# Differentiable Particle Filtering

Adrien Corenflos[1]
Aalto University
adrien.corenflos@aalto.fi

James Thornton[1]
Oxford University
james.thornton@spc.ox.ac.uk

Arnaud Doucet
Oxford University
doucet@stats.ox.ac.uk

August 27, 2020

**Abstract**

Particle Filters (PF) are a powerful class of methods for performing state inference in state-space models and for computing likelihood estimates for fixed parameters. Resampling is a key ingredient of PF, necessary to obtain low variance estimates. However, resampling operations result in the simulated likelihood function being non-differentiable with respect to parameters, even if the true likelihood is itself differentiable. These resampling operations also yield high variance gradient estimates of the Evidence Lower Bound (ELBO) when performing variational inference. By leveraging Optimal Transport (OT) ideas, we introduce differentiable PF, providing a differentiable simulated likelihood function. This allows one to perform parameter estimation via maximization of the simulated likelihood using gradient techniques and to compute low variance gradient estimates for variational inference. We demonstrate the performance of differentiable PF on various examples.

## 1 Introduction

In this section, we provide here a brief introduction to state-space modelling and PF. The key resampling operations provide both a non-differentiable estimate of the likelihood function and high-variance estimates of the ELBO gradients. We will review previous attempts to address these problems, outline their limitations and our contributions.

### 1.1 State-Space Models

State-space models (SSM) are a broad and expressive class of time-series models, used in numerous scientific domains including econometrics, ecology and robotics; see e.g. [17, 31, 51]. SSM may be characterized by a latent $\mathcal{X}$-valued discrete time Markov process $(X_t)_{t\geq 1}$ and a $\mathcal{Y}$-valued observation process $(Y_t)_{t\geq 1}$ satisfying $X_1 \sim \mu_\theta(\cdot)$ and for $t \geq 1$

$$X_{t+1}|\{X_t = x\} \sim f_\theta(\cdot|x), \quad Y_t|\{X_t = x\} \sim g_\theta(\cdot|x), \tag{1}$$

where $\theta \in \Theta$ is a parameter of interest. Given realized observations $(y_t)_{t\geq 1}$ and some parameter values $\theta$, one may perform state inference at time $t$ by computing the posterior of $X_t$ given $y_{1:t} := (y_1, ..., y_t)$ which satisfies $p_\theta(x_1|y_0) := \mu_\theta(x_1)$ and for $t > 1$

$$p_\theta(x_t|y_{1:t-1}) = \int f_\theta(x_t|x_{t-1})p_\theta(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}, \; p_\theta(x_t|y_{1:t}) = \frac{g_\theta(y_t|x_t)p_\theta(x_t|y_{1:t-1})}{\int g_\theta(y_t|x_t)p_\theta(x_t|y_{1:t-1})\mathrm{d}x_t}.$$

$p_\theta(x_t|y_{1:t-1})$ and $p_\theta(x_t|y_{1:t})$ are known as the predictive and filtering distributions respectively. The corresponding log-likelihood $\ell(\theta) = \log p_\theta(y_{1:T})$ is then given by

$$\ell(\theta) =) \log p_\theta(y_{1:T}) = \sum_{t=1}^{T} \log p_\theta(y_t|y_{1:t-1}), \tag{2}$$

[1] Equal contribution

where

$$p_\theta(y_1|y_{1:0}) := \int g_\theta(y_1|x_1)\mu_\theta(x_1)\mathrm{d}x_1, \qquad p_\theta(y_t|y_{1:t-1}) = \int g_\theta(y_t|x_t)p_\theta(x_t|y_{1:t-1})\mathrm{d}x_t \quad t \geq 2. \tag{3}$$

The filtering distributions and log-likelihood functions are only available analytically for a very restricted class of SSM such as linear Gaussian models and when $\mathcal{X}$ is finite. For non-linear SSM, one needs to approximate these quantities.

## 1.2 Particle Filtering

PFs are Monte Carlo methods providing such approximations using $N$ weighted particles $(w_t^i, X_t^i)_{i\in[N]}$, here $[N] := \{1, ..., N\}$ [18]. $X_t^i \in \mathcal{X}$ denotes the value of the $i^{\text{th}}$ particle at time $t$ and weights $\mathbf{w}_t := (w_t^1, ..., w_t^N)$ satisfy $w_t^i \geq 0$, $\sum_{i=1}^N w_t^i = 1$. Particles are sampled according to proposal distributions $q_\phi(x_1|y_1)$ at time $t = 1$ and $q_\phi(x_t|x_{t-1}, y_t)$ at time $t \geq 2$. One often sets $\phi = \theta$ but this is not necessarily the case; see e.g. [33, 38, 41]. The PF relies on the 'incremental weights' assumed here to be strictly positive for all $t \geq 1$ to avoid unnecessary technicalities:

$$\omega_{\theta,\phi}(x_1, y_1) := \frac{p_\theta(x_1, y_1)}{q_\phi(x_1|y_1)}, \qquad \omega_{\theta,\phi}(x_{t-1}, x_t, y_t) := \frac{p_\theta(x_t, y_t|x_{t-1})}{q_\phi(x_t|x_{t-1}, y_t)} \quad \text{for} \quad t \geq 2,$$

where $p_\theta(x_1, y_1) := g_\theta(y_1|x_1)\mu_\theta(x_1)$ and $p_\theta(x_t, y_t|x_{t-1}) := g_\theta(y_t|x_t)f_\theta(x_t|x_{t-1})$. A generic PF is described in Algorithm 1. More sophisticated schemes such as the auxiliary PF [44] and iterated auxiliary PF [26] could also be considered but are omitted here for simplicity.

**Algorithm 1** Particle Filter

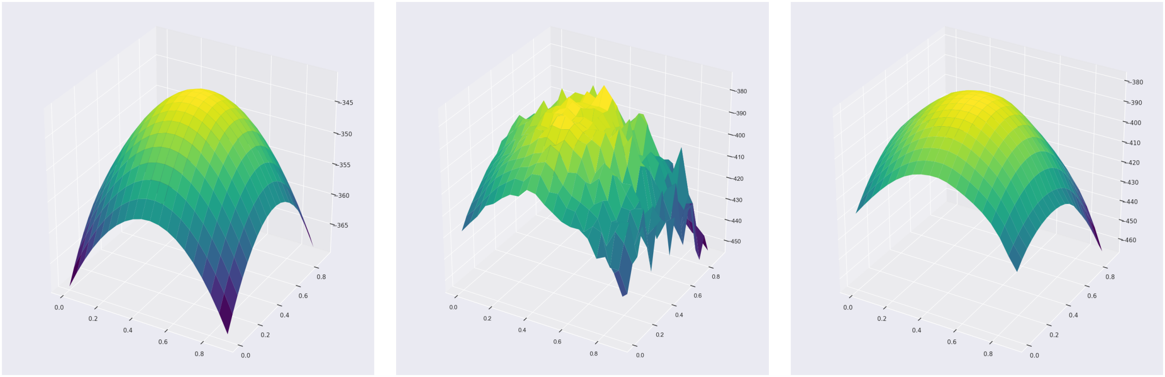
```
procedure PF($\theta$, $\phi$, $N$, $y_{1:T}$)
    Initialize:
    for $i \in [N]$ do
        Sample $X_1^i \overset{\text{i.i.d.}}{\sim} q_\phi(\cdot|y_1)$, set $w_1^i \propto \omega_{\theta,\phi}(X_1^i, y_1)$ and $\hat{\ell}(\theta) = \frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_1^i, y_1)$.
    for $t = 2, ..., T$ do
        Sample ancestors $\mathbf{A}_{t-1} \sim r(\cdot|\mathbf{w}_{t-1})$ and set $\tilde{X}_{t-1}^i = X_{t-1}^{A_{t-1}^i}$ for $i \in [N]$.
        for $i \in [N]$ do
            Sample $X_t^i \sim q_\phi(\cdot|\tilde{X}_{t-1}^i, y_t)$ and compute $w_t^i \propto \omega_{\theta,\phi}(\tilde{X}_{t-1}^i, X_t^i, y_t)$.
        $\hat{\ell}(\theta) \leftarrow \hat{\ell}(\theta) + \log \hat{p}_\theta(y_t|y_{1:t-1})$, where $\hat{p}_\theta(y_t|y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(\tilde{X}_{t-1}^i, X_t^i, y_t)$.
    return log-likelihood estimate $\hat{\ell}(\theta) = \log \hat{p}_\theta(y_{1:T})$.
```

Resampling operations may be characterized by a distribution $r(\cdot|\mathbf{w}_t)$ on $[N]^N$. For example, with the multinomial scheme, one has $r(\mathbf{A}_t|\mathbf{w}_t) = \prod_{i=1}^N \sum_{k=1}^N w_t^k \delta_k(A_t^i)$. Recall that $\tilde{X}_t^i = X_t^{A_t^i}$, a resampling scheme is said to be unbiased if

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \psi(\tilde{X}_t^i)\right] = \mathbb{E}\left[\sum_{i=1}^N w_t^i \psi(X_t^i)\right] \qquad \text{for any } \psi : \mathcal{X} \to \mathbb{R}. \tag{4}$$

All standard resampling schemes (multinomial, stratified, systematic) satisfy (4) [18]. This property guarantees $\exp(\hat{\ell}(\theta))$ is an unbiased estimate of the likelihood $\exp(\ell(\theta))$ for any $N$, and ensures computational efforts are focused on 'promising' regions of the latent space. Under weak assumptions, $\hat{\ell}(\theta)$ is a consistent estimate of $\ell(\theta)$ as $N \to \infty$ [15].

