Variational Resampling

Oskar Kviman
KTH

Nicola Branchini
University of Edinburgh

Víctor Elvira
University of Edinburgh

Jens Lagergren
KTH

Abstract

We cast the resampling step in particle filters (PFs) as a variational inference problem, resulting in a new class of resampling schemes: variational resampling. Variational resampling is flexible as it allows for choices of 1) divergence to minimize, 2) target distribution to input to the divergence, and 3) divergence minimization algorithm. With this novel application of VI to particle filters, variational resampling further unifies these two powerful and popular methodologies. We construct two variational resamplers that replicate particles in order to maximize lower bounds with respect to two different target measures. We benchmark our variational resamplers on challenging smoothing tasks, outperforming PFs that implement the state-of-the-art resampling schemes.

1 INTRODUCTION

Particle filters are a widely used class of algorithms to perform stochastic nonlinear filtering that, since their popularization in (Gordon et al., 1993), have found applications in countless domains, notably (probabilistic) mobile robot localization (Thrun et al., 2000, 2001; Maggio et al., 2023; Placed et al., 2023), epidemic tracking (Storvik et al., 2023) option pricing in mathematical finance (Creal, 2012), Bayesian phylogenetic inference (Bouchard-Côté et al., 2012; Moretti et al., 2021; Koptagel et al., 2022); and many more.

In PFs, a resampling step is used to filter out low-probability particles in order to avoid particle degeneracy. Classical resampling schemes, like multinomial, systematic or stratified resampling, are designed to produce resampled measures that do not deviate too much from the normalized importance weighted measure. This is achieved by making the resampling schemes produce resampled measures that are unbiased.

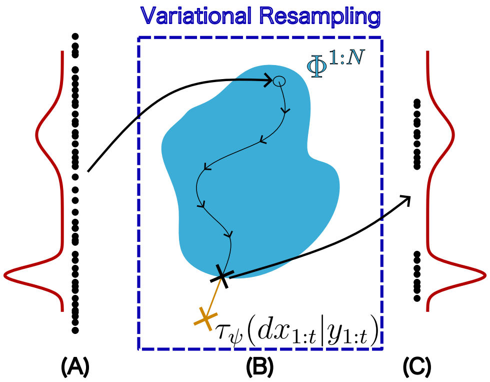


Figure 1: **(A)** At each time step t in a particle filter, N particles have been proposed (black points) in the latent space. The posterior density is shown in red. Each particle, i, has a single replica, i.e. $\phi_t^i = 1$. **(B)** Given the proposed particles, variational resampling infers the combination $\phi_t^{1:N} \in \Phi^{1:N}$ that minimizes a divergence with respect to a target measure, $\tau_\psi(dx_{1:t}|y_{1:t})$. **(C)** The optimal resampled measure produced by variational resampling is then returned. The surviving particles (black points) each have at least one replica.

The variational inference (VI; Jordan et al. (1999); Blei et al. (2017)) methodology allows for inferring a distribution by directly minimizing a divergence to a target distribution. As such, we cast the resampling step as a VI problem, where we recognize the simpler, rational-valued resampled measure to be a variational distribution, approximating a target measure. In contrast to the classical resampling schemes, this allows us to directly minimize a divergence between the two measures.

In doing so, we are proposing a novel class of resampling schemes, which we call variational resampling. Variational resampling is flexible as it allows the practitioner to design and choose among divergences to minimize, target measures to approximate and optimization algorithm for minimizing the divergence. It is principled as it relies on the VI methodology, and

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

further unifies VI and PFs. Indeed, variational resampling is a new paradigm for developing resampling schemes in PFs.

In this work, we construct two instances (resamplers) of variational resampling. The first resampler produces resampled measures which are accurate approximations of the normalized importance weighted measure, improving over the state-of-the-art methods.

The second resampler is explicitly designed to enhance smoothing-related estimates for the PF. Utilizing the flexibility of variational resampling, this is done by resampling based on the likelihoods of the particle trajectories up until the current time step. We think of this new resampler as performing *online smoothing*, resulting in smoothing estimates superior to the previous state-of-the-art methods.

Our contributions can be itemized as follows:

- We propose a new class of resampling schemes, variational resampling. This is a new paradigm for developing resampling schemes as it casts the resampling step as a VI problem, paving way for interesting future work.
- We design a deterministic optimization algorithm that maximizes a lower bound on the log-normalizing constant of the target measure, which we call the lower bound (LB) resampler.
- We give two target measures to use as input in the LB resampler, resulting in two variational resampling schemes. The first (Sec. 3.2.1) resamples particles by minimizing the Kullback-Leibler (KL) divergence between the resampled measure and the normalized importance weighted one. The second (Sec. 3.2.2) is carefully designed to enhance the smoothing estimates from the PF, resulting in impressive performances.

We illustrate the practical benefits of our framework first on toy experiments, providing intuition, and then experiment on challenging state-space models using synthetic and real data.

2 BACKGROUND

Let $x_{1:t}$ and $y_{1:t}$ be sequences of latent variables and observations, respectively, each of length t, and let $p(dx_{1:t}|y_{1:t})$ be a posterior measure with corresponding unnormalized density $\gamma(x_{1:t}|y_{1:t})$ which factorizes as follows

$$\gamma(x_{1:t}|y_{1:t}) = g(y_t|x_t)f(x_t|x_{t-1})\gamma(x_{1:t-1}|y_{1:t-1}). \quad (1)$$

We use the dx argument to distinguish measures from densities. Furthermore, let $\delta_{x^i_{1:t}}(dx_{1:t})$ be a Dirac measure on the latent space, N be the number of particles, $x^i_{1:t}$ be the positions in the particle trajectory of the i-th particle in $\mathbb{R}^t$, and $\phi^i_t \in [1, ..., N]$ the number of times the i-th particle is replicated at time step t. That is, $\phi^{1:N}_t$ is in $\Phi^{1:N}$, the set of all possible combinations of $\phi^{1:N}_t$, such that $\sum_{i=1}^N \phi^i_t = N$.

Assuming that resampling is performed at every time step in the particle filter, then the importance weight for particle i at time t is

$$w^i_t = \frac{g(y_t|x^i_t)f(x^i_t|x^i_{t-1})}{k(x^i_t|x^i_{t-1})}, \quad (2)$$

where x^i_t is simulated from a proposal density, $k(x_t|x^i_{t-1})$. From these weights, an estimate of the marginal log-likelihood can be computed

$$\log \widehat{Z}_t = \log \frac{1}{N}\sum_{i=1}^N w^i_t \approx \log p(y_t|y_{1:t-1}), \quad (3)$$

where

$$\log \widehat{Z}_{1:t} = \sum_{k=1}^t \log \widehat{Z}_k \approx \log p(y_{1:t}) \quad (4)$$

We now introduce the normalized importance weighted measure,

$$\pi_{\widetilde{w}^{1:N}_t}(dx_{1:t}|y_{1:t}) = \sum_{i=1}^N \widetilde{w}^i_t \delta_{x^i_{1:t}}(dx_{1:t}), \quad (5)$$

where $\widetilde{w}^i_t = w^i_t / \sum_{j=1}^N w^j_t$, which itself can be viewed as an approximation to the target measure of interest, $\pi_{\widetilde{w}^{1:N}_t}(dx_{1:t}|y_{1:t}) \approx p(dx_{1:t}|y_{1:t})$ (Doucet et al., 2009, Section 3.4).

