Amortization and Surrogate Gaps in Variational Sequential Monte Carlo

Sida Li
The University of Chicago
listar2000@uchicago.edu

Abstract

Recent developments in MCMC and variational inference (VI) methods have enabled scalable approximate inference on problems where exact posteriors are intractable. Especially, the variational sequential Monte Carlo (VSMC) proposed by Naesseth et al. offers a stochastic optimization-based solution to filtering problems in sequential models by replacing the proposal distributions in traditional SMC with variational approximations. Adopting the framework of amortized inference suboptimality, this paper provides theoretical and numerical analysis of (1) the amortization gap in VSMC as a result of sharing (amortizing) the variational parameters over the entire dataset and (2) the "surrogate gap" introduced by optimizing a surrogate ELBO as the lower bound of the true ELBO objective. Results from this paper give practical guidances for the design and improvement of VSMC algorithms.

1 Introduction

A central issue in the fields of statistics and probabilistic machine learning involves calculating probability distributions and expectations [Blei, Kucukelbir, and McAuliffe 2017]. This task forms the basis of Bayesian statistics and machine learning, where all inferences are considered as expectations over the posteriors. While exact inferences on modern Bayesian problems are usually intractable, due to either complex model structures or the scale of datasets, recent advances in Markov-Chain Monte Carlo (MCMC) and variational inference (VI) offer approximate and scalable solutions to these problems [Zhang et al. 2018].

In particular, variational sequential Monte Carlo (VSMC) [C. Naesseth et al. 2018] was proposed to enable approximate inference on sequential models (e.g. State-Space Models (SSM), Stochastic Volatility Models), which play significant roles in domains like dynamical systems and finance [Svensson and Schön 2017, Zeng and Wu 2013]. VSMC lies in the intersection between MCMC and VI: on one hand, it inherits the mature sequential Monte Carlo (SMC) framework which guarantees the asymptotic optimality of particle approximations to filtering (posterior) distributions; on the other hand, it proposes a new class of variational family, where each element in the family is viewed as a potential proposal distribution for SMC, and derives a variational lower bound so that its parameters are amendable to stochastic optimization. Overall, VSMC has demonstrated remarkable performance in approximating log-marginal likelihoods over other methods.

Despite exhibited success in benchmarks, there are several problems within VSMC where careful investigations are needed. Firstly, while VSMC replaces the fixed proposal distributions in SMC with variational ones, the proposed variational family is a generic one. In other words, without further case-by-case knowledge of the specific models, the generic family parameterized by neural networks might be inefficient and suffer from incapacity in providing good approximations to the posteriors. Secondly, VSMC optimizes the variational parameters over a **surrogate ELBO** that lower-bounds the true ELBO, thus introducing additional divergence from the marginal log-likelihood. As some

Final Paper for Advanced Bayesian Methodology Seminar at UChicago (2023).

literature points out, this practice renders inconsistent gradient estimates in VSMC and contributes to the high variances in the training process [Chen, Sanz-Alonso, and Willett 2022].

Inspired by the study of inference suboptimality in variational auto-encoders (VAEs), this work seeks to provide theoretical and numerical analysis of the above problems, which are denoted as the **amortization gap** and the **surrogate gap** respectively. The rest of the paper is organized as follows: section 2.3 introduces related works and concisely summarizes the VSMC algorithm based on the original paper; section 3 formalizes the notions of amortization/surrogate gaps in VSMC; section 4 involves extensive experiments and numerical results on several sequential models over both synthetic and real-world datasets; finally, based on the preceding results, section 5 gives some insights and tips for designing better VSMC algorithms in practice. The main contributions of this work are:

- We rigorously formulate the notions of amortization gap and surrogate gap, and connect them to identified problems within VSMC.
- We perform numerical experiments to analyze the relative scales of the gaps and mechanisms to mitigate them.
- We offer practical guidance in designing better VSMC algorithms based on the results.