Henceforth, let $\mathcal{X} = \mathbb{R}^{d_x}$, $\theta \in \Theta = \mathbb{R}^{d_\theta}$ and $\phi \in \Phi = \mathbb{R}^{d_\phi}$. We will further assume that $\theta \mapsto \mu_\theta(x)$, $\theta \mapsto f_\theta(x'|x)$ and $\theta \mapsto g_\theta(y_t|x)$ are differentiable for all $x, x'$ and $t \in [T]$. These assumptions cover a very large class of SSM

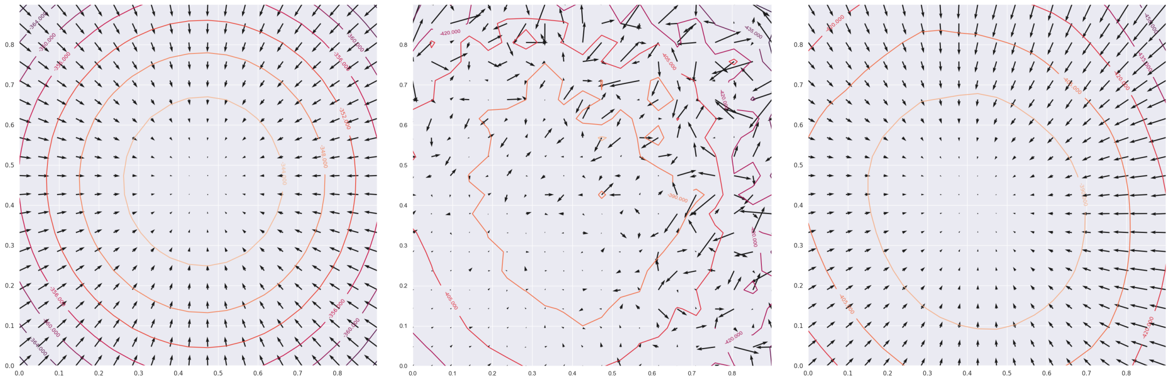


(a) Kalman Filter

(b) PF - Multinomial

(c) Differentiable PF

Figure 1: Log-likelihood $\ell(\theta)$ and PF estimates $\hat{\ell}(\theta; \mathbf{u})$ $(T = 150, N = 25)$ for linear Gaussian SSM



(a) Kalman Filter

(b) PF - Multinomial

(c) Differentiable PF

Figure 2: Vector fields for $\nabla_\theta \ell(\theta)$ and $\nabla_\theta \hat{\ell}(\theta; \phi, \mathbf{u})$ $(T = 150, N = 25)$ for linear Gaussian SSM

and imply that $\theta \mapsto \ell(\theta)$ is differentiable. We will also assume that we can use the reparameterization trick [23, 30] to sample from $q_\phi(\cdot)$; $X_1^i = \Xi_\phi(U_1^i, y_1) \sim q_\phi(\cdot|y_1)$ and $X_t^i = \Gamma_\phi(\theta, U_t^i, \tilde{X}_{t-1}^i, y_t) \sim q_\phi(\cdot|\tilde{X}_{t-1}^i, y_t)$ where $U_t^i \overset{\text{i.i.d.}}{\sim} \lambda$ with $\lambda$ independent of $(\theta, \phi)$ and denote $\mathbf{U} := \{U_t^i; i \in [N], t \in [T]\}$. To emphasize that $\hat{\ell}(\theta)$ depends on $\mathbf{U}$ and $\phi$, we will write $\hat{\ell}(\theta; \phi, \mathbf{U})$ for clarity. Finally, we will require $\phi \mapsto \Xi_\phi(u, y_1)$ and $\phi \mapsto \Gamma_\phi(\theta, u, x, y_t)$ be differentiable for $t \geq 2$ to ensure differentiability of $\hat{\ell}(\theta; \phi, \mathbf{U})$ with respect to (w.r.t) $\theta, \phi$.

## 1.3 Maximum Likelihood, Variational Inference and Contributions

A recent review of particle methods for parameter inference in SSM can be found in [28]. If $\ell(\theta)$ was available, one could estimate $\theta$ through Maximum Likelihood Estimation (MLE). As it is not, one may instead sample $\mathbf{U} = \mathbf{u}$, keep it fixed (i.e. use common random numbers) and then maximize $\theta \mapsto \hat{\ell}(\theta; \phi, \mathbf{u})$. For independent $(X_t)_{t \geq 1}$, i.e. $f_\theta(x'|x) = \mu_\theta(x')$, $\hat{\ell}(\theta; \phi, \mathbf{u})$ is differentiable and learning $\theta$ by maximizing this function is standard in econometrics [24, 35, 52]. However, for dependent $(X_t)_{t \geq 1}$, $\hat{\ell}(\theta; \phi, \mathbf{u})$ is only piecewise continuous and differentiable because the resampling steps introduce discontinuities in the particles that are selected across PFs when $\theta$ and $\phi$ vary; see Figure 1-b and Figure 2-b for a scenario where $d_x = 2$ and $d_\theta = 2$.

To make $\hat{\ell}(\theta; \phi, \mathbf{u})$ continuous w.r.t. $\theta$ and therefore easier to maximize, [43, 39] proposed to sort the particles and to then sample from a piecewise linear approximation of the resulting cumulative distribution function when $d_x = 1$. At a $O(N \log N)$ cost, this provides a continuous and piecewise differentiable function. Extensions [34] and [14] have been proposed for cases where $d_x > 1$. However, the scheme in [34] only outputs a piecewise continuous $\hat{\ell}(\theta; \phi, \mathbf{u})$, while the method in [14] has $O(N^2)$ complexity and requires using the same Gaussian proposal for all particles; i.e. $q_\phi(x_t|x_{t-1}, y_t) = \mathcal{N}(x_t; \mu_t^\phi, \Sigma_t^\phi)$. More recently, a modified resampling scheme has been proposed in [29, 36, 37]. The resulting PF-net is said to be differentiable but computes gradients that ignores the non-differentiable resampling scheme. The PF scheme in [27] is also said to be differentiable but ignores the resampling terms as acknowledged in [27, Section F].

An alternative approach to estimate $\theta$ consists of approximating the score vector using particle smoothing algorithms [28]. However, we will focus here on variational techniques which, contrary to these techniques, also allow us to learn the parameters $\phi$ of the "encoder". As $\exp(\hat{\ell}(\theta; \phi, \mathbf{U}))$ is an unbiased estimate of $\exp(\ell(\theta))$ for any $N \geq 1, \phi \in \Phi$ for standard resampling schemes then, by Jensen's inequality, one has

$$\ell^{\mathrm{ELBO}}(\theta, \phi) := \mathbb{E}_{\mathbf{U}}[\hat{\ell}(\theta; \phi, \mathbf{U})] \leq \ell(\theta). \tag{5}$$

The case $N = 1$ corresponds to the standard ELBO and, for specific classes of SSM, various variational distributions have been proposed; see e.g. [4, 32, 45]. Using $N > 1$ can be regarded as an extension of the popular Importance Weighted Auto-Encoders (IWAE) [7] particularly suited to dynamic models and can be used in combination with any of these methods. In particular, we have $\ell^{\mathrm{ELBO}}(\theta, \phi) \to \ell(\theta)$ as $N \to \infty$ even if the selected family of variational distributions is poor for the SSM considered. It is possible to maximize $\ell^{\mathrm{ELBO}}(\theta, \phi)$ w.r.t. both $\theta, \phi$ using stochastic gradient ascent as unbiased gradient estimates of the ELBO are available [33, 38, 41]. However these gradient estimates suffer from high variance however, as the terms corresponding to resampling steps cannot be computed using the reparameterization trick and rely instead on score estimators [23, 56]. Consequently, [33, 38, 41] use biased gradient estimates instead which ignore these resampling terms and report improvements as $N$ increases over standard variational approaches and IWAE. Nevertheless, because of the use of biased gradients, the objective optimized by these procedures is not well-defined and the impact of such gradients on the resulting parameter estimates remains unclear.

The contribution of this paper is three-fold.

- We show that the techniques ignoring the gradient terms due to resampling operations provide, even in the limit where $N \to \infty$, gradient estimates which can differ very significantly from $\nabla_\theta \ell(\theta)$.
- We propose the first class of differentiable PFs which do not ignore resampling terms but rely on differentiable resampling schemes derived from OT techniques. This provides a differentiable $\hat{\ell}(\theta; \phi, \mathbf{u})$; see Figure 1-c for an illustration. Using automatic differentiation, one can compute $\nabla_\theta \ell^{\mathrm{ELBO}}(\theta, \phi; \mathbf{u})$ to maximize this smooth function (see Figure 2-c) or $\nabla_{\theta,\phi} \ell^{\mathrm{ELBO}}(\theta, \phi; \mathbf{U})$ to maximize $\ell^{\mathrm{ELBO}}(\theta, \phi)$.
- We establish weak consistency results and demonstrate empirically the benefits of our approach.