After resampling at time step t, a resampled measure is obtained as

$$q_{\phi^{1:N}_t}(dx_{1:t}|y_{1:t}) = \sum_{i=1}^N \frac{\phi^i_t}{N} \delta_{x^i_{1:t}}(dx_{1:t}), \quad (6)$$

which is a normalized measure. The classical resampling schemes, like multinomial, systematic or stratified resampling, are designed to produce resampled measures that do not deviate too much from $\pi_{\widetilde{w}^{1:N}_t}(dx_{1:t}|y_{1:t})$. This is achieved by making the resampling schemes produce resampled measures that are unbiased, i.e.

$$\mathbb{E}_{R\left(\phi^{1:N}_t|\widetilde{w}^{1:N}_t\right)}[\phi^i_t] = N \cdot \widetilde{w}^i_t, \quad i = 1, \dots, N \quad (7)$$

where $R\left(\phi^{1:N}_t|\widetilde{w}^{1:N}_t\right)$ is the law of the resampling scheme. The discrepancy between the measures can, however, be large. This is quantified in Chopin et al. (2020, Chapter 9, Section 9.7), where the total variation (TV) distance of Eq. (5) and (6) is measured.

To constrain the analysis of the resampling methods in this work, we consider bootstrap PFs (BPFs; i.e. the prior density is utilized as the proposal density) and assume that resampling is employed at every time step. See Alg. 1 for an algorithmic description.

Algorithm 1 BPF with resampling every t

```
Initialization. Obtain initial particle positions $\{x_1^i\}_{i=1}^N$ as a sample from the prior pdf $p(x_1)$.
for $t = 1, \dots, T$ do
    Weighting. Compute the unnormalized weights as
        $w_t^i = g(y_t|x_t^i), \quad i = 1, \dots, N \qquad (8)$
    Resampling. Obtain $\phi_t^{1:N}$ and the resampled indices $\{r(i)\}_{i=1}^N$ based on $\widetilde{w}_t^{1:N}$
    Updating trajectories.
        $x_{1:t}^{r(i)} = x_{1:t}^i, \quad i = 1, \dots, N \qquad (9)$
    if $t < T$ then
        Propagation.
            $x_{t+1}^i \sim f(x_{t+1}|x_t^{r(i)}), \quad i = 1, \dots, N \qquad (10)$
    end if
end for
```

3 VARIATIONAL RESAMPLING

As discussed in the previous sections, a desirable property of a resampling scheme is that the resampled measure is a good approximation of the normalized importance weighted measure. The importance of this may be highlighted via the TV distance between the two measures, which can be seen as a quantification of how much the resampling scheme degenerates the approximation of the posterior measure obtained from $\pi_{\widetilde{w}_t^{1:N}}(dx_{1:t}|y_{1:t})$ (Chopin et al., 2020, Chapter 9, Section 9.7).

Variational resampling accommodates the possibility to construct resampling schemes that directly minimize a divergence between $q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})$ and $\pi_{\widetilde{w}_t^{1:N}}(dx_{1:t}|y_{1:t})$ with respect to the number of replicates of each particle, $\phi_t^{1:N}$. Therefore, casting the resampling step as a VI problem is a change of paradigm regarding how to develop resampling methods.

In our proposed variational resampling setting, the task is to infer $\phi_t^{1:N}$ such that some function $\mathcal{L}$ is optimized with respect to a target measure, $\tau_\psi(dx_{1:t}|y_{1:t})$, with Dirac measures at values $x_{1:t}^{1:N}$ and parameters ψ. For example, if we let $\pi_{\widetilde{w}_t^{1:N}}(dx_{1:t}|y_{1:t})$ be the target measure (giving, $\psi = \widetilde{w}_t^{1:N}$) and $\mathcal{L}$ is the KL divergence, then the resampled measure is one that minimizes the KL divergence to $\pi_{\widetilde{w}_t^{1:N}}(dx_{1:t}|y_{1:t})$. In Fig. 1, we illustrate an abstraction of variational resampling.

Given $\mathcal{L}$, $\tau_\psi(dx_{1:t}|y_{1:t})$ and $x_{1:t}^{1:N}$, one can devise optimization algorithms for inferring $\phi_t^{1:N}$ and, as a result, the resampled measure, $q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})$. Implementing variational resampling for a BPF requires minor modifications of the BPF in Alg 1. These can be implemented with few additional lines of code, as exemplified in Alg. 2. Next, we devise an optimization algorithm for a general target measure. Then, in Sec. 3.2.1-3.2.2, we propose two target measures and apply them to the optimization algorithm, leading to our two variational resamplers.

3.1 The Lower Bound Resampler

To construct our optimization algorithm, we let $\tau_\psi(dx_{1:t}|y_{1:t})$ be an unnormalized target measure with normalizing constant $\log Z(\tau_\psi)$, and we will below derive the function $\mathcal{L}$ as a lower bound on $\log Z(\tau_\psi)$. We start from the negated KL between the resampled measure and the normalized $\tau_\psi(dx_{1:t}|y_{1:t})$

$$
\begin{aligned}
&- \mathrm{KL}\left(q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t}) \| \tau_\psi(dx_{1:t}|y_{1:t})/Z(\tau_\psi)\right) && (11)\\
&= \int q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t}) \log \frac{\tau_\psi(dx_{1:t}|y_{1:t})/Z(\tau_\psi)}{q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})} && (12)\\
&= \int q_{\phi_t^{1:N}}(x_{1:t}|y_{1:t}) \log \frac{\tau_\psi(x_{1:t}|y_{1:t})/Z(\tau_\psi)}{q_{\phi_t^{1:N}}(x_{1:t}|y_{1:t})} dx_{1:t} && (13)\\
&= \sum_{i=1}^N \frac{\phi_t^i}{N} \log \frac{\tau_\psi(x_{1:t}^i|y_{1:t})}{\phi_t^i/N} - \log Z(\tau_\psi) \leq 0, && (14)
\end{aligned}
$$

where in Eq. (13) we replace the Radon-Nykodim derivative with the density ratio. Also, in Eq. (14) the integral becomes a sum since the measures are concentrated in $x_{1:t}^{1:N}$, and, by convention, $0 \log \frac{0}{0} = 0$. Finally, the negative KL is non-positive.