2 Background

2.1 Sequential Probabilistic Model

The sequential models within this paper's interest assume a sequence of latent variables $x_1, ..., x_T$ indexed by time (T is pre-defined). At each time t, the latent variable x_t evolves (stochastically) based on previous variables $x_1, ..., x_{t-1}$ according to some transition density $f(x_t|x_{1:t-1})$ [1]. In the following discussion we also assume the model to be Markovian so that $f(x_t|x_{1:t-1}) = f(x_t|x_{t-1})$, which holds in most situations like SSMs. At each time step, the data y_t is observed following some observation density $g(y_t|x_t)$. Altogether, the joint density is written as:

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

where $f(x_1)$ denotes the density (prior) for the beginning latent variable. As a concrete example, consider a linear Gaussian SSM. We have

$$x_t = Ax_{t-1} + e_t \tag{2}$$
$$y_t = Cx_t + \epsilon_t \tag{3}$$

where $x_t \in \mathbb{R}^k, y_t \in \mathbb{R}^d, A \in \mathbb{R}^{k\times k}, C \in \mathbb{R}^{k\times d}$. Also, $e_t \sim N(0, Q), \epsilon_t \sim N(0, R)$ are independent Gaussian noises. This setup gives explicit transition and observation densities since

$$x_t|x_{t-1} \sim N(Ax_{t-1}, Q) \quad y_t|x_t \sim N(Cx_t, R) \tag{4}$$

and the log marginal likelihood is tractable through methods like Kalman filters [Welch, Bishop, et al. 1995]. In general, however, calculating the marginal likelihood and the posterior $p(x_{1:T}|y_{1:T})$ is difficult. VSMC is thus proposed to perform approximate inference as a mixed variational and Monte Carlo method.

2.2 Sequential Monte Carlo

The sequential Monte Carlo (SMC), also known as Particle Filters, is an MCMC algorithm that sequentially approximates the **filtering distribution** $p(x_{1:t}|y_{1:t})$ with a collection of weighted samples (also called "particles") that evolves over time [Doucet, De Freitas, Gordon, et al. 2001]. Fixing the number of particles N as a hyper-parameter, at time $t = 1$, SMC samples $x_1^{(i)} \sim r_1(x_1)$ from some unconditional proposal density r_1 with importance weights $w_1^{(i)} = f(x_1^{(i)})/r_1(x_1^{(i)})$. For $t = 2, ..., T$, the following procedure is repeated:

[1] we use the shorthand $x_{1:t}$ in this paper for the collection $\{x_1, ..., x_t\}$

Discrete resampling $$a_t^{(i)} \sim \text{Categorical}(w_{t-1}^{(1)}, ..., w_{t-1}^{(N)}) \tag{5}$$

New Proposal $$x_t^{(i)} \sim r_t\Big(x_t | x_{t-1}^{(a_{t-1}^{(i)})}\Big) \tag{6}$$

Append State $$x_{1:t}^{(i)} = \Big(x_{1:t-1}^{(a_t^{(i)})}, x_t^{(i)}\Big) \tag{7}$$

Reweighting $$w_t^{(i)} = f(x_t^{(i)} | x_{t-1}^{(a_{t-1}^{(i)})}) \cdot g(y_t | x_t^{(i)}) / r_t(x_t | x_{t-1}^{(a_{t-1}^{(i)})}) \tag{8}$$

It is worth noting that in the third step we update the ith **trajectory** by concatenating the newly sampled $x_t^{(i)}$ with its ancestors (i.e. the $a_t^{(i)}$th trajectory at time $t-1$), and the new weights are updated for the entire trajectories instead of the particles at a single timestep. The correctness of the above construction is guaranteed by results from importance sampling and Monte Carlo theories, as we have:

Theorem 1 *As $N \to \infty$, we have the approximation*

$$p(x_{1:t}|y_{1:t}) \overset{d}{=} \hat{p}_N(x_{1:t}|y_{1:t}) := \sum_i^N \frac{w_t^{(i)}}{\sum_j w_t^{(j)}} \delta_{x_{1:t}^{(i)}} \tag{9}$$

where δ denotes the Dirac measure.

