# Fully Reparameterized Variational Sequential Monte Carlo

Kenny Chiu

**Abstract**

We review the main contributions and limitations of the original Variational Sequential Monte Carlo paper by Naesseth et al. (2018). We also propose a modification to the Variational Sequential Monte Carlo algorithm where we replace the Categorical resampling step with a continuous, reparameterizable approximation in order to reduce the variance of the noisy unbiased gradient estimator. Our empirical results suggest that using the Gumbel-Softmax approximation reduces the variance of the unbiased estimator and achieves a performance comparable to that of the biased gradient estimator.

## 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 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 our proposed modifications; 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 (SSM). 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.,

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

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 $\delta_x$ 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 $q_\lambda$ with variational parameters $\lambda$ 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 $q_\lambda$. 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 $\lambda$, 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 $g_{\text{score}}$ component is dependent on the Categorical distribution of the ancestors, which in turn is dependent on $\lambda$ through the weights $w_{1:T}^{1:N}$ and cannot be reparameterized. Naesseth et al. (2018) found that the stochastic estimator for $g_{\text{score}}$ had a much larger variance compared to that of $g_{\text{rep}}$, 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 $g_{\text{score}}$ (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 make 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 dwell on this point further and leave it as a possible direction of future work.

## 3 Reparameterizing the Ancestors

We propose a different strategy to address the large variance of the unbiased gradient estimator. Instead of resampling the ancestors from a Categorical distribution, we resample

them from a continuous, reparameterizable approximation with the idea being that reparameterizing the ancestors in terms of some noise distribution independent of the parameters will reduce the variance of the unbiased gradient.

We acknowledge that this idea is not entirely novel and has been explored in previous works. One VSMC variant that Lawson et al. (2018) explored used the Straight-Through Gumbel-Softmax estimator (Jang et al., 2017) to the Categorical when estimating the gradient, though the Categorical resampling was still used at run time. Though their empirical results suggest that this approach works well in practice, we suspect that the recovered variational parameters may not optimal due to a potential gap between the model being optimized and the model being used at run time. We instead propose to use the continuous resampling model for both when optimizing and when running VSMC.

To approximate an ancestor variable $a_t^i \in \mathbb{N}$, we rewrite the ancestor as a vector on the $(N-1)$-dimensional simplex, i.e., $a_t^i \in \Delta^{N-1} = \{(v_1, ..., v_N) | \sum_{i=1}^N v_i = 1, v_i \geq 0, i = 1, ..., N\}$. Ancestors sampled from the Categorical distribution would be in the form of one-hot vectors (one entry $v_i$ is 1 and all others 0). A continuous approximation relaxes this requirement and allows for any sample vectors that are on the simplex. Any approximating distribution that satisfies this relaxed requirement may be used, though the distribution should also be reparameterizable to be independent of the variational parameters as per our original motivation for this approach.

A modification to the SMC procedure is needed when using the continuous approximation as there is no longer an interpretation of a state having a single ancestor for general vectors on the simplex. This can be resolved by interpreting the ancestor vector as mixture weights and taking the ancestor state as the convex mixture of all $N$ states from the previous timestep, i.e.,

$$x_{t-1}^{a_t^i} = \sum_{\ell=1}^{N} a_t^i(\ell) x_{t-1}^{\ell}$$

where $a_t^i(\ell)$ denotes the $\ell$-th entry of the vector $a_t^i$. Note that this modification to VSMC implies that the particle construction procedure is no longer *true* SMC and that the VSMC objective becomes only approximate. However, we argue that as long as the ancestor vector is "one-hot enough", the difference coming from the approximation becomes negligible. This suggests that the approximating distribution should also be able to control the degree to which the sample vectors approximate one-hot vectors through some temperature parameter $\tau$. In the following sections, we discuss two candidate approximating distributions that satisfy all of our requirements above.