Adding $\log Z(\tau_\psi)$ on both sides of the inequality in Eq. (14), we define the $\mathcal{L}$ which we aim to optimize in our algorithm

$$
\begin{aligned}
&\mathcal{L}\left(\tau_\psi(dx_{1:t}|y_{1:t}), q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})\right) := && (15)\\
&\sum_{i=1}^N \frac{\phi_t^i}{N} \log \frac{\tau_\psi(x_{1:t}^i|y_{1:t})}{\phi_t^i/N} \leq \log Z(\tau_\psi), && (16)
\end{aligned}
$$

i.e., $\mathcal{L}$ is a lower bound on $\log Z(\tau_\psi)$.

We seek to maximize Eq. (15) w.r.t. $\{\phi_t^i\}_{i=1}^N$, which leads to the following optimization problem

$$\{\phi_t^{i,\star}\}_{i=1}^N = \arg\max_{\{\phi_t^i\}_{i=1}^N \in \Phi^{1:N}} \mathcal{L}\left(\tau_\psi(dx_{1:t}|y_{1:t}), q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})\right) \tag{17}$$

We will address Eq. (17) with our lower bound (LB) resampler, an optimization algorithm that finds $\phi_t^{1:N} \in \Phi^{1:N}$ which maximize $\mathcal{L}$.

Concretely, Eq. (15) is maximized by greedily replicating particles in a sequential manner. Initially, all particles have zero replicates. Until the total number of replicates is N, we choose to replicate the particle that contributes to the largest (intermediate) LB score. An algorithmic description is provided in Alg. 3. By using heapsort, extracting from the heap to compute the arg max function in line 6 in Alg. 3, the LB resampler has a time complexity of $\mathcal{O}(N \log N)$.

Algorithm 2 BPF with *variational resampling* every t

```
Initialization. Obtain initial particle positions $\{x_1^i\}_{i=1}^N$ as a sample from the prior pdf $p(x_1)$.
for $t = 1, \dots, T$ do
    Weighting. Set the unnormalized importance weights as
        $$w_t^i = g(y_t|x_t^i), \quad i = 1, \dots, N \tag{18}$$
    and recursively update Eq. (28).
    Variational resampling.
    if LB resampler with $\pi_{\gamma_t^{1:N}}(dx_{1:t}|x_{1:t}^{1:N})$ then
        Input $\{\gamma_t^i\}_{j=1}^N$ to Alg. 3 to obtain $\phi_t^i$ replicates of particle $i$ and resampled index $r(i)$.
    else if LB resampler with $\pi_{w_t^{1:N}}(dx_{1:t}|y_{1:t})$ then
        Input $\{w_t^j\}_{j=1}^N$ to Alg. 3 to obtain $\phi_t^i$ replicates of particle $i$ and resampled index $r(i)$.
    end if
    Updating trajectories. Set $x_{1:t}^{r(i)} = x_{1:t}^i$
    if $t < T$ then
        Propagation.
            $$x_{t+1}^i \sim f(x_{t+1}|x_t^{r(i)}), \quad i = 1, \dots, N \tag{19}$$
    end if
end for
```

3.2 Target Measures

Here we discuss two target measures that we will use as input to the LB resampler.

3.2.1 Importance Weighted Target Measure

The first target measure we construct is one that minimizes the KL divergence between the resampled measure and the normalized importance weighted measure in Eq. (5). As we will show, this is coincidentally achieved simultaneously by maximizing a lower bound on $\log \widehat{Z}_t$ (see Eq. (3)).

Let $\tau_\psi(dx_{1:t}|y_{1:t})$ be the (unnormalized) importance weighted target measure,

$$\tau_\psi(dx_{1:t}|y_{1:t}) = \pi_{w_t^{1:N}}(dx_{1:t}|y_{1:t}) \tag{20}$$

$$= \sum_{i=1}^N w_t^i \delta_{x_{1:t}^i}(dx_{1:t}), \tag{21}$$

i.e. the unnormalized version of the importance weighted measure in Eq. (5), therefore it is not a probability measure (but still a measure). Its log-normalizing constant is $\log Z(q_{w_t}) = \log \sum_{i=1}^N w_t^i$.

Plugging this target measure into the LB resampler, Eq. (15) becomes a lower bound on $\log Z(q_{w_t})$,

$$\mathcal{L}(\pi_{w_t^{1:N}}(dx_{1:t}|y_{1:t}), q_{\phi_t^{1:N}}(dx_{1:t}|y_{1:t})) \le \log \sum_{i=1}^N w_t^i. \tag{22}$$

Maximizing this bound is equivalent to minimizing the KL divergence between the resampled measure and the normalized importance weighted measure.

Interestingly, the LB resampler using this target measure simultaneously maximizes a lower bound on $\log \widehat{Z}_t$. To see this, start by plugging $\pi_{\widetilde{w}_t^{1:N}}(dx_{1:t}|y_{1:t})$ into Eq. (14), and observe that this gives a negated KL divergence between the resampled measure and the normalized importance weighted measure (both are normalized),

$$\sum_{i=1}^N \frac{\phi_t^i}{N} \log \frac{\pi_{\widetilde{w}_t^{1:N}}(x_{1:t}^i|y_{1:t})}{\phi_t^i/N} \le 0. \tag{23}$$

Then add $\log \frac{1}{N} \sum_{i=1}^N w_t^i$ on both sides to get a lower bound on the estimated marginal log-likelihood, induced by the resampling measure,

$$\widehat{\text{ResELBO}}_t := \sum_{i=1}^N \frac{\phi_t^i}{N} \log \frac{\pi_{w_t^{1:N}}(x_{1:t}^i|y_{1:t})/N}{\phi_t^i/N} \tag{24}$$

$$\le \log \frac{1}{N} \sum_{i=1}^N w_t^i = \log \widehat{Z}_t. \tag{25}$$

Note that the $1/N$ coefficients in the log ratio in the L.H.S. of the inequality can be discarded during maximization, and so maximizing $\widehat{\text{ResELBO}}_t$ is equivalent to maximizing the bound in Eq. (22).

In our state-space model experiments we will indeed see empirically that this variational resampler achieves the best approximation of the normalized importance weighted measure in terms of the TV distance, as well as the smallest KL divergences to the normalized importance weighted measure in Sec. 5.1.

Algorithm 3 The LB resampler

```
Input: ${u^i}_{i=1}^N$
set $\phi^i = 0, \forall i$
define $h(\phi^i, u^i) = \phi^i \log \frac{u^i}{\phi^i}$, where $h(0, u^i) := 0$
define $C^+(\phi^i, u^i) = h(\phi^i + 1, u^i) - h(\phi^i, u^i)$
while $\sum_{i=1}^N \phi^i < N$ do
    compute $m = \arg\max_i C^+(\phi^i, u^i)$
    set $\phi^m = \phi^m + 1$
end while
return ${\phi^i}_{i=1}^N$, ${r(i)}_{i=1}^N$ (based on the replicates)
```

3.2.2 Model Based Target Measure

Next, we define a target measure designed in order to produce powerful PF-based smoothing approximations. We call this measure the model based target measure,

$$\tau_\psi(dx_{1:t}|y_{1:t}) = \pi_{\gamma_t^{1:N}}(dx_{1:t}|y_{1:t}) \tag{26}$$

$$= \sum_{i=1}^N \gamma_t^i \delta_{x_{1:t}^i}(dx_{1:t}), \tag{27}$$

where

$$\gamma_t^i = \prod_{k=1}^t g(y_k|x_k^i) f(x_k^i|x_{k-1}^i), \tag{28}$$

and we denote $f(x_1^i|x_0) := p(x_1^i)$.

