1 Variational Sequential Monte Carlo

Naesseth et al. (2018) introduce variational sequential Monte Carlo (VSMC), a variational family for approximating the posterior distribution of any sequence of variables. VSMC combines both variational inference and standard SMC, and can approximate the posterior arbitrarily well with more computation. Naesseth et al. (2018) derive a variational lower bound and present a stochastic gradient descent algorithm for optimizing the parameters. Naesseth et al. (2018) also make a connection to the importance weighted auto-encoder (IWAE) and show that the IWAE lower bound is a special case of the VSMC bound.

1.1 Background

Let $p(x_{1:t}, y_{1:t})$ be a sequence of probabilistic models for latent $x_{1:t}$ and observed $y_{1:t}$ with $t = 1, ..., T$. The posterior distribution of interest is $p(x_{1:T}|y_{1:T})$. Assume that the joint density factorizes as

$$p(x_{1:T}, y_{1:T}) = f(x_1) \prod_{t=2}^{T} f(x_t|x_{t-1}) \prod_{t=1}^{T} g(y_t|x_t)$$

where f is a prior for x and g is the distribution of y_t given x_t.

In VI, the divergence between a variational family $q_\lambda(x_{1:T})$ with variational parameters λ and the posterior is minimized so thaat $q_\lambda(x_{1:T}) \approx p(x_{1:T}|y_{1:T})$. For the KL divergence, this minimization is equivalent to maximizing the ELBO

$$\mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda}[\log p(x_{1:T}|y_T) - \log q_\lambda(x_{1:T})]$$

SMC approximates the posterior through weighted samples. Each sample is a sequence of particles that are sequentially drawn for $t = 1, ..., T$. At $t = 1$, standard importance sampling (IS) is used to sample $x_1^i \sim r(x_1)$ from some proposal r. For $t > 1$, auxiliary ancestor variables $a_{t-1}^i \in \{1, ..., N\}$ are resampled with probability proportional to the weights w_{t-1}^j. New values are then proposed and appended to each sequence, and the weights are updated. This procedure is summarized as follows:

1. Resample: $a_{t-1}^i \sim \text{Categorical}\left(\frac{w_{t-1}^j}{\sum_\ell w_{t-1}^\ell}\right)$
2. Propose: $x_t^i \sim r(x_t|x_{t-1}^{a_{t-1}^i})$
3. Append: $x_{1:t}^i = \left(x_{1:t-1}^{a_{t-1}^i}, x_t^i\right)$
4. Reweight: $w_t^i = \frac{f(x_t^i|x_{t-1}^{a_{t-1}^i}) g(y_t|x_t^i)}{r(x_t^i|x_{t-1}^{a_{t-1}^i})}$

The final samples $x_{1:T}^i$ are referred to as trajectories. The posterior is approximated by these weighted trajectories, i.e.,

$$p(x_{1:t}|y_{1:t}) \approx \hat{p}(x_{1:t}|y_{1:t}) = \sum_{i=1}^{N} \frac{w_t^i}{\sum_\ell w_t^\ell} \delta_{x_{1:t}^i}$$

where δ_X is the Dirac measure at X. As the number of trajectories N increase, the posterior approximation becomes arbitrarily accurate. Note that SMC yields an unbiased estimate of the marginal likelihood of $y_{1:T}$ given by

$$\hat{p}(y_{1:T}) = \prod_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} w_t^i$$

The proposal distribution $r(x_t|x_{t-1})$ is the main design choice. Bootstrap particle filter (BPF) uses the model prior f as the proposal.

1.2 VSMC

VSMC addresses the problem of choosing a proposal distribution in SMC by instead learning a parameterized proposal.