The precise statements and proofs of results are given in the Appendices.

# 2 Resampling via Optimal Transport

## 2.1 Optimal Transport and the Wasserstein Metric

We present here the basics of OT [42, 54] at the core of our algorithms. Consider atomic probability measures $\alpha = \sum_{i=1}^N a_i \delta_{u_i}$ and $\beta = \sum_{i=1}^N b_i \delta_{v_i}$ on $\mathcal{X} = \mathbb{R}^{d_x}$ with weights $\mathbf{a} = (a_i)_{i \in [N]}$, $\mathbf{b} = (b_i)_{i \in [N]}$, and atoms $\mathbf{u} = (u_i)_{i \in [N]}$, $\mathbf{v} = (v_i)_{i \in [N]}$. In this case, the squared 2-Wasserstein metric between $\alpha$ and $\beta$ is given by

$$\mathrm{OT}(\alpha, \beta) = \min_{\mathcal{P} \in \mathcal{U}(\alpha, \beta)} \mathbb{E}_{(U,V) \sim \mathcal{P}} \left[ ||U - V||^2 \right] = \min_{\mathbf{P} \in \mathcal{S}(\mathbf{a}, \mathbf{b})} \textstyle\sum_{i=1}^N \sum_{j=1}^N c_{i,j} p_{i,j}, \tag{6}$$

where $\mathcal{U}(\alpha, \beta)$ the set of distributions on $\mathcal{X} \times \mathcal{X}$ with marginals $\alpha$ and $\beta$, any $\mathcal{P} \in \mathcal{U}(\alpha, \beta)$ is of the form $\mathcal{P}(\mathrm{d}u, \mathrm{d}v) = \sum_{i,j} p_{i,j} \delta_{u_i}(\mathrm{d}u) \delta_{v_j}(\mathrm{d}v)$, $c_{i,j} = ||u_i - v_j||^2$, $\mathbf{P} = (p_{i,j})_{i \in [N], j \in [N]}$ and $\mathcal{S}(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in [0,1]^{N \times N} | \sum_{j=1}^N p_{i,j} = a_i,\ \sum_{i=1}^N p_{i,j} = b_j\}$.

Any element $\mathcal{P} \in \mathcal{U}(\alpha, \beta)$ allows us to "transport" $\beta$ to $\alpha$ (and vice-versa), i.e.

$$\alpha(\mathrm{d}u) = \int \mathcal{P}(\mathrm{d}u, \mathrm{d}v) = \int \mathcal{P}(\mathrm{d}u|v)\beta(\mathrm{d}v), \quad \mathcal{P}(\mathrm{d}u|v) = \textstyle\sum_{i,j} b_i^{-1} p_{i,j} \delta_{v_i}(v) \delta_{u_j}(\mathrm{d}u). \tag{7}$$

The minimization problem (6) may be solved through linear programming at computational complexity $\mathcal{O}(N^3 \log N)$ [5]. Note that there is not necessarily a unique minimizing argument.

It is also possible to compute the squared 2-Wasserstein metric using a dual formulation

$$\mathrm{OT}(\alpha, \beta) = \max_{\mathbf{f}, \mathbf{g} \in \mathcal{R}(C)} \mathbf{a}^t \mathbf{f} + \mathbf{b}^t \mathbf{g}, \tag{8}$$

where $\mathbf{f} = (f_i)$, $\mathbf{g} = (g_i)$, $\mathbf{C} = (c_{i,j})$ and $\mathcal{R}(\mathbf{C}) = \{\mathbf{f}, \mathbf{g} \in \mathbb{R}^N | f_i + g_j \leq c_{i,j}, i, j \in [N]\}$.

### 2.2 Ensemble Transform Resampling

The use of OT for resampling in PF has been pioneered in [46]. Contrarily to standard resampling schemes, it relies not only on the particle weights but also on their locations. This method is motivated as follows. At time $t$, we obtain after the sampling step a particle approximation $\beta = \frac{1}{N} \sum \delta_{X_t^i}$ of the "prior" $q_{\phi,\theta}(x_t | y_{1:t}) := \int q_\phi(x_t | x_{t-1}, y_t) p_\theta(x_{t-1} | y_{1:t-1}) \mathrm{d}x_{t-1}$ while $\alpha = \sum w_t^i \delta_{X_t^i}$ approximates the posterior $p_\theta(x_t | y_{1:t})$. Under mild regularity conditions, the optimal transport between $q_{\phi,\theta}(x_t | y_{1:t})$ and $p_\theta(x_t | y_{1:t})$ is a (deterministic) transport map $T_t : \mathcal{X} \to \mathcal{X}$ such that if $X \sim q_{\phi,\theta}(x_t | y_{1:t})$ then $T_t(X) \sim p_\theta(x_t | y_{1:t})$.

It is shown in [46] how can one approximate this transport map by solving the OT problem (6) between $\beta$ and $\alpha$, then use the 'Ensemble Transform' (ET), i.e. set

$$\tilde{X}_t^i = \textstyle\sum_{k=1}^N r(A_t^i = k | \mathbf{w}_t) X_t^k = N \sum_{k=1}^N p_{i,k}^{\mathrm{OT}} X_t^k, \tag{9}$$

instead of sampling ancestors in Algorithm 1. Indeed, as $N \to \infty$, this provides a consistent approximation of $T$ [46, 48, 40].

We note that while $\mathbf{P}^{\mathrm{OT}} = (p_{i,j}^{\mathrm{OT}})$, like any element of $\mathcal{U}(\alpha, \beta)$, provides an unbiased resampling scheme defined by the resampling distribution $r(\mathbf{A}_t | \mathbf{w}_t) = \prod_{i=1}^N r(A_t^i | \mathbf{w}_t)$ with $r(A_t^i = k | \mathbf{w}_t) = N p_{i,k}^{\mathrm{OT}}$, the ET only satisfies (4) for affine functions $\psi$.

## 3 Differentiable Resampling via Regularized Optimal Transport

The OT approaches to resampling are attractive but their $O(N^3 \log N)$ cost is prohibitive. Moreover, the Wasserstein metric is not differentiable; see e.g. [13, Section 2].

### 3.1 Differentiable Regularized Optimal Transport

To address these problems, we rely instead on an entropy-regularized version of the squared 2-Wasserstein distance [10, 20, 42] defined for any $\epsilon > 0$ by

$$\mathrm{OT}_\epsilon(\alpha, \beta) = \min_{\mathbf{P} \in \mathcal{S}(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^N p_{i,j} \left( c_{i,j} + \epsilon \log \frac{p_{i,j}}{a_i b_j} \right). \tag{10}$$

The function minimized in (10) is strictly convex and thus only admits one single minimizing argument $\mathbf{P}_\epsilon^{\mathrm{OT}} = (p_{\epsilon,i,j}^{\mathrm{OT}})$. $\mathrm{OT}_\epsilon(\alpha, \beta)$ can also be computed using the regularized dual $\mathrm{DOT}_\epsilon(\mathbf{f}, \mathbf{g})$; i.e. $\mathrm{OT}_\epsilon(\alpha, \beta) = \max_{\mathbf{f}, \mathbf{g}} \mathrm{DOT}_\epsilon(\mathbf{f}, \mathbf{g})$ with

$$\mathrm{DOT}_\epsilon(\mathbf{f}, \mathbf{g}) := \mathbf{a}^t \mathbf{f} + \mathbf{b}^t \mathbf{g} - \epsilon \mathbf{a}^t \mathbf{M} \mathbf{b}, \quad \text{where} \quad (\mathbf{M})_{i,j} := \exp(\epsilon^{-1}(f_i + g_j - c_{i,j})) - 1, \tag{11}$$

where $\mathbf{f}, \mathbf{g}$ are now unconstrained. For the dual pair $(\mathbf{f}^*, \mathbf{g}^*)$ maximizing (11), one has

$$p^{\mathrm{OT}}_{\epsilon,i,j} = a_i b_j \exp\left(\epsilon^{-1}(f_i^* + g_j^* - c_{i.j})\right), \tag{12}$$

and $\nabla_{\mathbf{f},\mathbf{g}} \mathrm{DOT}_\epsilon(\mathbf{f},\mathbf{g})|_{(\mathbf{f}^*,\mathbf{g}^*)} = \mathbf{0}$. This first-order condition leads to

$$f_i^* = \mathcal{T}_\epsilon(\mathbf{b}, \mathbf{g}^*, \mathbf{C}_{i:}), \qquad g_i^* = \mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}^*, \mathbf{C}_{:i}), \tag{13}$$

where $\mathbf{C}_{i:}$ (resp. $\mathbf{C}_{:i}$) is the $i^{\mathrm{th}}$ row (resp. column) of $\mathbf{C}$. Here $\mathcal{T}_\epsilon : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ denotes the mapping $\mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}, \mathbf{C}_{:,i}) = -\epsilon \ \log \sum_k \exp\left\{\log a_k + \epsilon^{-1}(f_k - c_{k,i})\right\}$.