## 3.1 Gumbel-Softmax Distribution

The Gumbel-Softmax distribution (Jang et al., 2017; Maddison et al., 2017) was the approximating distribution explored by Lawson et al. (2018) and is a natural choice due to its direct translation from a Categorical distribution. To approximate a Categorical distribution with probabilities given by normalized weights $\tilde{w}_{t-1}^{1:N}$, the Gumbel-Softmax

distribution generates a sample vector $a_t^i$ through the transformation

$$a_t^i(j) = \frac{\exp\left(\frac{\log \tilde{w}_{t-1}^j + g_j^i}{\tau}\right)}{\sum_{\ell=1}^N \exp\left(\frac{\log \tilde{w}_{t-1}^\ell + g_\ell^i}{\tau}\right)} \qquad \text{for } j = 1, ..., N$$

where $g_1^i, ..., g_N^i$ are i.i.d. samples from a GUMBEL$(0, 1)$ distribution. This transformation reparameterizes the ancestors in terms of GUMBEL noise. Smaller values of $\tau$ lead to samples that more closely resemble one-hot vectors, and larger values encourage more mixing between the components.

## 3.2 Invertible Gaussian Reparameterization

The Invertible Gaussian Reparameterization (Potapczynski et al., 2020) (IGR) distribution is a more recent development for relaxing discrete distributions. The GUMBEL-SOFTMAX approximation is appealing as its parameters can be interpreted as the discrete distribution that is being approximated. IGR drops this property in exchange for more flexibility with the choice of the reparameterized distribution and the transformation function. For this project, we consider the IGR distribution as presented in (Potapczynski et al., 2020).

Let $\mathcal{S}^{N-1} = \{(v_1, ..., v_{N-1}) | \sum_{i=1}^{N-1} v_i < 1, v_i \geq 0, i = 1, ..., N-1\}$ denote an alternative representation of $\Delta^{N-1}$. To generate a sample $a_t^i \in \Delta^{N-1}$ that approximates a sample from a CATEGORICAL distribution with probability $\tilde{w}_{t-1}^{1:N}$, the IGR distribution generates a vector $\tilde{a}_t^i \in \mathcal{S}^{N-1}$ through the sequence of transformations

$$z_t^i = \mu_t + \text{diag}(\sigma_t)\varepsilon_t^i$$
$$\tilde{a}_t^i = g(z_t^i, \tau)$$

where $\varepsilon_t^i =$ is $\mathcal{N}(0, I_{N-1})$ noise, $\mu_t \in \mathbb{R}^{N-1}$, $\sigma_t \in (0, \infty)^{N-1}$, and $g$ is some invertible function with a tractable Jacobian. The vector $a_t^i$ is then recovered as

$$a_t^i = \left(\tilde{a}_t^i(1), \ldots, \tilde{a}_t^i(N-1), 1 - \sum_{\ell=1}^{N-1} \tilde{a}_t^i(\ell)\right) .$$

Potapczynski et al. (2020) provide several suggestions for $g$. For simplicity, we use their modified softmax function

$$g(z_t^i, \tau)(j) = \frac{\exp\left(\frac{z_t^i(j)}{\tau}\right)}{\sum_{\ell=1}^{N-1} \exp\left(\frac{z_t^i(\ell)}{\tau}\right) + \delta}$$

where $\delta > 0$. Like in GUMBEL-SOFTMAX, smaller values of $\tau$ discourage mixing.

An important detail to note is that because the parameters of the CATEGORICAL distribution do not directly translate over to IGR parameters, IGR requires solving the the optimization problem

$$(\mu_t, \sigma_t) = \arg\min_{(\mu,\sigma)} \mathbb{E}_{\tilde{a}_t}\left[\|\tilde{a}_t - w_{t-1}^{1:N-1}\|_2^2\right] .$$

In the context of VSMC, this optimization needs to be done before the resampling procedure in each time step of SMC. We solve this optimization problem using automatic differentiation, which we have found to be a major computational bottleneck in non-trivial problems. We suspect that there may a closed analytical form for the solution (at least for certain choices of the reparameterized distribution and the transformation) that would alleviate this limitation, but we do not pursue this further given the time frame of this project.

# 4 Experiments

We evaluate our proposed extension of VSMC on the same linear Gaussian SSM problems studied by Naesseth et al. (2018). These time series problems are useful for studying the properties of VSMC as the log marginal likelihood $\log p(y_{1:T})$ can be computed using the Kalman filter (Jong, 1988). This allows us to directly compare VSMC and our proposed extensions based on how well they can recover the marginal likelihood. The model for the linear Gaussian SSM is

$$\begin{aligned} x_t &= Ax_{t-1} + v_t \\ y_t &= Cx_t + e_t \end{aligned}$$

where $v_t \sim \mathcal{N}(0, Q)$, $e_t \sim \mathcal{N}(0, R)$, and $x_1 \sim \mathcal{N}(0, I)$.