To sample from the VSMC family, SMC is run with the proposals parameterized by λ to obtain a set of trajectories, and then a single trajectory $x_{1:T}^{b_T}$ is sampled with probability proportional to its weight. The variational distribution $q_\lambda(x_{1:T})$ marginalizes out all auxiliary variables created in the sampling process. The joint distribution for all variables generated by VSMC is given by

$$\tilde{\phi}_\lambda(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}, b_T) = \prod_{i=1}^{N} \left[r_\lambda(x_1^i)\right] \prod_{t=2}^{T} \prod_{i=1}^{N} \left[\frac{w_{t-1}^{a_{t-1}^i}}{\sum_\ell w_{t-1}^\ell} r_\lambda(x_t^i | x_{t-1}^{a_{t-1}^i})\right] \left[\frac{w_T^{b_T}}{\sum_\ell w_T^\ell}\right]$$

Note that the data $y_{1:T}$ enter through the weights and optionally through the proposal distribution.

Let $b_t = a_t^{b_{t+1}}$ for $t \leq T-1$ denote the ancestors for the final trajectory $x_{1:T}^{b_T}$. Let $\neg b_{1:T}$ denote all indices not equal to $(b_1, ..., b_T)$. Then the marginal distribution of $x_{1:T} = x_{1:T}^{b_{1:T}} = (x_1^{b_1}, ..., x_T^{b_T})$ is given by

$$q_\lambda(x_{1:T}|y_{1:T}) = p(x_{1:T}, y_{1:T}) \mathbb{E}_{\tilde{\phi}_\lambda(x_{1:T}^{\neg b_{1:T}}, a_{1:T-1}^{\neg b_{1:T-1}})} \left[\hat{p}(y_{1:T})^{-1}\right]$$

MCMC estimates of the expectation lead to biased estimates of $\log q_\lambda(x_{1:T}|y_{1:T})$ and the ELBO. Instead, a tractable lower bound to the ELBO that can be stochastically optimized is used. The surrogate ELBO is the expected log marginal likelihood estimate given by

$$\begin{aligned}
\tilde{\mathcal{L}}(\lambda) &= \sum_{t=1}^{T} \mathbb{E}_{\tilde{\phi}_\lambda(x_{1:t}^{1:N}, a_{1:t-1}^{1:N})} \left[\log\left(\frac{1}{N}\sum_{i=1}^{N} w_t^i\right)\right] \\
&= \mathbb{E}[\log \hat{p}(y_{1:T})] \\
&\leq \mathcal{L}(\lambda) \\
&\leq \log p(y_{1:T})
\end{aligned}$$

The surrogate ELBO and its gradient can be estimated unbiasedly and stochastically using quantities produced during sampling from VSMC. It is assumed that the proposals $r_\lambda(x_t|x_{t-1})$ are reparameterizable in terms of some distribution s that is not a function of λ, i.e., $x_t = h_\lambda(x_{t-1}, \epsilon_t)$ and $\epsilon_t \sim s(\epsilon_t)$. The gradient of the surrogate ELBO is then given by

$$\begin{aligned}
\nabla\tilde{\mathcal{L}}(\lambda) &= g_{rep} + g_{score} \\
g_{rep} &= \mathbb{E}[\nabla \log \hat{p}(y_{1:T})] \\
g_{score} &= \mathbb{E}\left[\log \hat{p}(y_{1:T}) \nabla \log \tilde{\phi}_\lambda(a_{1:T-1}^{1:N} | \epsilon_{1:T}^{1:N})\right]
\end{aligned}$$

Note that the ancestor variables are discrete and cannot be reparameterized. This may lead to high variance in the score term g_{score}. To lower the variance, Rao-Blackwellization (Robert & Roberts, 2021) can be applied by noting that the ancestor variables a_{t-1} have no effect on weights prior to time t. This leads to

$$g_{score} = \sum_{t=2}^{T} \mathbb{E}\left[\log \frac{\hat{p}(y_{1:T})}{\hat{p}(y_{1:t-1})} \left(\sum_{i=1}^{N} \nabla \log \frac{w_{t-1}^{a_{t-1}^i}}{\sum_\ell w_{t-1}^\ell}\right)\right]$$

The score function $\nabla \log \tilde{\phi}_\lambda(a_{1:T-1}^{1:N} | \epsilon_{1:T}^{1:N})$ with estimates of future log average weights is used as a control variate.