The dual can be maximized using the Sinkhorn algorithm introduced for OT in the seminal paper [10]. This algorithm has computational complexity $O(N^2)$ at each iteration and converges quickly [2]. The Sinkhorn has popularized the use of OT ideas in large-scale machine learning (ML) applications at it is is computationally cheaper than the $O(N^3 \log N)$ complexity of non-regularized OT and is particularly amenable to GPU implementation. We present in Algorithm 2 the implementation proposed in [20].[1]

**Algorithm 2** Sinkhorn algorithm

```
function POTENTIALS($\mathbf{a}, \mathbf{b}, \mathbf{u}, \mathbf{v}$)
    Local variables: $\mathbf{f}, \mathbf{g} \in \mathbb{R}^N$
    Initialize: $\mathbf{f} = \mathbf{0}$, $\mathbf{g} = \mathbf{0}$
    Set $\mathbf{C} \leftarrow \mathbf{u}\mathbf{u}^t + \mathbf{v}\mathbf{v}^t - 2\mathbf{u}\mathbf{v}^t$
    while stopping criterion not met do
        for $i \in [N]$ do
            $f_i \leftarrow \frac{1}{2}(f_i + \mathcal{T}_\epsilon(\mathbf{b}, \mathbf{g}, \mathbf{C}_{i:}))$
            $g_i \leftarrow \frac{1}{2}(g_i + \mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}, \mathbf{C}_{:i}))$
    return $\mathbf{f}, \mathbf{g}$
```

**Algorithm 3** DET Resampling

```
function ENSTRANSFORM($\mathbf{X}, \mathbf{w}, N$)
    $\mathbf{f}, \mathbf{g} \leftarrow$ POTENTIALS($\mathbf{w}, \frac{1}{N}\mathbf{1}, \mathbf{X}, \mathbf{X}$)
    for $i \in [N]$ do
        for $j \in [N]$ do
            $p^{\mathrm{OT}}_{\epsilon,i,j} = \frac{w_i}{N} \exp\left(\frac{f_i + g_j - c_{i,j}}{\epsilon}\right)$
    return $\tilde{\mathbf{X}} = N\mathbf{P}^{\mathrm{OT}}_\epsilon \mathbf{X}$
```

The pair of dual vectors $(\mathbf{f}^*, \mathbf{g}^*)$ solving (13) are differentiable with respect to $\mathbf{a}, \mathbf{b}, \mathbf{C}$ using the implicit function theorem; see [20] for details. This implies that the derivatives of $\mathbf{P}^{\mathrm{OT}}_\epsilon$ are readily accessible by combining the derivatives of (13) with the derivatives of (12), using automatic differentiation (autodiff).

This algorithm is particularly useful when evaluating the Sinkhorn divergence.

$$\mathrm{SD}_\epsilon(\alpha, \beta) = \mathrm{OT}_\epsilon(\alpha, \beta) - \tfrac{1}{2}\mathrm{OT}_\epsilon(\alpha, \alpha) - \tfrac{1}{2}\mathrm{OT}_\epsilon(\beta, \beta). \tag{14}$$

Unlike $\mathrm{OT}_\epsilon$, $\mathrm{SD}_\epsilon$ is a symmetric positive definite and differentiable loss function that is convex in each of its input variables; see [20, Theorem 1].

## 3.2 Differentiable Ensemble Transform Resampling

The ET (9) is not differentiable as it relies on OT. We obtain a differentiable ET (DET) by using the entropy-regularized OT computed by running Algorithm 3 using $(\mathbf{X}_t, \mathbf{w}_t, N)$:

$$\tilde{X}_t^i = N \textstyle\sum_{k=1}^N p^{\mathrm{OT}}_{\epsilon,k,i} X_t^k. \tag{15}$$

This was first suggested in [11] but it has never been realized before that this could be exploited to obtain a differentiable PF when combined to the reparameterization trick.

If we rescale $\mathbf{X}_t$ by a factor $\eta > 0$ then $\mathbf{P}^{\mathrm{OT}}$ is not impacted by this rescaling but this property does not hold for $\mathbf{P}^{\mathrm{OT}}_\epsilon$ which is undesirable. In practice, for $\mathbf{X}_t \in \mathbb{R}^{N \times d_x}$, we compute $\delta(\mathbf{X}_t) = \sqrt{d_x} \max_{k \in [d_x]} \mathrm{std}_i(X_{t,k}^i)$ and rescale accordingly $\mathbf{C}$ to ensure that $\epsilon$ is approximately independent of the scale and dimension of the problem. The DET does not satisfy (4) but the following result holds. For a measure $\gamma$ and test function $\psi$, we write $\gamma(\psi) = \int \gamma(\mathrm{d}x)\psi(x)$.

[1] Updates in Algorithm 2 are a variation on Equation (13) where the fixed point updates have been stabilized. If $\alpha = \beta$ then this converges faster than the standard Sinkhorn algorithm, for more details see Appendix 6.1

**Proposition 3.1.** Let $\mathbf{P}^{\mathrm{OT}}$, resp. $\mathbf{P}^{\mathrm{OT}}_{\epsilon}$, be the OT [2], resp. regularized OT, between random measures $\alpha_N = \sum_{i=1}^{N} w^i \delta_{X^i}$ and $\beta_N = \frac{1}{N}\sum_i \delta_{X^i}$. Let $\bar{\mathbf{X}} = N\mathbf{P}^{\mathrm{OT}}\mathbf{X}$, resp. $\tilde{\mathbf{X}} = N\mathbf{P}^{\mathrm{OT}}_{\epsilon}\mathbf{X}$, be the ET, resp. DET and $\bar{\alpha}_N = \frac{1}{N}\sum_i \delta_{\bar{X}^i}$ and $\tilde{\alpha}_N = \frac{1}{N}\sum_i \delta_{\tilde{X}^i}$. Then, under regularity conditions, for any bounded Lipschitz function $\psi$, there exists an increasing sequence $(N_k)_{k\geq 1}$ such that

$$\left|\mathbb{E}\Big[\bar{\alpha}_{N_k}(\psi) - \alpha_{N_k}(\psi)\Big]\right| \leq \kappa_{N_k}, \qquad \left|\mathbb{E}\Big[\tilde{\alpha}_{N_k}(\psi) - \alpha_{N_k}(\psi)\Big]\right| \leq \kappa_{N_k} + \mathcal{E}_{N_k,\epsilon}. \tag{16}$$

where $\lim_{k\to\infty} \kappa_{N_k} = 0$, $\lim_{\epsilon\to 0} \mathcal{E}_{N_k,\epsilon} = 0$ and the expectations are w.r.t. to the distribution of the random measures.

### 3.3 Discussion and Analysis

Given independent data $(Y_t)_{t\in[T]}$, it is standard in ML to estimate $\theta$ by finding the maximizer $\hat{\theta}_{\mathrm{ELBO}}$ of the ELBO (5) using stochastic gradient with mini-batches. This is computationally more efficient than finding the Simulated MLE (SMLE), which is the maximizer $\hat{\theta}_{\mathrm{SMLE}}$ of $\hat{\ell}(\theta;\phi,\mathbf{u})$ using (deterministic) full batch gradient as favored in econometrics [24, 35, 52][3]. However, we cannot use mini-batches for SSM. Each gradient iteration requires a pass over the whole dataset. Hence estimating $\hat{\theta}_{\mathrm{ELBO}}$ rather than $\hat{\theta}_{\mathrm{SMLE}}$ for SSM is not a cheaper computational alternative. We compare empirically these estimates in Section 4.

For standard resampling schemes, i.e. satisfying (4), one has $\ell^{\mathrm{ELBO}}(\theta,\phi) = \mathbb{E}_{\mathbf{U}}[\hat{\ell}(\theta;\phi,\mathbf{U})] \leq \ell(\theta)$. As mentioned previously, one can compute unbiased estimates of the gradient but these suffer from high variance so these gradient terms are typically omitted; see e.g. [33, 38, 41]. However, we show here that neglecting the gradient of the resampling terms yields gradient estimates significantly different from $\nabla_\theta \ell(\theta)$.

**Proposition 3.2.** Consider a PF with multinomial resampling where $\phi$ is distinct from $\theta$ then, under regularity assumptions, the ELBO gradient estimate omitting the resampling terms converges as $N \to \infty$ in probability to

$$\int \nabla_\theta \log p_\theta(x_1, y_1)\ p_\theta(x_1|y_1)\mathrm{d}x_1 + \sum_{t=2}^{T} \int \nabla_\theta \log p_\theta(x_t, y_t|x_{t-1})\ p_\theta(x_{t-1:t}|y_{1:t})\mathrm{d}x_{t-1:t} \tag{17}$$

whereas Fisher's identity gives

$$\nabla_\theta \ell(\theta) = \int \nabla_\theta \log p_\theta(x_1, y_1)\ p_\theta(x_1|y_{1:T})\mathrm{d}x_1 + \sum_{t=2}^{T} \int \nabla_\theta \log p_\theta(x_t, y_t|x_{t-1})\ p_\theta(x_{t-1:t}|y_{1:T})\mathrm{d}x_{t-1:t}. \tag{18}$$

It is clear that (17) and (18) only coincide if we had $p_\theta(x_{t-1:t}|y_{1:t}) = p_\theta(x_{t-1:T}|y_{1:T})$; i.e. for models where the latent variables are independent. When the SSM is mixing quickly so that future observations $y_{t+1:T}$ do not bring significant information about the latent state $X_t$ given $y_{t:T}$, (17) can be a decent approximation of (18). When (17) differs significantly from (18) such as for slow mixing SSM, there is no reason why a gradient algorithm using (17) would provide parameter estimates close to the MLE.