Proof: *we refer interested readers to p.14 in C. A. Naesseth, Lindsten, Schön, et al. 2019.*

This theorem justifies using the (weighted) empirical distribution of the trajectories as an approximation to the filtering distribution. When $t = T$, we have an approximate posterior for all latent variables. Another desirable property of SMC lies in its ability to provide an unbiased estimator of the marginal likelihood directly with particle weights [C. A. Naesseth, Lindsten, Schön, et al. 2019]:

$$p(y_{1:T}) \approx \hat{p}(y_{1:T}) := \prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^{(i)} \tag{10}$$

2.3 Variational Inference

The SMC algorithm detailed above enjoys the theoretical guaranteed in providing a particle-based approximation to the posterior. In practice, however, the proposal distributions ($r_1(x_1)$, $r_t(x_t|x_{t-1})$ for $t > 1$) need to be carefully chosen to balance between enough exploration of the latent space and exploitation of high-density regions - a tricky tradeoff faced by many importance sampling algorithms [Neal 2001].

VSMC improves over traditional SMC by leveraging **variational proposal densities** $r_1(x_1; \lambda)$ and $r_t(x_t|x_{t-1}; \lambda)$ parameterized by λ [C. Naesseth et al. 2018]. Through stochastic optimization of λ through a surrogate-ELBO loss (see next section), the good proposals will be learnt automatically without the need of manual designs. The following definitions and theorems are proposed in the original VSMC paper, and presented here as facts.

Proposition 1 *By using the variational proposal densities, the joint distribution of all variables sampled in the VSMC algorithm (based on (5)-(8)) can be expressed as:*

$$\phi(x_{1:T}^{(1:N)}, a_{1:T}^{(1:N)}; \lambda) = \left[\prod_{i=1}^N r(x_1^{(i)}; \lambda)\right] \cdot \left(\prod_{t=2}^T \prod_{i=1}^N \left[\frac{w_{t-1}^{(a_{t-1}^{(i)})}}{\sum_\ell w_{t-1}^{(\ell)}} r(x_t^{(i)} | x_{t-1}^{(\ell)}; \lambda)\right]\right) \tag{11}$$

with shorthand ϕ.

The next theorem establishes the variational distribution for the posterior $p(x_{1:T}|y_{1:T})$ through the variational proposals.

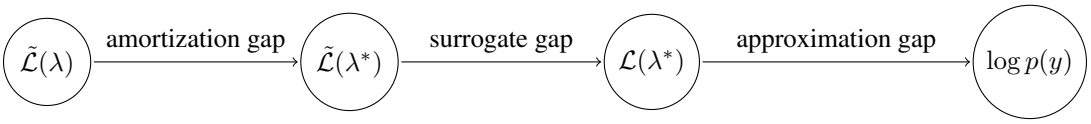
Theorem 2 *The variational approximation to the posterior of all latent variables is:*

$$q(x_{1:T}|y_{1:T}) = p(x_{1:T}, y_{1:T})\mathbb{E}_{\phi(\hat{x}_{1:T};\lambda)}\left[\mathbf{1}\{x_{1:T} \neq \hat{x}_{1:T}\}\hat{p}(y_{1:T})^{-1}\right]$$

where $\hat{p}(y_{1:T})$ is the unbiased marginal likelihood estimator derived by SMC.

Intuition: *we give an intuitive explanation here. Ignoring the indicator, the expectation on the RHS serves as an "averaged" marginal likelihood over the data $y_{1:T}$. The expectation accounts for different trajectories and their corresponding weights which are used to estimate $\hat{p}$. The indicator function means that the expectation is only taken over trajectories other than the current $x_{1:T}$ under evaluation; this can be thought about a "CAVI-like" coordinate ascent process where each coordinate is a complete trajectory. We refer the complete proof to Appendix A on C. Naesseth et al. 2018.*

3 Methods for Quantifying "Gaps"



In section 2, an overview of the sequential models and the variational formulation of VSMC are given. In order to make VSMC useful in practice, certain compromises and design choices are taken, which give rise to "gaps" in estimating the true marginal likelihood and optimizing for the variational parameters. This section is dedicated to unveiling two major gaps.