When used in the LB resampler, $\pi_{\gamma_t^{1:N}}(dx_{1:t}|y_{1:t})$ gives a lower bound on its self-normalized constant. Analogous to the derivations in the previous section, this means that this resampler minimizes the KL between the resampled measure and the self-normalized version of $\pi_{\gamma_t^{1:N}}(dx_{1:t}|y_{1:t})$.

The model based target measure has two powerful properties. First, as shown in Eq. (28), the target factorises from time t to the first time step *regardless whether resampling was performed before time t*. This is significantly different from the other target measures and classical resampling in general,[1] where this information is lost as the previous importance weights are set to uniform post resampling. This is an interesting property when approximating the smoothing measure as it will enforce replication of particle trajectories with higher unnormalized smoothing density scores over the full sequence up to time t. We think of this as an *online smoothing* property.

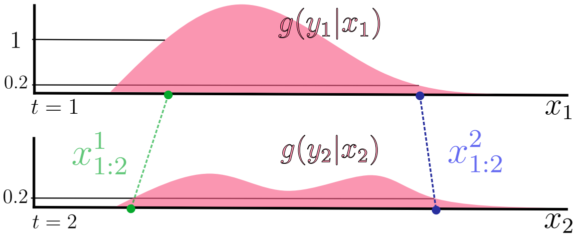


Figure 2: The likelihood functions (pink) in two time steps for two particle trajectories (purple) are show. We are using a BPF, resampling at every time step and, for simplicity, a uniform prior. The importance weights at $t = 2$, w_2^1 and w_2^2, are equal, while the model based target measure assigns more weight to the green trajectory based on its history—*regardless if resampling was performed at $t = 1$*. We refer to this property as online smoothing, and it allows our LB resampler with π_γ to resample particles with higher posterior-density trajectories. See Sec. 3.2.2 for details.

Secondly, the target does not include the evaluation of the proposal density, implying that one can use this target measure to construct a variational-resampling based PF with a proposal density which is difficult to evaluate, but possible to sample from. This is not possible for the other target measures and in classical resampling, as the importance weights are functions of the evaluation of the proposal density.

We give the following example to demonstrate the power of the online smoothing property.

Example 1 (Online Smoothing) In Fig. 2 we consider two particle trajectories over two time steps and their corresponding likelihoods. To simplify the illustration, we assume uniform prior densities, $f(x_2|x_1) = f(x_1) = c$. However, the example generalizes to any number of time steps and prior densities.

At $t = 1$, particle 1 (x_1^1; green) has a higher observation likelihood than particle 2 (x_1^2; blue), $g(y_1|x_1^1) = 1$ and $g(y_1|x_1^2) = 0.2$, respectively. In $t = 2$, they have equal observation likelihoods, $g(y_2|x_2^1) = g(y_2|x_2^2) = 0.2$.

Using a BPF and performing resampling at every time step, the classical resampling schemes and the LB resampler with π_w will not discriminate between x_2^1 and x_2^2 when resampling, as their importance weights will be equal, $w_2^1 = g(y_2|x_2^1) = g(y_2|x_2^2) = w_2^2 = 0.2$.

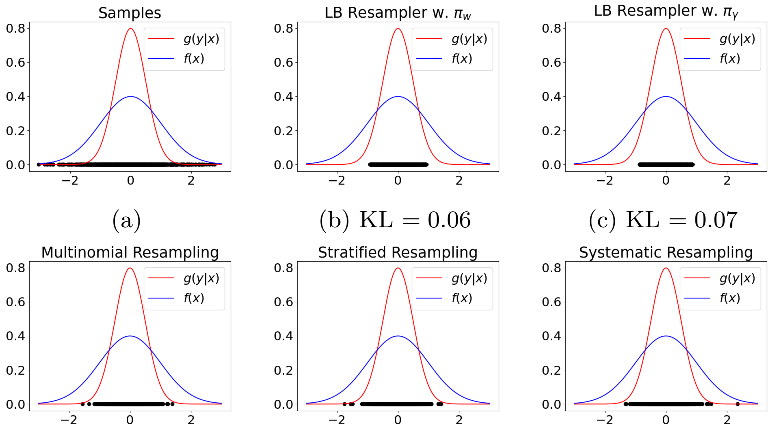
Meanwhile, the LB resampler with π_γ assigns higher weight to $x_{1:2}^1$, since

$$\gamma_2^1 = g(y_2|x_2^1) f(x_2^1|x_1^1) g(y_1|x_1^1) f(x_1^1) = 0.2c^2,$$

is greater than

$$\gamma_2^2 = g(y_2|x_2^2) f(x_2^2|x_1^2) g(y_1|x_1^2) f(x_1^2) = 0.04c^2.$$

[1] Note that this also differs from a PF with adaptive resampling (Doucet et al., 2009, Section 3.5).



(a)

(b) KL = 0.06

(c) KL = 0.07

(d) KL = 0.42

(e) KL = 0.17

(f) KL = 0.12

Figure 3: The KL divergences between the resampled and the normalized importance weighted measure are reported in the captions of the subplots (lower is better). (a) $N = 1000$ samples from the proposal $k(x) = f(x)$. The remaining subplots display the resampled particles using (b) the LB resampler with π_w, (c) the LB resampler with π_γ, (d) multinomial resampling, (e) stratified resampling, and (f) systematic resampling. In a typical VI manner, the two variational resampling schemes (b-c) concentrate the mass of their measures around the mode of the posterior density.

As we will see in the subsequent experiments, this online smoothing property results in impressive smoothing performances in challenging state-space models.

4 RELATED WORK

Deterministic resampling. While resampling is a well-studied topic in particle filtering (Douc and Cappé, 2005; Hol et al., 2006; Li et al., 2015), few works have previously considered deterministic resampling schemes, notably Li et al. (2012). However, Li et al. (2012) proposes methods that are based on partitioning the state space in grids, which does not scale beyond very low-dimensional problems. Crisan and Lyons (2002) presented a theoretical analysis of an algorithm that minimizes the relative entropy similarly to our Algorithm 3, but the method (i) is applicable only to discrete state spaces (ii) is randomized instead of deterministic and (iii) is not implemented in experiments.

Variational inference and particle filters. More broadly, previous works have considered variational lower bounds in the context of particle filters (e.g., Maddison et al. (2017); Naesseth et al. (2018); Le et al. (2018); Moretti et al. (2021)), however, all of these use traditional resampling schemes. In Saeedi et al. (2017), a variational objective is maximized by finding optimal positions and weights of a particle distribution, i.e., VI is not used as a resampling step. Grover et al. (2018) proposes a method with a similar name to ResELBO, but their method uses accept/reject steps. Futhermore, backward simulation methods (Lindsten et al., 2013) can be constructed to improve the smoothing approximations, however these require post-hoc (offline) analysis of PFs and are not alternatives to resampling schemes.