By using the differentiable resampling scheme presented before, we will demonstrate experimentally that we can obtain unbiased gradient estimates of significantly reduced variance. However, as this scheme does not satisfy (4), we cannot guarantee that $\ell^{\mathrm{ELBO}}_{\epsilon}(\theta,\phi) := \mathbb{E}_{\mathbf{U}}[\hat{\ell}_{\epsilon}(\theta;\phi,\mathbf{U})] \leq \ell(\theta)$ where $\epsilon$ is the DET parameter. However, the approach is principled in the sense of the proposition below.

**Proposition 3.3.** Under regularity assumptions, there exists an increasing sequence $(N_k)_{k\geq 1}$

$$\frac{1}{T}|\ell^{\mathrm{ELBO}}_{\epsilon}(\theta,\phi) - \ell(\theta)| \leq K\left[\frac{C}{N_k} + \bar{\kappa}_{N_k} + \bar{\mathcal{E}}_{N_k,\epsilon}\right],$$

where $C, K < \infty$ only depend on $\theta, \phi$ and $\bar{\kappa}_{N_k}, \bar{\mathcal{E}}_{N_k,\epsilon}$ constitute the maximum error term over $t \in [T]$ introduced by the DET with $\lim_{k\to\infty} \bar{\kappa}_{N_k} = 0$, $\lim_{\epsilon\to 0} \bar{\mathcal{E}}_{N_k,\epsilon} = 0$.

[2] If $\mathbf{P}^{\mathrm{OT}}$ is not unique, select the OT with maximal entropy.

[3] Strong guarantees are available for $\hat{\theta}_{\mathrm{SMLE}}$. It can be shown to be, like the MLE, asymptotically consistent and efficient if $N \propto T^{1/2+\gamma}$ for any $\gamma > 0$ [24, 35, 52].

In all our experiments, $|\ell_{\epsilon}^{\mathrm{ELBO}}(\theta,\phi)] - \ell^{\mathrm{ELBO}}(\theta,\phi)|$ is significantly smaller than $\ell(\theta) - \ell^{\mathrm{ELBO}}(\theta,\phi)$ even for moderate $N$ so $\ell_{\epsilon}^{\mathrm{ELBO}}(\theta,\phi) < \ell(\theta)$ hence we keep the ELBO terminology.

### 3.4 Extensions

We have focused on the use of the DET to obtain a differentiable resampling scheme. For large $\epsilon$, the DET can yield particles underestimating the variance of the posterior so [1] proposes to modify the DET to ensure that $\frac{1}{N}\sum_{i=1}^{N}\delta_{\tilde{X}_t^i}$ has the same covariance as $\sum_{i=1}^{N} w^i \delta_{X_t^i}$. In Appendix 6, we show how to compute a differentiable version of this scheme and provide a computationally cheaper alternative.

Alternatively, one may also attempt to find particle locations $\tilde{\mathbf{X}}_t = (\tilde{X}_t^i)_{i\in[N]}$ minimizing $\mathrm{SD}_{\epsilon}(\alpha,\beta)$ in (14) at time $t$ between $\alpha = \sum w_t^i \delta_{X_t^i}$ and $\beta = \frac{1}{N}\sum \delta_{\tilde{X}_t^i}$; i.e. find the best equally weighted distribution $\beta$ approximating $\alpha$ (which itself approximates $p_{\theta}(x_t|y_{1:t})$). Here $\mathbf{a} = \mathbf{w}_t$, $\mathbf{b} = \mathbf{1}/N$, $\mathbf{u} = \mathbf{X}_t$ and $\mathbf{v} = \tilde{\mathbf{X}}_t$. We recall that $\mathrm{SD}_{\epsilon}(\alpha,\beta)$ is a symmetric positive definite and differentiable loss function [20, Theorem 1] unlike $\mathrm{OT}_{\epsilon}$. This is a non-convex optimization problem but one could attempt to find a local minimum using a gradient procedure [42, Section 9.1.2]. Note that even if the non-convex optimization problem could be solved exactly, such resampling schemes do not satisfy (4). This approach is attractive but computationally expensive. Implementation details, including how to compute gradients $\nabla\tilde{\mathbf{X}}_t$, are provided in Appendix 6.

## 4 Experiments

The experiments presented here all use the DET (15) with $\epsilon = 0.5$, which provides a good trade-off between likelihood bias and stability of the gradient calculations. Further experiments using different parameters, alternative differentiable resampling techniques and other SSMs can be found in Appendix 6 and Appendix 10. The code to replicate the experiments can be found at `https://github.com/jtt94/filterflow`.

### 4.1 Linear Gaussian model

We first consider the following 2-dimensional linear Gaussian SSM for which exact inference can be carried out using Kalman techniques:

$$X_t|\{X_{t-1}=x\} \sim \mathcal{N}\left(\mathrm{diag}(\theta_1\ \theta_2)x, 0.5\mathbf{I}_2\right), \quad Y_t|\{X_t = x\} \sim \mathcal{N}(x, 0.1\cdot\mathbf{I}_2). \tag{19}$$

We simulate $T = 150$ observations using $\theta = (\theta_1,\theta_2) = (0.5, 0.5)$. Figure 1, displayed earlier, shows $\ell(\theta)$ obtained by Kalman and $\hat{\ell}(\theta;\mathbf{u})$ computed using PF with multinomial (MUL) and DET resampling for $N = 25$ particles using $q_{\phi}(x_t|x_{t-1},y_t) = f_{\theta}(x_t|x_{t-1})$. The corresponding gradient vector fields are given in Figure 2, where the gradient is computed using finite difference for PF MUL. We present the empirical mean and standard deviation (std) of $\frac{1}{T}(\hat{\ell}(\theta;\mathbf{U}) - \ell(\theta))$ in Table 1 for $\theta_1 = \theta_2 \in \{0.25, 0.5, 0.75\}$ computed using 100 realizations of $\mathbf{U}$. The empirical mean approximates the difference between the average ELBO and $\frac{1}{T}\ell(\theta)$. We observe that $\frac{1}{T}|\ell^{\mathrm{ELBO}}(\theta) - \ell_{\epsilon}^{\mathrm{ELBO}}(\theta)|$ is negligible.

We now compare the performance of the estimators $\hat{\theta}_{\mathrm{SMLE}}$ (for DET) and $\hat{\theta}_{\mathrm{ELBO}}$ (for both MUL and DET) learned using gradient with learning rate $10^{-4}$ on 100 steps using $N = 25$ to $\hat{\theta}_{\mathrm{MLE}}$ computed using Kalman derivatives. We simulate $M = 50$ realizations of $T = 150$ observations using $\theta = (\theta_1,\theta_2) = (0.5, 0.5)$. The ELBO stochastic gradient estimates are computed using biased gradient estimates of $\ell_{\mathrm{ELBO}}(\theta)$ ignoring the contributions of resampling steps as in [38, 41, 33] (we recall that unbiased estimates suffer from very high variance) and unbiased gradients of $\ell_{\epsilon}^{\mathrm{ELBO}}(\theta)$ using PF DET. We average $B$ parallel PFs to reduce the variance of these gradients of the ELBO and also $B$ PFs (with fixed random seeds) to compute the gradient of $\hat{\ell}_{\mathrm{SMLE}(\theta;\mathbf{u}_{1:B})} := \frac{1}{B}\sum_{b=1}^{B}\hat{\ell}(\theta;\mathbf{u}_b)$. For this example, $\hat{\theta}_{\mathrm{ELBO}}^{\mathrm{DET}}$ maximizing $\ell_{\epsilon}^{\mathrm{ELBO}}(\theta)$ outperforms $\hat{\theta}_{\mathrm{ELBO}}^{\mathrm{MUL}}$ and $\hat{\theta}_{\mathrm{SMLE}}$. However, as $B$ increases, $\hat{\theta}_{\mathrm{SMLE}}$ gets closer to $\hat{\theta}_{\mathrm{ELBO}}^{\mathrm{DET}}$ which is to be expected as $\hat{\ell}_{\mathrm{SMLE}}(\theta;\mathbf{u}_{1:B}) \longrightarrow \ell_{\epsilon}^{\mathrm{ELBO}}(\theta)$.