3.1 The Surrogate Gap

Following the expression of variational density q in Theorem 2, it is natural to think about leveraging the ELBO for optimizing λ, as in many other VI algorithms. Despite having a clean variational form, the expectation in q only allows biased Monte Carlo estimation with sampled trajectories; together with the need to avoid calculating the joint density $p(x_{1:T}, y_{1:T})$, VSMC proposes to use and perform gradient descent on a **surrogate ELBO**, which has the form:

$$\tilde{\mathcal{L}}(\lambda) := \sum_{t=1}^{T} \mathbb{E}_{\phi(x_{1:t}^{(1:N)}, a_{1:t}^{(1:N)};\lambda)}\left[\log \hat{p}(y_{1:t})\right] = \mathbb{E}_{\phi}\left[\log \hat{p}(y_{1:T})\right] \tag{12}$$

$$\leq \mathcal{L}(\lambda) \leq \log p(x_{1:T}) \tag{13}$$

where $\mathcal{L}(\lambda)$ and the first equality can be established based on a Jensen-inequality argument (for the sake of this paper we omit the proof but instead focus on numerically measure this gap). The **surrogate gap** is thus defined as the difference:

$$D_{\text{surrogate}}(\lambda) = \mathcal{L}(\lambda) - \tilde{\mathcal{L}}(\lambda) \tag{14}$$

In theory, the more trajectories are sampled for Monte Carlo estimation of $\mathcal{L}$, the smaller $D_{\text{surrogate}}(\lambda)$ and vice versa. This surrogate gap can be connected to a common criticism on VSMC that it gives unstable estimate of the log marginal likelihood[2], especially when sample size N is small [Chen, Sanz-Alonso, and Willett 2022].

3.2 The Amortization Gap

Another important, yet often ignored, aspect that is compromised in practice is the amortized parameterization of proposals. Even though the variational proposal densities r_t can be automatically learnt through stochastic optimization, the common choice of parameterization behind λ, e.g. neural networks, needs to amortize the parameters over the entire sequence of data - making it sub-optimal against carefully designed variational family that optimizes for data in each timestep t individually. The key takeaway is that this becomes an inevitable price to pay when we want to make VSMC

[2]the unstability on one hand exists because of the gap exists, but also because the gradient $\nabla_\lambda \tilde{\mathcal{L}}(\lambda)$ is not a consistent estimator of $\nabla_\lambda \mathcal{L}(\lambda)$. Nevertheless they are all due to the surrogate ELBO.

generic - i.e. we choose to fit λ with neural networks instead of trying to come up with case-by-case proposal parameters (even if the latter might work better). If we let λ^* to denote the best parameters among all possible proposals, while λ denotes the best (optimized) parameters for models restricted to neural networks with parameters amortized over the sequence. We define the **amortization gap** as:

$$D_{\text{amortize}}(\lambda) = \tilde{\mathcal{L}}(\lambda^*) - \tilde{\mathcal{L}}(\lambda) \tag{15}$$

Note that both terms on the RHS are surrogate ELBOs, so overall we have the relationship

$$\mathcal{L}(\lambda^*) \geq \tilde{\mathcal{L}}(\lambda^*) \geq \tilde{\mathcal{L}}(\lambda) \tag{16}$$

The concept of amortization gap is adapted from inference suboptimality analysis in variational auto-encoders [Cremer, Li, and Duvenaud 2018], which mentions a similar amortization effect when sharing the variational parameters across the entire training dataset instead of optimizing each data point individually (here we are more concerned about amortization over time horizon). This gap is mainly accountable for VSMC's inability to provide consistently good log marginal likelihood estimates, especially when the model parameters need to be jointly optimized and the proposals are simple neural networks with limited expressiveness.