Biased resampling. Notably, many recent works have considered schemes that introduce a bias in the estimates of the marginal likelihood either usinga biased resampling scheme or replacing resampling with another operation, for example optimal transport based resampling (Reich, 2013; Corenflos et al., 2021; Li et al., 2022), nudging (Akyildiz and Míguez, 2020) or measure transport maps (Arbel et al., 2021; Maken et al., 2022) (in particular, Arbel et al. (2021) corrects for the bias using separate sets of particles).

5 EXPERIMENTS

Here we conduct a series of experiments highlighting the superiority of variational resampling with respect to different metrics. We compare the LB resampler with π_w (Sec. 3.2.1) and with π_γ (Sec. 3.2.2) to multinomial resampling and the state-of-the art resampling schemes, systematic and stratified.

First, we provide toy experiments where we explore the properties of the variational resamplers. We then examine the smoothing estimation qualities of PFs using variational resampling on challenging state-space models in three settings: the stochastic volatility model with and without real data, and the Lorenz 63 model.

The code to reproduce our results are publicly available at GitHub: https://github.com/okviman/Variational-Resampling.

5.1 Non-Sequential Toy Experiments

Here we are considering the following posterior density,

$$p(x|y=0) \propto g(y=0|x)f(x), \tag{29}$$

where $g(y=0|x) = \mathcal{N}(y=0|x, 0.25)$ and $f(x) = \mathcal{N}(x|0,1)$. In accordance with the rest of this work, we set $k(x) = f(x)$.

In Fig. 3, we visualize the distribution of the resampled particles when using the different resampling schemes, and report the corresponding KL divergences between the resampled and the normalized importance weighted measure (see Eq. (23)). Interestingly, the two variational resampling schemes—the LB resampler with π_γ or with π_w—both concentrate their distributions of resampled particles around the mode of the posterior density, in a typical ELBO-based VI fashion.

The LB resampler with π_w achieves the best KL scores. As shown in Sec. 3.2.1, this resampler is indeed resampling particles in order to minimize this KL. As such, the scores verify empirically that our optimization algorithm is working as intended.

5.2 State-Space Models

First, let $p(dx_{1:T}|y_{1:T})$ be a ground truth smoothing measure generated by running a BPF with multinomial resampling and 50,000 particles, and $x^*_{1:T}$ be the true latent sequence. For all state-space models, we evaluate the two following smoothing-related mean-squared errors (MSEs). **MSE***: the MSE between the mean of the smoothing distribution from the PF and $x^*_{1:T}$, and **MSE**: the MSE between the smoothing distribution from the PF and the mean of $p(dx_{1:T}|y_{1:T})$.

Additionally, following Chopin et al. (2020, Chapter 9, Section 9.7), we evaluate the TV distance between the resampled measures and the normalized importance weighted measures at every time step,

$$\mathrm{TV}(q_{\phi^{1:N}}, \pi_{\widetilde{w}^{1:N}}) = \tag{30}$$
$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\sum_{x_t \in S[\pi_{\widetilde{w}_t^{1:N}}]}\left|\sum_{i=1}^{N}\frac{\phi_t^i}{N}\delta_{x_t^i}(x_t) - \sum_{i=1}^{N}\widetilde{w}_t^i\delta_{x_t^i}(x_t)\right|,$$

where $S[\pi_{\widetilde{w}_t^{1:N}}]$ denotes the support of the normalized importance weighted measure. As we will demonstrate, our results are aligned with the findings in Chopin et al. (2020) that systematic produces small TV distances compared to the existing schemes. However, our LB resampler with π_w outperforms systematic, advancing the state-of-the-art with respect to this metric.

Finally, in the real-data experiment, we also investigated the marginal log-likelihood scores produced by our algorithms. Relative to the estimates from the PF using stratified, the estimates from our two schemes are indeed similar to those from an expensive PF.

All reported results are averages of ten independent runs. In all our state-space model experiments, we found that the resampling schemes performed the same w.r.t. effective sample size. Additional experimental results and details are provided in the Supplementary Material, where it is evident that the results are consistent across multiple seeds.

5.2.1 Stochastic Volatility Model

The stochastic volatility (SV) model is a common model in financial econometrics which is popular to use for benchmarking PFs (Doucet et al., 2009; Lindsten et al., 2014; Naesseth et al., 2018). The model is formulated as follows

$$g(y_t|x_t^i) = \mathcal{N}\left(y_t|0, \beta^2 e^{x_t^i}\right) \tag{31}$$

$$f(x_t^i|x_{t-1}^i) = \mathcal{N}\left(x_t^i|m x_{t-1}^i, \sigma^2\right) \tag{32}$$

$$p(x_1^i) = \mathcal{N}\left(x_1^i\middle|0, \frac{\sigma^2}{1-m^2}\right). \tag{33}$$

We first test the performance of the resampling methods when generating observations from the model for a given set of parameters. This allows us to evaluate MSE*.

Then we follow Lindsten et al. (2014) and experiment on Standard and Poor's (S&P) 500 data between 2006-04-03 and 2014-03-31, i.e. $T = 2011$. We run an expensive grid search using a BPF with multinomial resampling and $N = 10,000$, finding the set of parameters yielding the highest marginal log-likelihood estimates. Using these parameters, we evaluate the performances of the resampling schemes. This real data experiment will inform us whether the LB resamplers are useful for parameterized models used in practice.

The Synthetic Data-Generation Setting We parameterize the model by setting $(\sigma, \beta, m) = (1, 0.5, 0.91)$, following Doucet et al. (2009, Section 2.1). The results are presented in Table 1. Our LB resampler with π_γ outperforms the other methods with ten times less particles, with respect to the smoothing-related MSEs. Meanwhile, the LB resampler with π_w produces the smallest TV distances.

Resampling scheme	$N = 100$		$N = 1000$		
	MSE*	MSE	MSE*	MSE	TV
LB resampler w. π_w	1.79	0.91	1.17	0.35	**0.13**
LB resampler w. π_γ	**1.14**	**0.22**	**1.09**	**0.15**	0.29
Multinomial	1.75	0.87	1.40	0.51	0.37
Systematic	1.67	0.79	1.18	0.29	0.16
Stratified	1.66	0.79	1.18	0.29	0.21

Table 1: Results on the **SV** model when $T = 1000$ and $(\sigma, \beta, m) = (1, 0.5, 0.91)$. The reported scores are averaged over ten runs for the same sequence of observations. MSE* denotes the MSE between the mean of the smoothing distribution from the PF and the true data-generating latent sequence, and MSE denotes the MSE between the smoothing distribution from the PF and the mean of $p(dx_{1:T}|y_{1:T})$. Lower scores are better and in bold font.