Table 1: Mean & std of $\frac{1}{T}(\hat{\ell}(\theta; \mathbf{U}) - \ell(\theta))$

| | Multinomial | | DET | |
|---|---|---|---|---|
| $\theta_1$, $\theta_2$ | mean | std | mean | std |
| 0.25 | -1.02 | 0.18 | -1.02 | 0.18 |
| 0.50 | -0.84 | 0.17 | -0.85 | 0.17 |
| 0.75 | -0.79 | 0.18 | -0.79 | 0.18 |

Table 2: $10^3\times$ RMSE [4] over 50 datasets

| $B$ | $\hat{\theta}^{\text{MUL}}_{\text{ELBO}}$ | $\hat{\theta}^{\text{DET}}_{\text{ELBO}}$ | $\hat{\theta}_{\text{SMLE}}$ |
|---|---|---|---|
| 1 | 4.86 | 3.94 | 16.87 |
| 4 | 4.94 | 3.37 | 7.01 |
| 10 | 4.79 | 2.72 | 4.53 |
| 25 | 4.83 | 2.23 | 2.74 |

We now reproduce a similar example as in [41] where one learns the parameters $\phi$ of the proposal using the ELBO for the following linear Gaussian SSM:

$$X_t|\{X_{t-1} = x\} \sim \mathcal{N}\left(\mathbf{A}x, \mathbf{I}_{d_x}\right), \quad Y_t|\{X_t = x\} \sim \mathcal{N}(\mathbf{I}_{d_y,d_x}x, \mathbf{I}_{d_y}), \tag{20}$$

with $\mathbf{A} = (0.42^{|i-j|+1})_{1\leq i,j\leq d_x}$, $\mathbf{I}_{d_y,d_x}$ is a $d_y \times d_x$ matrix with 1 on the diagonal for the $d_y$ first rows and zeros elsewhere. For $\phi \in \mathbb{R}^{d_x+d_y}$, we consider

$$q_\phi(x_t|x_{t-1}, y_t) = \mathcal{N}(x_t|\Delta_\phi^{-1}\left(\mathbf{A}x_{t-1} + \Gamma_\phi y_t\right), \Delta_\phi), \tag{21}$$

with $\Delta_\phi = \text{diag}(\phi_1, \ldots, \phi_{d_x})$ and $\Gamma_\phi = \text{diag}_{d_x,d_y}(\phi_1, \ldots, \phi_{d_x})$ is a $d_x \times d_y$ matrix with $\phi_i$ on the diagonal for $d_x$ first rows and zeros elsewhere. The locally optimal proposal $p(x_t|x_{t-1}, y_t) \propto g(y_t|x_t)f(x_t|x_{t-1})$ in [19] corresponds to $\phi = \mathbf{1}$, the vector with unit entries of dimension $d_\phi = d_x + d_y$. For $d_x = 25, d_y = 1$, $M = 10$ realizations of $T = 100$ observations using (20), we learn $\phi$ on each realization using 100 steps of stochastic gradient ascent with learning rate $10^{-1}$ on the $\ell^{\text{ELBO}}(\phi)$ using PF MUL with biased gradients and $\ell^{\text{ELBO}}(\phi)$ using PF DET with $N = 25$ and $B = 4$ PFs. While $p(x_t|x_{t-1}, y_t)$ is not guaranteed to maximize the ELBO, our experiments showed that it outperforms optimized proposals. We therefore report the RMSE of $\phi - \mathbf{1}$ and the average Effective Sample Size (ESS) [18] as proxy performance metrics. On both metrics, PF DET outperforms PF MUL. The RMSE is 0.115 for PF DET vs 0.13 for PF MUL while the average ESS after convergence is around 15 for PF DET vs 10 for PF MUL[5].

## 4.2 Variational Recurrent Neural Network (VRNN)

A VRNN is an SSM with latent state $X_t = (R_t, Z_t)$ where $R_t$ is an RNN state and $Z_t$ a latent Gaussian variable, $Y_t$ is here defined as a vector of binary observations [9]. The model is detailed below. $\text{RNN}_\theta$ denotes the forward call of an LSTM (Long Short Term Memory) RNN cell which at time $t$ emits the next RNN state $R_{t+1}$ and output $O_{t+1}$. $E_\theta$, $h_\theta$, $\mu_\theta$, $\sigma_\theta$ are fully connected neural networks; see Appendix 10.2 for more details. We train this model on the polyphonic music benchmark datasets [6], whereby $Y_t$ represents which notes are active. The observation sequences vary between length 127 and 347 for each dataset, with each observation of dimension 88. We chose latent states $Z_t$ and $R_t$ to be of dimension $d_z = 10$ and $d_r = 5$ respectively so $d_x = 15$. We use $q_\phi(x_t|x_{t-1}, y_t) = f_\theta(x_t|x_{t-1})$.

$$\begin{aligned} (R_{t+1}, O_{t+1}) &= \text{RNN}_\theta(R_t, Y_{1:t-1}, E_\theta(Z_t)) & \hat{p}_t &= h_\theta(E_\theta(Z_t), O_t) \\ Z_{t+1} &\sim \mathcal{N}(\mu_\theta(O_{t+1}), \sigma_\theta(O_{t+1})) & Y_t|X_t &\sim \text{Ber}(\hat{p}_t) \end{aligned}$$

[4] The Root Mean Square Error (RMSE) is defined as $\sqrt{\frac{1}{M}\sum_{i=1}^{2}\sum_{k=1}^{M}(\hat{\theta}_i^k - \hat{\theta}_{\text{MLE},i}^k)^2}$.

[5] The average time per iteration was 55 seconds for PF DET and 30 seconds for PF MUL.

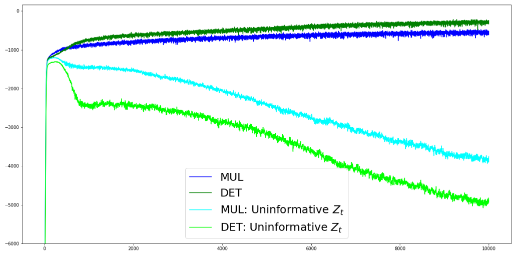


Figure 3: ELBO during training with and without informative $Z_t$, for musedata

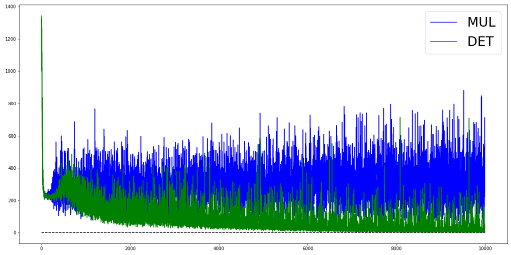


Figure 4: $-\log p_{\hat{\theta}^{(i)}}(y_T|y_{1:T-1})$ across iteration $i$ for musedata

The VRNN model is trained by maximizing $\ell^{\text{ELBO}}(\theta)$ and $\ell^{\text{ELBO}}_{\epsilon}(\theta)$, using PF MUL (with biased gradient estimates) and PF DET respectively, for $N = 25$. Although both methods appear to provide roughly similar ELBO values (see Figure 3), the model learned with DET resampling places a greater influence on the stochastic state $Z_t$ than the model learned with MUL resampling. Indeed, when using the trained models to compute the ELBOs but with $Z_t$ instead replaced with noise, i.e. $Z_t \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{d_z})$, we observe that the negative impact of using uninformative $Z_t$ is significantly larger for the model learned by DET. As discussed in [9], utilising the stochastic latent state is key to performance for such models. It is therefore not surprising that the model learned by DET appears to demonstrate significantly better predictive performance during training for the musedata dataset; see Figure 4. Similar performance improvement can be seen across the other datasets investigated, with metrics available in Appendix 10.2.

# 5 Conclusion

This paper introduces differentiable particle filtering (DPF) based on resampling using regularized OT. DPF retains some of the theoretical guarantees of particle filters and facilitates efficient parameter inference for time-series models using end-to-end gradient descent optimization. Additionally, DPF may be used with neural network parametrized encoders for complex data. Since DPF is end-to-end differentiable, such encoders may be learned jointly with parameters through gradient updates. OT-based differentiable resampling is more expensive than standard resampling but advances in the very active field of computational OT could be used to reduce this complexity; see e.g. [3]. From a theoretical and methodological point of view, it would also be beneficial to obtain quantitative bounds on the bias of the likelihood estimate introduced by differentiable resampling to propose principled guidelines on parameter choice.

# Acknowledgements

This material is based upon work supported in part by the U.S. Army Research Laboratory and the U. S. Army Research Office, and by the U.K. Ministry of Defence (MoD) and the U.K. Engineering and Physical Research Council (EPSRC) under grant number EP/R013616/1. Arnaud Doucet is also supported by the EPSRC grants EP/R018561/1 and EP/R034710/1.

# References

[1] Walter Acevedo, Jana de Wiljes, and Sebastian Reich. Second-order accurate ensemble transform particle filters. *SIAM Journal on Scientific Computing*, 39(5):A1834–A1850, 2017.

[2] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[3] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems*, pages 4429–4439, 2019.

[4] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*, 2015.

[5] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[6] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1881—-1888, 2012.

[7] Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations*, 2016.

[8] Ralph Byers. Solving the algebraic Riccati equation with the matrix sign function. *Linear Algebra and Its Applications*, 85:267 – 279, 1987.

[9] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, pages 2980–2988, 2015.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[11] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. *arXiv preprint arXiv:1310.4375v1*, 2013.

[12] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, Jan 2018.

[13] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, pages 6858–6868, 2019.

[14] David N DeJong, Roman Liesenfeld, Guilherme V Moura, Jean-François Richard, and Hariharan Dharmarajan. Efficient likelihood evaluation of state-space representations. *Review of Economic Studies*, 80(2): 538–567, 2013.