4 Experiments

Table 1: Numerical results from linear SSM experiments (SD in parenthesis)

Metrics	$Dx = 5, Dy = 1$	$Dx = 5, Dy = 3$	$Dx = 10, Dy = 1$	$Dx = 10, Dy = 10$
$\log p(y_{1:T})$	-18.29	-29.78	-57.57	-98.39
$\mathcal{L}(\lambda^*)$	-19.11	-83.45	-65.38	-121.75
$\tilde{\mathcal{L}}(\lambda^*)$	-19.56(2.01)	-106.11(15.36)	-71.72(17.93)	-132.56(19.41)
$\tilde{\mathcal{L}}(\lambda)$	-26.15(3.88)	-288.02(71.25)	-224.12(14.75)	-441.02(16.74)
$D_{\text{surrogate}}(\lambda^*)$	0.45	22.66	6.34	10.81
$D_{\text{amortize}}(\lambda)$	**6.59**	**181.91**	**152.40**	**308.46**

In the experiments, through concrete numerical experiments, we will quantify the two major gaps identified above and gain insights into their relative sizes inside two important models: linear SSMs and stochastic volatility models [Chib, Omori, and Asai 2009]. In both experiments, the vanilla ELBO $\mathcal{L}(\lambda)$ is calculated in the same way as surrogate $\hat{\mathcal{L}}(\lambda)$ but uses a considerably large N (as explained earlier, they are the same when $N \to \infty$), while the surrogate use a smaller N that is closer to its value in real-world application ($N \approx 10$). We parameterize λ as a 3-layer feed-forward neural network that takes in the timestep t and latent x_{t-1} and output the mean & variance of a Gaussian proposal kernel $r_t(x_t|x_{t-1})$. On the other hand, we deliberately parameterize λ^* by a (non-neural network) variational family that can factorize over the different timesteps and track the transition dynamics $f(x_t|x_{t-1})$ closely. This can be thought of a case-by-case manual design that generic VSMC algorithm cannot achieve. It is worth noting that both the models λ and λ^* receive the timestep t as an input, but the former (neural network) has to share parameters (FFN weights and biases) across all computations while the later can deliberately optimize a subset of its parameters (i.e. $\lambda^* = \{\lambda_1^*, ..., \lambda_T^*\}$) for a single step.

4.1 Linear Gaussian SSM

The setup of linear Gaussian SSM has been given in 2.1. Below we concisely summarize the experiment details.

- **model parameters**: we take $T = 10, N = 5$, noise matrices $Q = R = I$. The transition matrix A is designed as $(A)_{ij} = \alpha^{|i-j|}$ for $\alpha = 0.5$. Across the four experiments, we play with different dimensions of $x_t^{(i)}$ and $y_t^{(i)}$ - denoted as Dx and Dy in the plots.

- **Ground-truth log marginal**: for linear SSMs, $\log p(y_{1:T})$ is actually tractable through the use of Kalman filters [Welch, Bishop, et al. 1995]. We also report this value for each experiment as a reference upper bound.
- **choice of λ^***: the manually engineered variational proposals take the form
$$r_t(x_t|x_{t-1};\lambda^*) = N(x_t|\mu_t + diag(\beta_t)Ax_{t-1}, diag(\sigma_t^2)) \tag{17}$$
so $\lambda^* = \{\mu_t, \beta_t, \sigma_t\}_{t=1}^T$.

Table 1 summarizes the experiment results across four configurations with different latent and data dimensions. Each configuration is run a total of 10 times (5000 iterations for each time) so the standard deviations of the ELBOs are included as well. The most obvious result is that the **amortization gap** is much larger, if not one magnitude higher, than the **surrogate gap**. In fact, for simple model like config 1 ($Dx = 5, Dy = 1$), the differences among true marginal, ELBO, and surrogate ELBO for λ^* is neglectible. However, when we try to run generic VSMC with neural network parameterization λ, the surrogate ELBO plummets drastically. On the other hand, as model complexity increases, the surrogate gap also becomes non-trivial while the amortization gap keeps enlarging.