In all of our experiments, we use $\tau = 0.05$ for the Gumbel-Softmax and IGR approximations unless otherwise specified. For IGR, we use estimate the expectation of the inner optimization using 100 samples and optimize the parameters using 100 iterations of stochastic gradient descent.

## 4.1 Scalar Linear Gaussian SSM

We consider the scalar linear Gaussian SSM problem with $A = 0.5$, $Q = 1$, $C = 1$, $R = 1$, and $T = 2$. Following Naesseth et al. (2018), we use the proposal $q_\lambda(x_t|x_{t-1}) = \mathcal{N}(\lambda + 0.5x_{t-1}, 1)$ with $\lambda \in \mathbb{R}$ and $x_0 = 0$. We use $N = 2$ particles to approximate the posterior.

**Variance of gradient estimator.** We first investigate whether our proposed extension reduces the variance of the unbiased surrogate ELBO gradient estimator. Figure 1 shows the mean and the standard deviation over 100 samples of each gradient estimator as $\lambda$ varies for a single generated dataset. The four gradient estimators are VSMC ($g_{\text{rep}}$), uVSMC ($g_{\text{rep}} + g_{\text{score}}$), GS-VSMC (VSMC with Gumbel-Softmax), and IGR-VSMC (VSMC with IGR). The observed behaviour of VSMC and uVSMC are similar to what was observed with $g_{\text{rep}}$ and $g_{\text{score}}$ in (Naesseth et al., 2018), though we note that the standard deviation of our estimators appears to be larger than what they observed (for $N = 2$). We were unable to find the source of this difference in spread, but we suspect it to be a numerical stability issue coming from automatic differentiation. The figure clearly shows that GS-VSMC and IGR-VSMC reduce the variance of the unbiased gradient estimator (uVSMC), and that the behaviour of GS-VSMC and IGR-VSMC are very similar to that

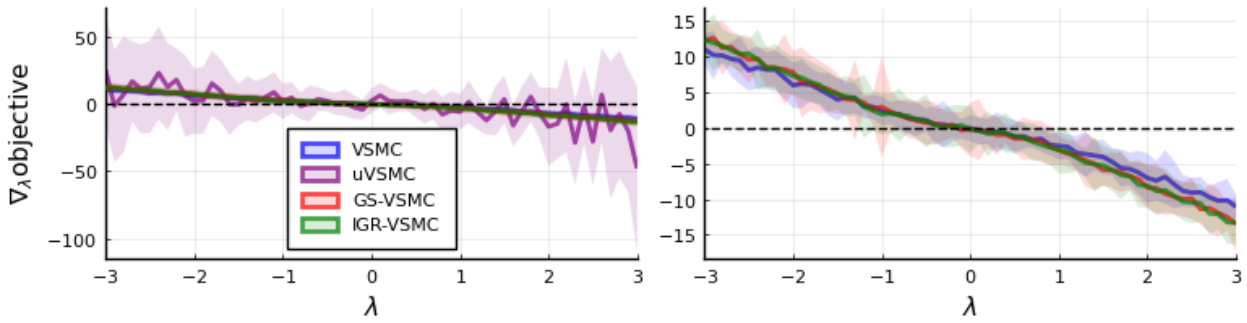


Figure 1: **(Left)** Mean and standard deviation over 100 samples of the gradient estimators for varying values of $\lambda$ for a single scalar linear Gaussian SSM problem. The number of particles is $N = 2$. The optimal value of $\lambda$ is the point at which the mean gradient is 0. **(Right)** The same figure but without UVSMC for better visibility.

of the biased gradient estimator (VSMC).

**Tightness of ELBO estimate.** We next evaluate how tight of a lower bound our proposed extension is able to estimate. Figure 2 shows the mean and standard deviation of the estimated surrogate ELBO across 100 runs. Following Naesseth et al. (2018), we try both the Adam step sequence and the adaptive $\rho$ step sequence (Kucukelbir et al., 2017) given by

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

where $k$ is the iteration number, $\delta = 10^{-16}$, $t = 0.1$, and $\eta = 0.1$. We find that all four VSMC variants are able to closely estimate the ELBO after a reasonable number of iterations, although VSMC with the unbiased gradient takes notably longer to converge. There is no notable difference between VSMC, GS-VSMC, and IGR-VSMC. We also note that the $\rho$ step sequence appears to work better than Adam and so we use the $\rho$ steps in our following experiments unless otherwise specified.