[15] Pierre Del Moral. *Feynman-Kac Formulae.* Springer, 2004.

[16] Randal Douc, Eric Moulines, et al. Limit theorems for weighted samples with applications to sequential monte carlo methods. *The Annals of Statistics*, 36(5):2344–2376, 2008.

[17] Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples.* CRC press, 2014.

[18] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704, 2009.

[19] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[20] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, 2019.

[21] Jim Gatheral. *The Volatility Surface: a Practitioner's Guide*, volume 357. John Wiley & Sons, 2011.

[22] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

[23] Paul Glasserman and Yu-Chi Ho. *Gradient Estimation via Perturbation Analysis*, volume 116. Springer Science & Business Media, 1991.

[24] Christian Gourieroux and Alain Monfort. *Simulation-Based Econometric Methods.* Oxford University Press, 1996.

[25] Matthew M Graham and Alexandre H Thiery. A scalable optimal-transport based local particle filter. *arXiv preprint arXiv:1906.00507*, 2019.

[26] Pieralberto Guarniero, Adam M Johansen, and Anthony Lee. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112(520):1636–1647, 2017.

[27] Rico Jonschkowski, Divyam Rastogi, and Oliver Brock. Differentiable particle filters: End-to-end learning with algorithmic priors. In *Proceedings of Robotics: Science and Systems*, 2018.

[28] Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, and Nicolas Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30(3):328–351, 2015.

[29] Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In *Conference on Robot Learning*, 2018.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[31] Genshiro Kitagawa and Will Gersch. *Smoothness Priors Analysis of Time Series*, volume 116. Springer Science & Business Media, 1996.

[32] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2101–2109, 2017.

[33] Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential Monte Carlo. In *ICLR*, 2018.

[34] Anthony Lee. Towards smooth particle filters for likelihood estimation with multivariate latent variables. Master's thesis, University of British Columbia, 2008.

[35] Lung-Fei Lee. On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models. *Econometric Theory*, 8(4):518–552, 1992.

[36] Xiao Ma, Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter recurrent neural networks. In *AAAI Conference on Artificial Intelligence*, 2020.

[37] Xiao Ma, Peter Karkus, Nan Ye, David Hsu, and Wee Sun Lee. Discriminative particle filter reinforcement learning for complex partial observations. In *ICLR*, 2020.

[38] Chris J Maddison, Dietrich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, 2017.

[39] Sheheryar Malik and Michael K Pitt. Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2):190–209, 2011.

[40] Aaron Myers, Alexandre H Thiery, Kainan Wang, and Tan Bui-Thanh. Sequential ensemble transform for Bayesian inverse problems. *arXiv preprint arXiv:1909.09591*, 2019.

[41] Christian A Naesseth, Scott W Linderman, Rajesh Ranganath, and David M Blei. Variational sequential Monte Carlo. In *AISTATS*, 2018.

[42] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[43] Michael K Pitt. Smooth particle filters for likelihood evaluation and maximisation. Technical report, Warwick Economics Research Paper no. 651, 2002.

[44] Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

[45] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in neural information processing systems*, pages 7785–7794, 2018.

[46] Sebastian Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[47] Sebastian M Schmon, George Deligiannidis, Arnaud Doucet, and Michael K Pitt. Large sample asymptotics of the pseudo-marginal method. *Biometrika, to appear - arXiv preprint arXiv:1806.10060*, 2020.

[48] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *ICLR*, 2018.

[49] Eduardo D Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer Science & Business Media, 2013.

[50] Vladislav ZB Tadić and Arnaud Doucet. Bias of particle approximations to optimal filter derivative. *arXiv preprint arXiv:1806.09590*, 2018.

[51] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. Probabilistic robotics. 2005.

[52] Kenneth E Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[53] Ramon van Handel. Hidden Markov Models. *Lecture Notes. Princeton University*, 2008.

[54] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[55] Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *Proceedings of the 31st Conference On Learning Theory*, 2018.

[56] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

[57] Walter Wonham. On a matrix Riccati equation of stochastic control. *SIAM J. Control Optim.*, 6:681–697, 11 1968.

# 6 Appendix A: Differentiable Transport Methodologies

We present here some details about the implementation of DET transform and alternative differentiable resampling methodologies.

## 6.1 Differentiable Ensemble Transforms (DET)

**Sinkhorn iterates** As described in Section 3.1 the first-order minimization condition for the dual vectors in $\mathrm{DOT}_\epsilon(\mathbf{f}, \mathbf{g})$ is given by (13):

$$f_i^* = \mathcal{T}_\epsilon(\mathbf{b}, \mathbf{g}^*, \mathbf{C}_{i:}), \qquad g_i^* = \mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}^*, \mathbf{C}_{:i}). \tag{22}$$

However, Algorithm 2 uses the "symmetrized" fixed-point iterates given by the equivalent condition

$$f_i^* = \frac{1}{2}\left(f_i^* + \mathcal{T}_\epsilon(\mathbf{b}, \mathbf{g}^*, \mathbf{C}_{i:})\right), \qquad g_i^* = \frac{1}{2}\left(g_i^* + \mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}^*, \mathbf{C}_{:i})\right). \tag{23}$$

which is known to lead to faster convergence when the source and the target distribution are the same. In our case, while they are not exactly the same, they should be similar for cases when the source is not completely degenerate, and intuitively converge faster which was confirmed empirically.

**Gradient computation** Additionally, the gradient can be computed at convergence only: using the implicit function theorem, it suffices to propagate only one step of (13) do to the special form of the fixed-point solution, see [? ] for details.

Naive backpropagation through $\mathbf{P}_\epsilon^{\mathrm{OT}} = \mathbf{P}_\epsilon^{\mathrm{OT}}(\mathbf{X}, \mathbf{w})$ lead to explosive gradients due to the propagation of numerical imprecision. As a consequence we clipped the gradient of $\mathbf{P}_\epsilon^{\mathrm{OT}}$ before running the reverse mode differentiation.

## 6.2 Covariance Adjusted Ensemble Transform (C-DET)

We describe the approach proposed in [1, 25] which modifies the ET to ensure that $\frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}^i}$ has the same covariance as $\sum_{i=1}^N w^i \delta_{X^i}$. This is achieved by using $\tilde{\mathbf{X}} = (N\mathbf{P} + \mathbf{\Delta})\mathbf{X}$ instead of $\tilde{\mathbf{X}} = N\mathbf{P}\mathbf{X}$ where $\mathbf{\Delta}$ is determined as follows.

**Proposition 6.1.** For $\mathbf{P} \in S(\mathbf{w}, \frac{1}{N}\mathbf{1})$, let $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\mathbf{B} = N\mathbf{P}^T$ and $\mathbf{A} = N\mathbf{W} - \mathbf{B}\mathbf{B}^T$. Let $\mathbf{\Sigma}(\mathbf{w}, \mathbf{X})$ denote the covariance of $\sum_{i=1}^N w^i \delta_{X^i}$ and $\mathbf{\Sigma}(\frac{1}{N}\mathbf{1}, \tilde{\mathbf{X}})$ the covariance of $\frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}^i}$ for $\tilde{\mathbf{X}} = (N\mathbf{P} + \mathbf{\Delta})\mathbf{X}$ then

$$\forall \mathbf{X} \in \mathbb{R}^{N \times d_x}, \quad \mathbf{\Sigma}(\mathbf{w}, \mathbf{X}) = \mathbf{\Sigma}(\tfrac{1}{N}\mathbf{1}, \tilde{\mathbf{X}}) \iff \mathbf{\Delta} \text{ is a solution of } \mathbf{A} = \mathbf{B}\mathbf{\Delta} + \mathbf{\Delta}\mathbf{B}^T + \mathbf{\Delta}\mathbf{\Delta}. \tag{24}$$

Moreover, the solution $\mathbf{\Delta}$ to (24) is unique, real symmetric and $\mathbf{\Delta}\mathbf{1} = \mathbf{\Delta}^T\mathbf{1} = \mathbf{0}$.

*Remark.* This parametrization of (24) differs from the one in [1] where the authors solve the (equivalent) Ricatti equation given by $\mathbf{B} = N\mathbf{P} - \mathbf{w}\mathbf{1}^T$ and $\mathbf{A} = N(\mathbf{W} - \mathbf{w}\mathbf{w}^T) - \mathbf{B}\mathbf{B}^T$. This parametrization allows us to prove the well-posedness of (24) in terms of controllability of the matrices pair $[\mathbf{A}, \mathbf{B}]$, see Section 6.2.1.

A number of methods exist to solve the Ricatti Equation (24). In [1], an ODE is discretized to find the stabilizing solution of $\frac{d}{dt}\mathbf{\Delta} = \mathbf{A} - \mathbf{B}\mathbf{\Delta} - \mathbf{\Delta}\mathbf{B}^T - \mathbf{\Delta}\mathbf{\Delta}$. We found this scheme is sometimes numerically unstable and use instead the sign matrix method from [8] detailed in Section 6.2.2.

### 6.2.1 Proof of Proposition 6.1

**Lemma 6.1.** The matrix $\mathbf{A}$ defined in Theorem 6.1 is non singular.

*Proof.* Recall that $\mathbf{A} = N\mathbf{W} - \mathbf{B}\mathbf{B}^T$ where $\mathbf{B} = N\mathbf{P}^T$, $\mathbf{P}\mathbf{1} = \frac{1}{N}\mathbf{1}$ and $\mathbf{P}^T\mathbf{1} = \mathbf{w}$.