For models like linear Gaussian SSM, understanding the dynamics and designing specific variational parameters is crucial for the success of VSMC algorithm.

4.2 Stochastic Volalility Models

Table 2: Numerical results from stochatic volatility models (SD in parenthesis)

Metrics	$N = 5$	$N = 10$	$N = 50$
$\mathcal{L}(\lambda^*)$	-47.51	-47.51	-47.51
$\tilde{\mathcal{L}}(\lambda^*)$	-81.56(4.15)	-62.17(3.98)	-50.13(3.13)
$\tilde{\mathcal{L}}(\lambda)$	-103.54(10.56)	-82.27(8.21)	-61.89(7.69)
$D_{\text{surrogate}}(\lambda^*)$	**34.06**	14.66	2.62
$D_{\text{amortize}}(\lambda)$	21.98	**20.10**	**11.76**

Stochastic Volatility Models are advanced financial models used to analyze and predict the volatility of assets in financial markets. These models account for the random nature of market volatility, making them crucial tools in understanding and managing financial risks. A general setup is

$$x_t = \mu + \phi(x_{t-1} - \mu) + e_t \tag{18}$$
$$y_t = \beta \exp\left(\frac{x_t}{2}\right)\epsilon_t \tag{19}$$

with $e_t \sim N(0, Q), \epsilon_t \sim N(0, I)$ and $x_1 \sim N(\mu_0, Q)$. Both x_t and y_t have the same dimensionality and all operations above are element-wise when x_t is multivariate. Instead of using synthetic data, our experiment runs on Federal Reserves exchange rate dataset for the past 24 months ($T = 24$) [Federal Reserve Board 2023]. We pick 5 currencies' exchange rate with respect to USD (so $y_t \in \mathbb{R}^5$) and run three experiments with identical configurations except for different sample size N. Some experiment details include:

- **model parameters**: since this is a real-world dataset, we don't have access to the model parameters θ in advance. Instead we use a separate expectation-maximization algorithm [Dempster, Laird, and Rubin 1977] to estimate $\hat{\theta}$. Even though VSMC allows joint optimization of $\tilde{\mathcal{L}}(\theta, \lambda)$, such training is unstable and not of this paper's concern.
- **choice of λ^***: the manually chosen variational proposals take the form
$$r_t(x_t|x_{t-1};\lambda^*) \propto f(x_t|x_{t-1})N(x_t;\mu_t,\Sigma_t) \tag{20}$$
so $\lambda^* = \{\mu_t, \Sigma_t\}_{t=1}^T$. Since $x_t|x_{t-1} \sim N(\mu + \phi(x_{t-1} - \mu, Q)$, the resulting proposal is also Gaussian. And the product formulation of r also captures the multiplication of noise ϵ_t in the observation y_t [Collins 2013].

Base on the results from Table 2, we see that the surrogate gap no longer dominates the amortization gap by a far margin for more complicated models. When we take small $N = 5$, we even see a smaller amortization gap. As we increase the number of sample trajectories taken, the surrogate gap drops much quicker than the amortization gap: when $N = 50$, the difference from the true ELBO (which is approximated by using an unrealistic $N = 1000$) becomes fairly small. These experiments confirm the theoretical understanding that the surrogate gap is mainly attributable to approximation errors due to insufficient sample sizes.

Based on the above analysis and experiments, we are able to give out a few guidelines in implementing VSMC for practicle use.