**Role of $\tau$.** We also investigate the role of the temperature $\tau$ specifically in GS-VSMC. Figure 3 shows the mean and standard deviation of the ELBO gradient estimate as well as the mean and standard deviation of the ELBO estimate for GS-VSMC with three levels of $\tau$. We observe that the standard deviation of the gradient estimator can become unstable for small $\tau$ (e.g., $\tau = 0.001$). Somewhat surprisingly however, the value of $\tau$ does not seem to greatly impact estimation of the ELBO. Even when the number of particles is large (e.g., $N = 50$), the impact of greater mixing on the ELBO estimate is marginal considering the relative error. Our results suggest that using a $\tau$ "small enough" (e.g., $\tau = 0.1$) is likely sufficient for reasonable estimation.

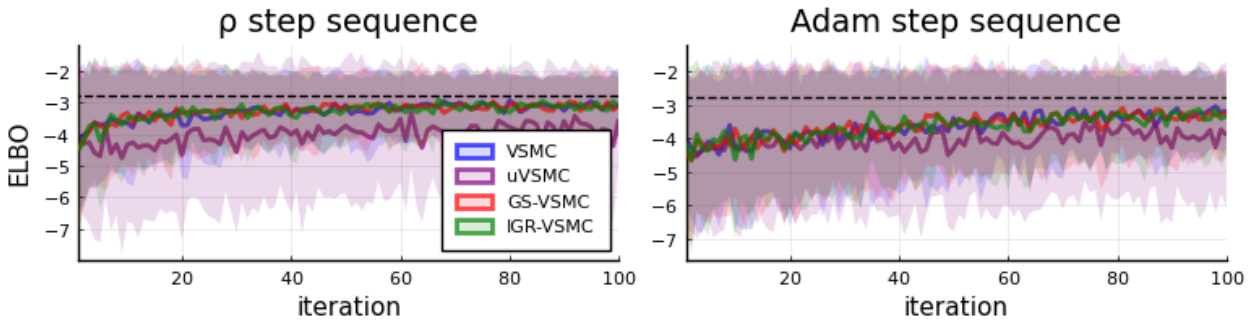


Figure 2: Mean and standard deviation of the estimated surrogate ELBO for a scalar linear Gaussian SSM problem across 100 runs. The black line is the log marginal likelihood $\log p(y_{1:T})$. The left figure shows using the $\rho$ step sequence and the right shows using the Adam step sequence.

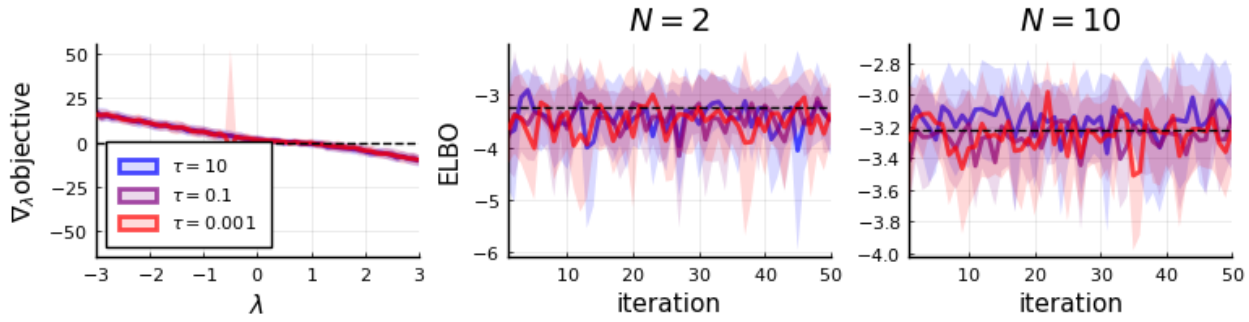


Figure 3: **(Left)** Mean and standard deviation over 100 gradient estimates of GS-VSMC with different values of $\tau$ for a scalar linear Gaussian SSM problem. **(Center, Right)** Mean and standard deviation over 100 ELBO estimates of GS-VSMC with different values of $\tau$ and using $N$ particles.

