum_{j\neq i} w_2^j}\right)\left(\nabla\log w_1^{a_1^i} - \sum_{\ell=1}^{N}\frac{w_1^{\ell}}{\sum_m w_1^m}\nabla\log w_1^{\ell}\right)\right].$$

Score Function Gradient Below we provide the derivation of a score function-like estimator that is applicable in very general settings. However, we have found that in practice the variance tends to be quite high.

$$\begin{aligned}\nabla\widetilde{\mathcal{L}}(\lambda) &= \nabla\mathbb{E}_{\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda)}\left[\log\widehat{p}(y_{1:T})\right] \\ &= \mathbb{E}_{\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda)}\left[\nabla\log\widehat{p}(y_{1:T}) + \log\widehat{p}(y_{1:T})\nabla\log\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda)\right],\end{aligned}$$

with

$$\nabla\log\widehat{p}(y_{1:T}) = \nabla\sum_{t=1}^{T}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_t^i\right) = \sum_{t=1}^{T}\sum_{i=1}^{N}\frac{w_t^i}{\sum_{\ell} w_t^{\ell}}\nabla\log w_t^i,$$

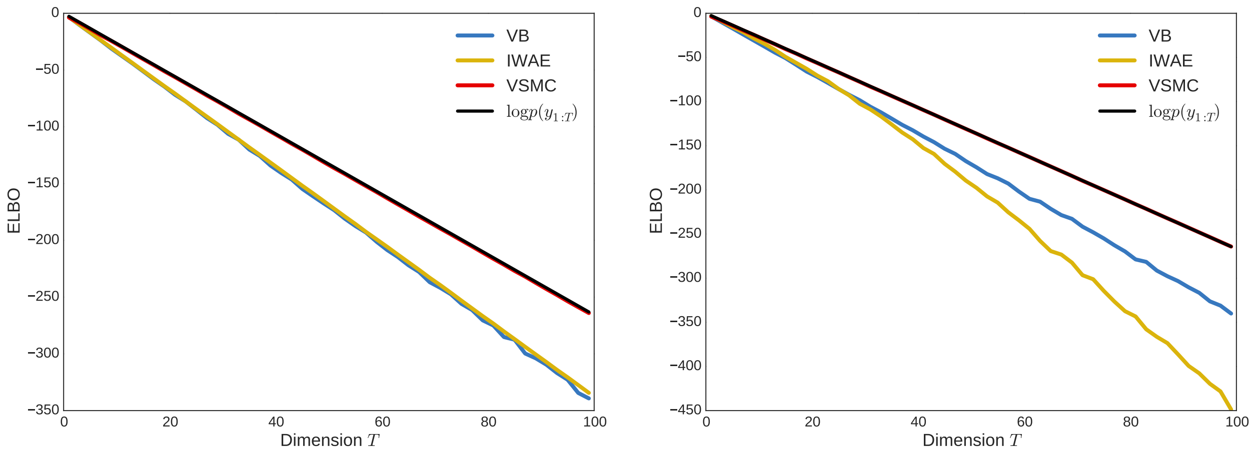
and

$$\begin{aligned}&\nabla\log\widetilde{\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}\,;\lambda) \\ &= \sum_{i=1}^{N}\left[\nabla\log r(x_1^i\,;\lambda) + \sum_{t=2}^{T}\left[\nabla\log r(x_t^i|x_{t-1}^{a_{t-1}^i}\,;\lambda) + \nabla\log w_{t-1}^{a_{t-1}^i} - \sum_{\ell=1}^{N}\bar{w}_{t-1}^{\ell}\nabla\log w_{t-1}^{\ell}\right]\right].\end{aligned}$$

A.4 Scaling With Dimension

In this section we study how the methods compare on a simple toy model defined by

$$p(x_{1:T}, y_{1:T}) = \prod_{t=1}^{T}\mathcal{N}(x_t\,;0,1)\mathcal{N}(y_t\,;x_t^2,1).$$



(a) $\lambda_{\text{IWAE}} = \lambda^{\star}_{\text{VB}}$ (b) $\lambda_{\text{IWAE}} = \lambda^{\star}_{\text{VSMC}}$

Figure 6: ELBO, for standard VB, IWAE, and VSMC, as a function of the dimension T of a toy problem. Here we set the number of samples in IWAE and VSMC to be $N = 2T$.

We study the data set $y_t = 3, \forall t$. Figure 6 shows the result when we let the number of samples in importance weighted auto-encoder (IWAE) (variational importance sampling (VIS)) and VSMC grow with the dimension $N = 2T$. For low T the optimal parameters for IWAE are close to $\lambda^{\star}_{\text{VSMC}}$. On the other hand for high T, the optimal parameters for IWAE are close to those of standard variational Bayes (VB), i.e. $\lambda^{\star}_{\text{VB}}$. Figure 6 indicates that just by letting $N \propto T$, VSMC can achieve arbitrarily good approximation of $p(x_{1:T} \mid y_{1:T})$ even if $T \to \infty$. This holds, under some regularity conditions, even if $p(x_{1:T}, y_{1:T})$ is a state space model [Bérard et al., 2014]. This asymptotic approximation property is not satisfied by VIS, we see in Figure 6 that the approximation deteriorates as T increases. Note that this does not hold if the dimension of the latent space, i.e. $\dim(x_t)$, tends to infinity rather than the number of time points T.