Resampling scheme	$N = 100$	$N = 1000$	
	MSE	MSE	TV
LB resampler w. π_w	0.82	0.34	**0.13**
LB resampler w. π_γ	**0.20**	**0.16**	0.28
Multinomial	0.82	0.59	0.37
Systematic	0.74	0.31	0.16
Stratified	0.73	0.30	0.21

Table 2: Results on the S&P 500 data using the **SV** model when $T = 2011$. The reported scores are averaged over ten runs for the same sequence of observations. Lower scores are better.

The S&P 500 Data Setting As described above, we ran a coarse grid search over 252 parameter combinations to find the SV model parameters that best fit the real data. The tested parameters in the grid search are shown in the Supplementary Material. Given the optimal parameters, we evaluated the MSE (note that MSE* is not applicable here as we do not have access to the true latent sequence) and TV distance, and the results are shown in Table 2.

For this real data experiment, the LB resampler with π_γ produces impressive smoothing MSE performances, again requiring a factor ten less particles to match the other resamplers. The LB resampler with π_w achieves the best TV to the normalized importance measure.

Furthermore, we computed the marginal log-likelihood scores, i.e. $\log \widehat{Z}_{1:T}$, of PFs using our LB resamplers, and using stratified resampling. The scores were then compared to that of an $N = 50,000$ PF with multinomial resampling. The other algorithms used $N = 1000$, and the results are shown in Table 3. Our resamplers here produce $\log \widehat{Z}_{1:T}$ scores that are very close to the ground truth, given the score of systematic. The results indicate that although the LB resamplers are deterministic, this is not necessarily prohibitive in practice.

Resampling scheme	$\log \widehat{Z}_{1:T}$
LB resampler w. π_w	5473.37
LB resampler w. π_γ	5470.68
Stratified	5468.77
Ground truth	5469.39

Table 3: Results on the S&P 500 data using the **SV** model when $T = 2011$. Comparison between our methods and the PF using stratified resampling ($N = 1000$ for the three methods). The ground truth is a $N = 50,000$ PF with multinomial resampling.

5.3 Lorenz 63 system

The Lorenz 63 dynamical system is a well-established benchmark for stochastic nonlinear filters in the metereological sciences (Van Leeuwen, 2009, 2010), and for PF methodologies, more generally (Crisan et al., 2018; Branchini and Elvira, 2024). The system has a three-dimensional latent space whose temporal evolution is described by the following stochastic differential equations (SDEs)

$$\mathrm{d}x_1 = \sigma(x_2 - x_1)\mathrm{d}t + \mathrm{d}w_1, \tag{34}$$

$$\mathrm{d}x_2 = (x_1(\rho - x_3) - x_2)\mathrm{d}t + \mathrm{d}w_2, \tag{35}$$

$$\mathrm{d}x_3 = (x_1 x_2 - \beta x_3)\mathrm{d}t + \mathrm{d}w_3, \tag{36}$$

where t denotes continuous time and $\{w_i(s)\}_{s\in(0,\infty)}$ for $i = 1, 2, 3$ are one-dimensional independent independent Wiener processes. As common in the PF literature, we discretize the SDEs following the Euler-Maruyama method. This is a challenging model due to its choatic nature.

The LB resampler with π_γ consistently outperforms the other resamplers using ten times less particles in terms of both MSE* and MSE. See the results in Table 4. Once again, the LB resampler with π_w achieves the smallest TV distance.

Resampling scheme	$N = 100$		$N = 1000$		
	MSE*	MSE	MSE*	MSE	TV
LB resampler w. π_w	642.97	323.25	485.45	234.54	**0.12**
LB resampler w. π_γ	**339.65**	**85.93**	**276.80**	**85.42**	0.44
Multinomial	634.24	333.55	658.51	269.79	0.34
Systematic	886.57	417.58	549.51	244.40	0.15
Stratified	594.01	394.56	388.95	247.29	0.19

Table 4: Results on the **Lorenz 63** model when $T = 1000$. The reported scores are averaged over ten runs for the same sequence of observations. MSE* denotes the MSE between the mean of the smoothing distribution from the PF and the true data-generating latent sequence, and MSE denotes the MSE between the smoothing distribution from the PF and the mean of $p(dx_{1:T}|y_{1:T})$. Lower scores are better and in bold font.

6 DISCUSSION

As stated in Sec. 3.1, the LB resampler runs with time complexity $\mathcal{O}(N \log N)$, while multinomial resampling runs in $\mathcal{O}(N)$. The results in the previous section, however, show that our algorithms require orders of magnitude less particles to outperform the baselines.

Variational resampling is here justified via empirical results and compelling ideas of bridging popular methodologies. Devising guarantees such as convergence rates in the VI methodology is an active field of research (Domke et al., 2024; Kim et al., 2024; Hotti et al., 2024), and so characterization of the biasedness of the LB resamplers (stemming from their deterministic nature) was outside the scope of this work. Nonetheless, establishing theoretical guarantees is an important future-work direction.

Other exciting research directions include the application of variational resampling to resampling-based adaptive IS methods (Bugallo et al., 2017; Elvira et al., 2017; Elvira and Chouzenoux, 2022), or mixture-learning in black-box VI (Kviman et al., 2022, 2023a,b) where it could be used to infer the mixture weights.

7 CONCLUSION

We have proposed a variational resampling, a new paradigm for developing resampling schemes in PFs by casting the resampling step as a VI problem. We design two variational resamplers that outperform the baselines in terms of 1) producing resampled measures with smaller TV distance to the corresponding normalized importance weighted measures, or 2) resulting in PFs with superior smoothing estimation performances, respectively. Additionally, we derive a novel lower bound, induced by the resampled measure, on the estimated marginal log-likelihood.

We limited our work to analyses based on the performances of the resampling schemes using BPFs and resampling at every time step. Apart from extending the analyses to more advanced PF algorithms, there is plenty of interesting future work in terms of developing a more theoretical understanding of variational resampling. New target measures and optimization algorithms can be constructed, and other divergences than the KL can be considered.

Acknowledgements

We acknowledge the very insightful comments provided by the reviewers, which have helped improve our work. This project was made possible through funding from the Swedish Foundation for Strategic Research grants BD15-0043 and ID19-0052, and from the Swedish Research Council grant 2018-05417_VR.

References

Ö. D. Akyildiz and J. Míguez. Nudging the particle filter. *Statistics and Computing*, 30:305–330, 2020.

M. Arbel, A. Matthews, and A. Doucet. Annealed flow transport monte carlo. In *International Conference on Machine Learning*, pages 318–330. PMLR, 2021.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877, 2017.

A. Bouchard-Côté, S. Sankararaman, and M. I. Jordan. Phylogenetic inference via sequential monte carlo. *Systematic biology*, 61(4):579–593, 2012.

N. Branchini and V. Elvira. An adaptive mixture view of particle filters. *Foundations of Data Science*, 2024.

M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

N. Chopin, O. Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

A. Corenflos, J. Thornton, G. Deligiannidis, and A. Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, pages 2100–2111. PMLR, 2021.