## 4.2 Linear Gaussian SSM

We evaluate our proposed extension on higher dimensional linear Gaussian SSM problems. We use the model $(A)_{i,j} = \alpha^{|i-j|+1}$ where $\alpha = 0.42$, $Q = I$, $R = I$, $C \sim \mathcal{N}(0, I)$, and $T = 10$. We try two settings of $d_x = \dim(x_t)$ and $d_y = \dim(y_t)$. We use the proposal

$$q_\lambda(x_t|x_{t-1}) = \mathcal{N}\left(\mu_t + \mathrm{diag}(\beta_t)Ax_{t-1}, \mathrm{diag}(\sigma_t^2)\right)$$

and set the number of particles $N = 4$.

Figure 4 shows the mean and standard deviation of the estimated surrogate ELBO for VSMC, uVSMC and GS-VSMC over 100 runs. Note that we found the optimization step of IGR-VSMC to be prohibitively expensive in this context and so we did not bother to consider it. We also found that uVSMC converged at a much slower in this context (if at all), though we note that we did not fine tune the step sizes beyond trying the Adam and $\rho$ step sequences. The performance of GS-VSMC seems comparable to VSMC in the first setting but does not converge as fast in the second setting. TODO

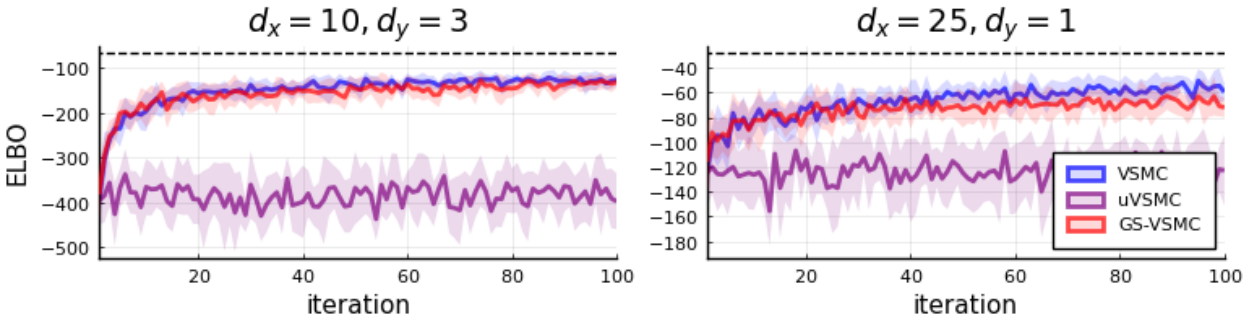


Figure 4: Mean and standard deviation of the estimated surrogate ELBO for two linear Gaussian SSM problem across 100 runs. The black line is the log marginal likelihood $\log p(y_{1:T})$. Note: uVSMC is using the Adam step sequence.

### 4.3 Capital Asset Pricing Model

We lastly evaluate our proposed extension on the Capital Asset Pricing Model (Capm) dataset (Verbeek, 2004).

## 5 Discussion

# References

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

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*. https://openreview.net/forum?id=rkE3y85ee

Jong, P. D. (1988). The likelihood for a state space model. *Biometrika*, *75*(1), 165–169. http://www.jstor.org/stable/2336450

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(14), 1–45. http://jmlr.org/papers/v18/16-107.html

Lawson, D., Tucker, G., Naesseth, C. A., Maddison, C., Adams, R. P., & Teh, Y. W. (2018). Twisted variational sequential monte carlo. In *Third workshop on bayesian deep learning (neurips)*.

Maddison, C. J., Mnih, A., & Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables.

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

Potapczynski, A., Loaiza-Ganem, G., & Cunningham, J. P. (2020). Invertible gaussian reparameterization: Revisiting the gumbel-softmax (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Eds.). In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems*. https://proceedings.neurips.cc/paper/2020/file/90c34175923a36ab7a5de4b981c1972f-Paper.pdf

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

Verbeek, M. (2004). *A guide to modern econometrics* (2nd). John Wiley & Sons.