Naesseth et al. (2018) found that ignoring the score term g_{score} from the ancestor variables leads to faster convergence while retaining a good approximation of the ELBO. This leads to the approximation

$$\nabla\tilde{\mathcal{L}}(\lambda) \approx \mathbb{E}[\nabla \log \hat{p}(y_{1:T})] = g_{rep}$$

To optimize VSMC, the algorithm is as follows:

1. Estimate $\nabla\tilde{\mathcal{L}}(\lambda)$ using a single sample from $s(\cdot)\tilde{\phi}_\lambda(\cdot|\cdot)$ which is obtained as a byproduct from sampling VSMC.

2. Compute the step-size. Naesseth et al. (2018) use Adam given by

$$\rho^n = \eta n^{-1/2+\delta}(1+\sqrt{s^n})^{-1}$$
$$s^n = t\left(\hat{\nabla}\tilde{\mathcal{L}}(\lambda^n)\right)^2 + (1-t)s^{n-1}$$

   for iteration n, $\delta = 10^{-16}$, $t = 0.1$, and various values for η.

3. Update the variational parameters

$$\lambda^{n+1} = \lambda^n + \rho^n\hat{\nabla}\tilde{\mathcal{L}}(\lambda^n)$$

If the target distribution $p_\theta(x_{1:T}|y_{1:T})$ depends on a set of unknown parameters θ, the parameters can be fit using variational EM. The surrogate ELBO is then

$$\log p_\theta(y_{1:T}) \geq \tilde{\mathcal{L}}(\lambda, \theta)$$

where the normalization constant $p_\theta(y_{1:T})$ is now a function of θ. $\tilde{\mathcal{L}}(\lambda, \theta)$ can be optimized with respect to both θ and λ using stochastic optimization.

1.3 Perspectives of VSMC

- When $N = 1$, VSMC reduces to a structured variational approximation (no resampling and so the variational distribution is the proposal).
- When $T = 1$, VSMC reduces to "variational" importance sampling. The surrogate ELBO for VIS is equal to the IWAE lower bound.
- If $\log\hat{p}(y_{1:T})$ is uniformly integrable, then as $N \to \infty$, $\tilde{\mathcal{L}}(\lambda) = \mathcal{L}(\lambda) = \log p(y_{1:T})$. A bound on the KL is given by

$$KL\left(q_\lambda(x_{1:T})\|p(x_{1:T}|y_{1:T})\right) \leq \frac{c(\lambda)}{N}$$

  for some constant $c(\lambda) < \infty$.

- VSMC scales well with T. Assuming that $\log\hat{p}(y_{1:T})$ is uniformly integrable, it can be shown that by taking $N = bT$, $b > 0$,

$$KL\left(q_\lambda(x_{1:T})\|p(x_{1:T}|y_{1:T})\right) \leq -\mathbb{E}\left[\log\frac{\hat{p}(y_{1:T})}{p(y_{1:T})}\right] \to \frac{\sigma^2(\lambda)}{2b}$$

  where $\sigma^2(\lambda) \in (0, \infty)$ as $T \to \infty$.

1.4 Questions

- How are the weights computed? See SMC.
- Why do MCMC estimates of the expected inverse normalization constant lead to biased estimates of the log marginal density and the ELBO?
- How is the gradient of the surrogate ELBO derived?
- What are the expectations in the gradient of the surrogate ELBO with respect to? $\epsilon_{1:T}^{1:N}$? The ancestor variables?

References

Naesseth, C., Linderman, S., Ranganath, R., & Blei, D. (2018). Variational sequential monte carlo (A. Storkey & F. Perez-Cruz, Eds.). In A. Storkey & F. Perez-Cruz (Eds.), *Proceedings of the twenty-first international conference on artificial intelligence and statistics*. http://proceedings.mlr.press/v84/naesseth18a.html

Robert, C., & Roberts, G. (2021). *Rao-blackwellization in the mcmc era*.