D. Creal. A survey of sequential monte carlo methods for economics and finance. *Econometric Reviews*, 31(3):245–296, 2012. doi: 10.1080/07474938.2011.607333.

D. Crisan and T. Lyons. Minimal entropy approximations and optimal algorithms. 8(4):343–356, 2002. doi: doi:10.1515/mcma.2002.8.4.343. URL https://doi.org/10.1515/mcma.2002.8.4.343.

D. Crisan, J. Míguez, and G. Ríos-Muñoz. On the performance of parallelisation schemes for particle filtering. *EURASIP Journal on Advances in Signal Processing*, 2018:1–18, 2018.

J. Domke, R. Gower, and G. Garrigos. Provable convergence guarantees for black-box variational inference. *Advances in Neural Information Processing Systems*, 36, 2024.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *Ispa 2005. proceedings of the 4th international symposium on image and signal processing and analysis, 2005.*, pages 64–69. IEEE, 2005.

A. Doucet, A. M. Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

V. Elvira and E. Chouzenoux. Optimized population monte carlo. *IEEE Transactions on Signal Processing*, 70:2489–2501, 2022.

V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Improving population monte carlo: Alternative weighting and resampling schemes. *Signal Processing*, 131:77–91, 2017.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET, 1993.

A. Grover, R. Gummadi, M. Lazaro-Gredilla, D. Schuurmans, and S. Ermon. Variational rejection sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 823–832. PMLR, 2018.

J. D. Hol, T. B. Schon, and F. Gustafsson. On resampling algorithms for particle filters. In *2006 IEEE nonlinear statistical signal processing workshop*, pages 79–82. IEEE, 2006.

A. Hotti, L. Van der Goten, and J. Lagergren. Benefits of non-linear scale parameterizations in black box variational inference through smoothness results and gradient variance bounds. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2024.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

K. Kim, J. Oh, K. Wu, Y. Ma, and J. Gardner. On the convergence of black-box variational inference. *Advances in Neural Information Processing Systems*, 36, 2024.

H. Koptagel, O. Kviman, H. Melin, N. Safinianaini, and J. Lagergren. Vaiphy: a variational inference based algorithm for phylogeny. *Advances in Neural Information Processing Systems*, 35:14758–14770, 2022.

O. Kviman, H. Melin, H. Koptagel, V. Elvira, and J. Lagergren. Multiple importance sampling elbo and deep ensembles of variational approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 10687–10702. PMLR, 2022.

O. Kviman, R. Molén, A. Hotti, S. Kurt, V. Elvira, and J. Lagergren. Cooperation in the latent space: The benefits of adding mixture components in variational autoencoders. In *International Conference on Machine Learning*, pages 18008–18022. PMLR, 2023a.

O. Kviman, R. Molén, and J. Lagergren. Improved variational bayesian phylogenetic inference using mixtures. *arXiv preprint arXiv:2310.00941*, 2023b.

T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood. Auto-encoding sequential monte carlo. In *International Conference on Learning Representations*, 2018.

T. Li, T. P. Sattar, and S. Sun. Deterministic resampling: unbiased sampling to avoid sample impoverishment in particle filters. *Signal Processing*, 92(7):1637–1645, 2012.

T. Li, M. Bolic, and P. M. Djuric. Resampling methods for particle filtering: classification, implementation, and strategies. *IEEE Signal processing magazine*, 32(3):70–86, 2015.

Y. Li, W. Wang, K. Deng, and J. S. Liu. Stratification and optimal resampling for sequential monte carlo. *Biometrika*, 109(1):181–194, 2022.

F. Lindsten, T. B. Schön, et al. Backward simulation methods for monte carlo statistical inference. *Foundations and Trends® in Machine Learning*, 6(1):1–143, 2013.

F. Lindsten, M. I. Jordan, and T. B. Schon. Particle gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15:2145–2184, 2014.

C. J. Maddison, J. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. Teh. Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4018–4025. IEEE, 2023.

F. A. Maken, F. Ramos, and L. Ott. Stein particle filter for nonlinear, non-gaussian state estimation. *IEEE Robotics and Automation Letters*, 7(2):5421–5428, 2022.

A. K. Moretti, L. Zhang, C. A. Naesseth, H. Venner, D. Blei, and I. Pe'er. Variational combinatorial sequential monte carlo methods for bayesian phylogenetic inference. In *Uncertainty in Artificial Intelligence*, pages 971–981. PMLR, 2021.

C. Naesseth, S. Linderman, R. Ranganath, and D. Blei. Variational sequential monte carlo. In *International conference on artificial intelligence and statistics*, pages 968–977. PMLR, 2018.

J. A. Placed, J. Strader, H. Carrillo, N. Atanasov, V. Indelman, L. Carlone, and J. A. Castellanos. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, 2023.

S. Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

A. Saeedi, T. D. Kulkarni, V. K. Mansinghka, and S. J. Gershman. Variational particle approximations. *The Journal of Machine Learning Research*, 18(1):2328–2356, 2017.

G. Storvik, A. D.-L. Palomares, S. Engebretsen, G. O. Isaksson Rø, K. Engø-Monsen, A. B. Kristoffersen, B. F. de Blasio, and A. Frigessi. A sequential Monte Carlo approach to estimate a time varying reproduction number in infectious disease models: the Covid-19 case. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnad043, 04 2023. ISSN 0964-1998. doi: 10.1093/jrsssa/qnad043. URL https://doi.org/10.1093/jrsssa/qnad043.

S. Thrun, D. Fox, W. Burgard, et al. Monte carlo localization with mixture proposal distribution. In *AAAI/IAAI*, pages 859–865, 2000.

S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust monte carlo localization for mobile robots. *Artificial intelligence*, 128(1-2):99–141, 2001.

P. J. Van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

P. J. Van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

Variational Resampling
Supplementary Materials

A ADDITIONAL RESULTS AND EXPERIMENTAL DETAILS

Here we include some additional results and details about the experiment setups, starting with those relating to the stochastic volatility model.

A.1 Stochastic Volatility Model

In Table 5 we show the estimation results when $T = 500$. As expected, all estimation scores improve with the number of particles. However, note that our LB resampler with π_γ outperforms the other resampling schemes with ten times less particles ($N = 10$ vs. $N = 100$).

In Table 6 we present the results of the same experiment setup as in the main text, but with another random seed for generating the data. The outcome of the experiment is the same.

To find the model parameters for the real data experiment we ran a grid search (as described in the main text) over the following parameters

$$m \in (0.01, 0.1, 0.2, 0.5, 0.6, 0.8, 0.9),$$

$$\beta \in (0.01, 0.1, 0.5, 1, 1.5, 2),$$

and

$$\sigma \in (0.01, 0.1, 0.5, 1, 1.5, 2).$$

All 252 possible combinations were tested, and the best combination was $(\sigma, \beta, m) = (1, 0.01, 0.8)$, measured by the marginal log-likelihoods estimated by the $N = 10,000$ PFs with multinomial resampling.

