Fully Reparameterized Variational Sequential Monte Carlo

Kenny Chiu

Abstract

TODO

1 Introduction

The paper by Naesseth et al. (2018) introduces Variational Sequential Monte Carlo (VSMC) as a flexible variational family for approximating the posterior distribution of a sequence of random variables. In this project, we discuss the main contributions and limitations of the paper. We also review VSMC and propose a fully reparameterized version of VSMC (which we refer to as FR-VSMC) aimed at addressing its problems of noisy gradient estimation. We evaluate our proposed extension through experiments similar to the ones in the original VSMC paper.

This report is organized as follows: Section 2 provides a review of the paper and VSMC; Section 3 describes our proposed modification to VSMC; Section 4 discusses the results of our experiments evaluating FR-VSMC; and Section 5 summarizes our main points and concludes with a discussion of what we have learned.

2 Variational Sequential Monte Carlo

The main contributions of the paper by Naesseth et al. (2018) include the introduction of VSMC variational family, as well as the derivation of a tractable bound for optimizing VSMC. Following (Naesseth et al., 2018), we provide an overview of these ideas in the context of a state space model. Let $x_{1:T}$ and $y_{1:T}$ denote sequences of T latent variables and T observations, respectively, and assume that their joint distribution factorizes as

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

The target distribution is then the posterior distribution $p(x_{1:T}|y_{1:T})$.

2.1 Sequential Monte Carlo

VSMC is heavily based on Sequential Monte Carlo (SMC). SMC is a MCMC method that approximates the posterior distribution of a sequence of random variables using a set of N weighted particles. SMC constructs the set of particles through the following procedure:

1. Initialize N samples $x_1^{1:N}$ from some proposal distribution $q(x_1)$ and assign a weight $w_1^i = \frac{p(x_1^i)p(y_1^i|x_1^i)}{q(x_1^i)}$ to each sample x_1^i, $i = 1, ..., N$.

2. For iterations $t = 2, ..., T$, do the following three steps:

   (a) **Resample**: sample *ancestor* variables $a_t^i \in \{1, ..., N\}$, $i = 1, ..., N$, from a CATEGORICAL distribution with probabilities proportional to $w_{t-1}^{1:N}$.

   (b) **Propose**: propose new states x_t^i, $i = 1, ..., N$, according to some proposal distribution $q(x_t^i|x_{t-1}^{a_t^i})$.

   (c) **Reweight**: assign new weights $w_t^i = \frac{p(x_t^i|x_{t-1}^{a_t^i})p(y_t^i|x_t^i)}{q(x_t^i|x_{t-1}^{a_t^i})}$, $i = 1, ..., N$, to each new state x_t^i.

Let $x_{1:T}^i$, $i = 1, ..., N$ denote the sequence of states that gave rise to x_T^i, i.e.,

$$x_{1:T}^i = \left(x_1^{\cdots}, \ldots, x_{T-2}^{a_{T-1}^{\cdots}}, x_{T-1}^{a_T^i}, x_T^i \right) .$$

Then the above procedure returns the set of particles $x_{1:T}^{1:N}$ along with weights $w_T^{1:N}$. SMC then approximates the posterior distribution with the discrete measure

$$p(x_{1:T}|y_{1:T}) \approx q(x_{1:T}|y_{1:T}) = \sum_{i=1}^{N} \frac{w_T^i}{\sum_{\ell=1}^{N} w_T^\ell} \delta_{x_{1:T}^i}$$

where δ_x is the Dirac measure at x. Notice that the approximation $q(x_{1:T}|y_{1:T}) \to p(x_{1:T}|y_{1:T})$ as $N \to \infty$. Also note that while we cannot evaluate the density of the approximation, we can sample from it by sampling a particle $x_{1:T}^i$ with probability w_T^i.

2.2 VSMC Objective

The key design step of SMC is choosing the proposal distribution q. Instead of specifying a distribution, VSMC postulates a parametric variational family q_λ with variational parameters λ for the proposal and learns the optimal proposal that minimizes

$$\mathrm{KL}\left(q_\lambda(x_{1:T}|y_{1:T}) \| p(x_{1:T}|y_{1:T})\right) .$$

This objective and the corresponding evidence lower bound (ELBO) is intractable, however, due to the intractable posterior density of q_λ. Naesseth et al. (2018) derive a *surrogate* ELBO given by

$$\tilde{\mathcal{L}}(\lambda) = \mathbb{E}_{x_{1:T}^{1:N}, a_{2:T}^{1:N}} \left[\log \hat{p}(y_{1:T})\right]$$

where $\hat{p}(y_{1:T})$ is an unbiased estimator of $p(y_{1:T})$ given by

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

The surrogate ELBO is a lower bound to the ELBO and can be optimized stochastically using samples obtained from running SMC. Assuming that the proposal $q_\lambda(x_t|x_{t-1})$ can be reparameterized in terms of some noise distribution $p(\varepsilon_t)$ independent of the parameters λ, the gradient of the surrogate ELBO can then be rewritten as

$$\begin{aligned}\nabla \tilde{\mathcal{L}}(\lambda) &= g_{\text{rep}} + g_{\text{score}} \\ g_{\text{rep}} &= \mathbb{E}_{\varepsilon_{1:T}^{1:N}, a_{2:T}^{1:N}} \left[\nabla \log \hat{p}(y_{1:T})\right] \\ g_{\text{score}} &= \mathbb{E}_{\varepsilon_{1:T}^{1:N}} \left[\log \hat{p}(y_{1:T}) \nabla \log p_\lambda(a_{2:T}^{1:N} | \varepsilon_{1:T}^{1:N})\right] \ .\end{aligned}$$

The g_{score} component is dependent on the CATEGORICAL distribution of the ancestors, which in turn is dependent on λ through the weights $w_{1:T}^{1:N}$ and cannot be reparameterized. Naesseth et al. (2018) found that the stochastic estimator for g_{score} had a much larger variance compared to that of g_{rep}, which impacts the convergence rate of the stochastic optimization. Naesseth et al. (2018) proposed several strategies to reduce the variance in practice, such as ignoring g_{score} (resulting in a biased gradient estimator), Rao-Blackwellization (Robert & Roberts, 2021), and using control variates. Overall, the noisy unbiased gradient estimator due to the non-reparameterizability of the ancestor variables is the main limitation of VSMC. We aim to address this issue with our proposed extension in Section 3.

2.3 Other Contributions and Limitations

We conclude this section with a brief mention of the other contributions and limitations of the paper. Naesseth et al. (2018) also made a connection between VSMC and Importance Weighted Autoencoders (Burda et al., 2016) (IWAE) where the surrogate ELBO is exactly the IWAE lower bound when $T = 1$. This leads to a reinterpretation of IWAE as *variational importance sampling* and extends known results of IWAE to special cases of VSMC (when $T = 1$), namely, that the surrogate ELBO is tighter than that of standard Variational Bayes (when $N = 1$).

One notable limitation of the paper is that the theory and experiments are all presented in the context of state space model problems despite the claim that VSMC is applicable to any sequence of models. It would be interesting to see how VSMC translates to other contexts and if any properties observed in the state space model context do not, but we do not focus on this idea further and leave it as a possible direction of future work.

3 Approximation of the Categorical Distribution

3.1 Gumbel-Softmax Distribution

3.2 Invertible Gaussian Reparameterization

4 Experiments

5 Discussion

5.1 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.

References

Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). Importance weighted autoencoders.

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*.