1. For models with known and simple dynamics (i.e. f and g are known), it might be desirable to check out whether there exists optimal (non-variational) proposal distributions. VSMC can be applied by adding variational parameters on top of these proposals, and this would usually outperform cases when the proposals are parameterized by more generic architectures (e.g. neural networks). This practice has been shown to be effective in greatly reducing the **amortization gap**.
2. For models where the T is not too large and sampling from the forward process (i.e. SMC transition dynamics) is not costly, try to increase the sample size N as much as possible. This practice unsurprisingly will reduce the **surrogate gap**; more importantly, this effect will be more obvious when combined with practice 1.

5 Discussion & Future Works

In this paper, we gave a general introduction to the sequential probablistic models and the corresponding VSMC algorithm for solving the filtering (posterior inference) problems. We innovatively borrowed the inference suboptimality framework from VAE literature into the analysis of VSMC. By carefully examining the model setup, we identified the existence of amortization and surrogate gaps in VSMC, and assosicated them with known issues with the algorithm. Finally, based on the numerical experiments, we successfully quantified these gaps and provided practical suggestions for anyone wishing to apply VSMC to their problems.

Due to the time limit, there are a few things that we failed to include in this paper and can be left for future works. Firstly, the original paper suggests the use of various variance-reduction techniques to make gradient estimates of the surrogate ELBO more accurate [C. Naesseth et al. 2018]. The effect of these techniques on the aforementioned gaps will be interesting to figure out. Secondly, other than linear Gaussian SSMs and stochastic volatility models, there are other interesting sequential models such as deep markov models [Khurana et al. 2020] that are not covered. Finally, the model parameters θ in this paper are assumed to be known; in practice, however, partially known or unknown dynamics [Andrieu and Doucet 2002] would give rise to new challenges. It is thus desirble to understand VSMC's behavior when the model and variational parameters are jointly optimized.

References

Andrieu, Christophe and Arnaud Doucet (2002). "Particle filtering for partially observed Gaussian state space models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.4, pp. 827–836.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Chen, Yuming, Daniel Sanz-Alonso, and Rebecca Willett (2022). "Autodifferentiable ensemble Kalman filters". In: *SIAM Journal on Mathematics of Data Science* 4.2, pp. 801–833.

Chib, Siddhartha, Yasuhiro Omori, and Manabu Asai (2009). "Multivariate stochastic volatility". In: *Handbook of financial time series*. Springer, pp. 365–400.

Collins, John Parnell (2013). "Comparison of Methods for Estimating Stochastic Volatility". In.

Cremer, Chris, Xuechen Li, and David Duvenaud (2018). "Inference suboptimality in variational autoencoders". In: *International Conference on Machine Learning*. PMLR, pp. 1078–1086.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.

Doucet, Arnaud, Nando De Freitas, Neil James Gordon, et al. (2001). *Sequential Monte Carlo methods in practice*. Vol. 1. 2. Springer.

Federal Reserve Board (2023). *Foreign Exchange Rates - H.10*. Accessed: 2023-12-12. URL: `https://www.federalreserve.gov/releases/h10/`.

Khurana, Sameer et al. (2020). "A convolutional deep markov model for unsupervised speech representation learning". In: *arXiv preprint arXiv:2006.02547*.

Naesseth, Christian et al. (2018). "Variational sequential monte carlo". In: *International conference on artificial intelligence and statistics*. PMLR, pp. 968–977.

Naesseth, Christian A, Fredrik Lindsten, Thomas B Schön, et al. (2019). "Elements of sequential monte carlo". In: *Foundations and Trends® in Machine Learning* 12.3, pp. 307–392.

Neal, Radford M (2001). "Annealed importance sampling". In: *Statistics and computing* 11, pp. 125–139.

Svensson, Andreas and Thomas B Schön (2017). "A flexible state–space model for learning nonlinear dynamical systems". In: *Automatica* 80, pp. 189–199.

Welch, Greg, Gary Bishop, et al. (1995). "An introduction to the Kalman filter". In.

Zeng, Yong and Shu Wu (2013). *State-space models: Applications in economics and finance*. Vol. 1. Springer.

Zhang, Cheng et al. (2018). "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8, pp. 2008–2026.