The S&P 500 data is publicly available here https://finance.yahoo.com/quote/%5EGSPC/history?ltr=1, but can also be found in our GitHub repository.

A.2 Lorenz 63 Model

We reran the the experiment setup in the main text using another random seed for generating the data, shown in Table 7. We draw the same conclusions here as we did in the main text—our LB resampler with π_γ gives superior performance w.r.t. the MSEs (again, with ten times less particles), while the LB with π_w achieves the smallest TV distance.

In the experiment in the main text all PFs had averaged ESS scores (Doucet et al., 2009, Section 3.5) of 622 for the variational resamplers or 621 for the others. In the experiment given here, the variational resamplers also had marginally higher average ESS scores (all scores were between 624 and 626).

Resampling scheme	$N = 10$		$N = 100$		$N = 1000$	
	MSE*	MSE	MSE*	MSE	MSE*	MSE
LB resampler w. π_w	2.19	1.34	1.62	0.73	1.16	0.22
LB resampler w. π_γ	**1.45**	**0.55**	**1.19**	**0.21**	**1.12**	**0.15**
Multinomial	2.11	1.19	1.76	0.79	1.27	0.29
Systematic	1.99	1.04	1.60	0.64	1.13	0.17
Stratified	2.02	1.07	1.64	0.67	1.13	0.17

Table 5: Results on the SV model when $T = 500$ and $(\sigma, \beta, m) = (1, 0.5, 0.91)$. The reported scores are averaged over ten runs for the same sequence of observations. Lower scores are better.

Resampling scheme	$N = 100$		$N = 1000$		
	MSE*	MSE	MSE*	MSE	TV
LB resampler w. π_w	1.76	0.88	1.26	0.40	**0.13**
LB resampler w. π_γ	**1.13**	**0.24**	**1.07**	**0.18**	0.29
Multinomial	1.68	0.84	1.39	0.52	0.36
Systematic	1.68	0.80	1.25	0.37	0.16
Stratified	1.65	0.80	1.24	0.37	0.21

Table 6: Results on the **SV** model when $T = 1000$ and $(\sigma, \beta, m) = (1, 0.5, 0.91)$ using another random for generating the data. The reported scores are averaged over ten runs for the same sequence of observations. MSE* denotes the MSE between the mean of the smoothing distribution from the PF and the true data-generating latent sequence, and MSE denotes the MSE between the smoothing distribution from the PF and the mean of $p(dx_{1:T}|y_{1:T})$. Lower scores are better and in bold font.

Resampling scheme	$N = 100$		$N = 1000$		
	MSE*	MSE	MSE*	MSE	TV
LB resampler w. π_w	528.57	403.25	377.67	181.92	**0.12**
LB resampler w. π_γ	**335.09**	**67.94**	**331.14**	**50.17**	0.44
Multinomial	470.08	284.44	426.29	196.77	0.34
Systematic	557.66	229.65	401.15	193.15	0.15
Stratified	621.02	438.56	544.10	200.49	0.19

Table 7: Results on the **Lorenz 63** model when $T = 1000$ using another random seed for generating the data. The reported scores are averaged over ten runs for the same sequence of observations. MSE* denotes the MSE between the mean of the smoothing distribution from the PF and the true data-generating latent sequence, and MSE denotes the MSE between the smoothing distribution from the PF and the mean of $p(dx_{1:T}|y_{1:T})$. Lower scores are better and in bold font.

In Fig. 4 we visualize the 3D latent sequence used in the experiment in the main text, whose parameters we describe below.

Lorenz 63 parameters. The Euler-Maruyama discretization of the Lorenz63 model leads to the following state-space model with paramters s, ρ, β,

$$\begin{aligned}
x^{(1)}_{t+\Delta t} &= x^{(1)}_t + s(x^{(2)}_t - x^{(1)}_t)\Delta t + \epsilon_1 \\
x^{(2)}_{t+\Delta t} &= x^{(2)}_t + (x^{(1)}_t(\rho - x^{(3)}_t) - x^{(2)}_t)\Delta t + \epsilon_2 \\
x^{(3)}_{t+\Delta t} &= x^{(3)}_t + (x^{(1)}_t x^{(2)}_t - \beta x^{(3)}_t)\Delta t + \epsilon_3,
\end{aligned}$$

where $\epsilon_1, +\epsilon_2, +\epsilon_3$ are all (independent) additive Gaussian noises whose variance depends on Δt, $\mathcal{N}(0, \sigma_x^2(\Delta t))$. Specifically, we set $\sigma_x^2(\Delta t) = \frac{1}{2}\Delta t$. We obtain observations, as typical in experiments on the Lorenz 63 system, by observing the first coordinate corrupted by noise as

$$y_{t+\Delta t} = x^{(1)}_{t+\Delta t} + \epsilon_y, \tag{37}$$

where $\epsilon_y \sim \mathcal{N}(0, \sigma_y^2)$ and we used. Finally, to obtain meaningful latent trajectories, we use the following equations relating total continuous time $\mathcal{T}$ and discrete time T (length of the time series that we will use to generate data)

$$T = \mathcal{T}/\Delta t. \tag{38}$$

Therefore, the full set of parameters we used is $\{\sigma = 10, \beta = 8/3, \rho = 28, \Delta t = 0.01, \mathcal{T} = 10\}$.

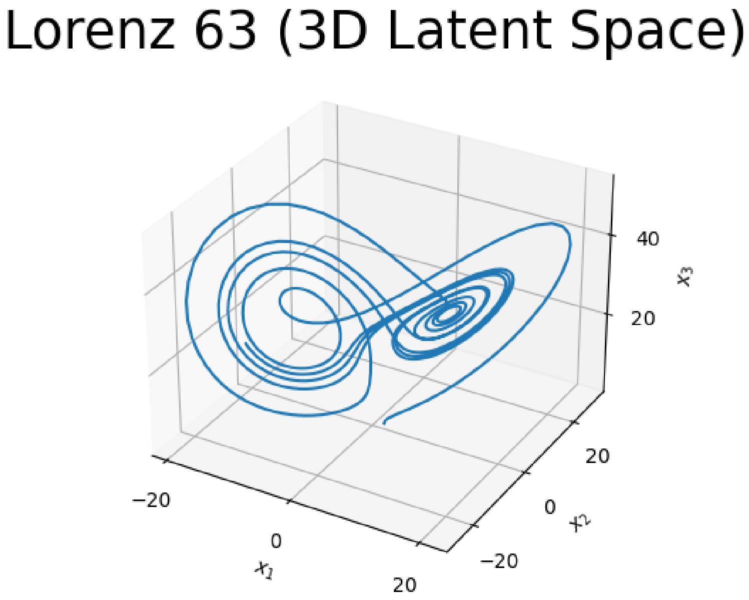


Figure 4: Visualization of the Lorenz 63 3D latent space realization used in the experiments in the main text.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Not Applicable**
   (b) Complete proofs of all theoretical results. **Not Applicable**
   (c) Clear explanations of any assumptions. **Not Applicable**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. **Not Applicable**

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**