For $i \in [N]$, we have $A_{i,i} = Nw_i - N^2 \sum_{k=1}^N p_{k,i}^2$. Moreover, we have

$$\begin{aligned}
\sum_{j\neq i} |A_{ij}| &= N^2 \sum_{j\neq i} \sum_{k=1}^N p_{k,i}p_{k,j} \\
&= N^2 \sum_{k=1}^N \left( \sum_j p_{k,i}p_{k,j} - p_{k,i}^2 \right) \\
&= N^2 \sum_{k=1}^N \left( p_{k,i} \sum_j p_{k,j} - p_{k,i}^2 \right) \\
&= N^2 \sum_{k=1}^N \left( \frac{p_{k,i}}{N} - p_{k,i}^2 \right) \\
&= A_{i,i}
\end{aligned}$$

Therefore the matrix $\mathbf{A}$ is diagonally dominant. Moreover it is symmetric and has non-negative diagonal entries, hence it is non-singular. □

Let us introduce the initial value differential algebraic Riccati equation

$$\frac{d}{dt}\mathbf{\Delta} = \mathbf{A} - \mathbf{B}\mathbf{\Delta} - \mathbf{\Delta}\mathbf{B}^T - \mathbf{\Delta}\mathbf{\Delta} \quad \text{with} \quad \mathbf{\Delta}(0) = \mathbf{0} \in \mathbb{R}^{N\times N} \tag{25}$$

**Lemma 6.2.** There exists a unique solution to (25), and its solution $\mathbf{\Delta}(t) \rightarrow_{t\rightarrow\infty} \Delta$ the unique solution of (24).

*Proof.* The proof follows directly from Theorem 2.1 in [57]. Indeed, (25) is equivalent to the terminal value differential equation $\frac{d}{dt}\mathbf{\Delta} - \mathbf{B}\mathbf{\Delta} - \mathbf{\Delta}\mathbf{B}^T - \mathbf{\Delta}\mathbf{\Delta} + \mathbf{A} = 0$. Moreover the following conditions hold:

1. the pair $(-\mathbf{B}^T, \mathbf{I}_N)$ is stabilizable: clearly $\mathbf{K} = \mathbf{B}^T + \mathbf{I}_N$ for example is such that $-\mathbf{B}^T - \mathbf{K}\mathbf{I}_N = -\mathbf{I}_N$ has negative eigenvalues.
2. the pair $(\mathbf{B}, \sqrt{\mathbf{A}})$ is observable. Indeed, because $\mathbf{A}$ is non singular, $\sqrt{\mathbf{A}}$ isn't either, and the matrix $\mathbf{R}(\lambda) = \left[\lambda\mathbf{I}_N - \mathbf{B}, \sqrt{\mathbf{A}}\right]$ is therefore of rank $N$ for all $\lambda \in \mathbb{C}$, using lemma 3.3.7 in [49], the pair is controllable.

□

Let $\mathcal{S}_0^N(\mathbb{R})$ be the set of real-valued symmetric matrix whose elements $\mathbf{M}$ satisfy $\mathbf{M}\mathbf{1} = \mathbf{M}^T\mathbf{1} = \mathbf{0}$.

**Lemma 6.3.** The solution $\mathbf{\Delta}$ of the Ricatti Equation (24) is real symmetric and $\mathbf{\Delta} \in \mathcal{S}_0^N(\mathbb{R})$.

*Proof.* We use a discretized version of the characterisation given by Lemma 6.2. Let

$$\mathbf{\Delta}_{n+1} = \mathbf{\Delta}_n + \eta \left( \mathbf{A} - \mathbf{B}\mathbf{\Delta}_n - \mathbf{\Delta}_n\mathbf{B}^T - \mathbf{\Delta}_n\mathbf{\Delta}_n \right)$$

By induction one can see that if $\mathbf{\Delta}_n \in \mathcal{S}_0^N(\mathbb{R})$ then $\mathbf{\Delta}_{n+1} \in \mathcal{S}_0^N(\mathbb{R})$. Moreover, $\Delta$ is a stable solution of (25), hence for a small enough time-step $\eta$, $\mathbf{\Delta}_n \rightarrow \mathbf{\Delta}$. $\mathcal{S}_0^N(\mathbb{R})$ being a closed subspace of $\mathbb{R}^{N\times N}$, we have $\mathbf{\Delta} \in \mathcal{S}_0^N(\mathbb{R})$. □

*Proof.* Proof of Proposition 6.1. The only thing left to prove is the equivalence in (24) for $\mathbf{P} \in S(\mathbf{w}, \frac{1}{N}\mathbf{1})$ and $\mathbf{\Delta} \in \mathcal{S}_0^N(\mathbb{R})$. We first note that $\overline{\mathbf{X}}^{\mathbf{w}}$ and $\mathbf{\Sigma}^{\mathbf{w}}(\mathbf{X})$ can be rewritten more compactly respectively as $\mathbf{w}^T\mathbf{X}$ and $\mathbf{X}^T(\mathbf{W} - \mathbf{w}\mathbf{w}^T)\mathbf{X}$. The left hand side of (24) can therefore be rewritten:

$\forall \mathbf{X} \in \mathbb{R}^{N\times d_x}$

$$
\begin{aligned}
& & \mathbf{\Sigma}^{\mathbf{w}}(\mathbf{X}) &= \mathbf{\Sigma}\left((\mathbf{P}+\mathbf{\Delta})\mathbf{X}\right) & \\
&\Longleftrightarrow & \mathbf{X}^T\left(\mathbf{W}-\mathbf{w}\mathbf{w}^T\right)\mathbf{X} &= \mathbf{X}^T(\mathbf{P}+\mathbf{\Delta})^T\left(\frac{1}{N}\mathbf{I}_N-\frac{1}{N^2}\mathbf{1}\mathbf{1}^T\right)(\mathbf{P}+\mathbf{\Delta})\mathbf{X} & (26)\\
&\Longleftrightarrow & \mathbf{W}-\mathbf{w}\mathbf{w}^T &= (\mathbf{P}+\mathbf{\Delta})^T\left(\frac{1}{N}\mathbf{I}_N-\frac{1}{N^2}\mathbf{1}\mathbf{1}^T\right)(\mathbf{P}+\mathbf{\Delta}) & (27)\\
&\Longleftrightarrow & N\mathbf{W} &= (\mathbf{P}+\mathbf{\Delta})^T(\mathbf{P}+\mathbf{\Delta}) & \\
&\Longleftrightarrow & N\mathbf{W}-\mathbf{P}^T\mathbf{P} &= \mathbf{P}^T\mathbf{\Delta}+\mathbf{\Delta}\mathbf{P}+\mathbf{\Delta}\mathbf{\Delta} &
\end{aligned}
$$

The equivalence between lines (26) and (27) comes from the fact that (26) must be true for all $\mathbf{X} \in \mathbb{R}^{N\times d_x}$.
□

### 6.2.2 Numerical solution of the Riccati equation (24)

We detail here the algorithm used to compute the solution of (24). The technique employed comes from [8].

**Definition 6.1.** Let $\mathbf{A} \in \mathbb{R}^{M\times M}$, the sign of matrix $\mathbf{A}$ is defined as $\text{sign}(\mathbf{A}) = \mathbf{A}(\mathbf{A}^2)^{-\frac{1}{2}}$.

The following theorem is a direct application of [8].

**Theorem 6.1.** Let $\mathbf{H} \in \mathbb{R}^{2N\times 2N}$ be defined by block as $\mathbf{H} = \begin{bmatrix} -\mathbf{B}^T & -\mathbf{I}_N \\ -\mathbf{A} & \mathbf{B} \end{bmatrix}$. Let $\mathbf{S} = \text{sign}(\mathbf{H})$ and $\mathbf{QR}$ be the QR decomposition of $\frac{1}{2}\left(\mathbf{I}_{2N} - \mathbf{S}\right)$. If $\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}$, then $\mathbf{\Delta} = \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}$ solves the Riccati Equation (24).

We use Newton iterates to compute $\text{sign}(\mathbf{H})$.

**Theorem 6.2.** For $\mathbf{M} \in \mathbb{R}^{M\times M}$ define the sequence $\mathbf{M}_{n+1} = \frac{1}{2}\left(\mathbf{M}_n + \mathbf{M}_n^{-1}\right), \quad \mathbf{M}_0 = \mathbf{M}$, then $\mathbf{M}_n$ converges globally to $\text{sign}(\mathbf{M})$.

*Remark.* The iterative method highlighted by Theorem 6.2 has the benefit of coming together with an estimation of the current error. Indeed, let $\mathbf{Y}_n = \frac{1}{2}(\mathbf{X}_n^2 + \mathbf{X}_n)$ and $\epsilon_n = ||\mathbf{Y}_n^2 - \mathbf{Y}_n||_2$, then $\epsilon_n \approx ||\text{sign}(\mathbf{X}) - \mathbf{X}_n||_2$. This allows one to stop the algorithm at convergence.

*Remark.* All of the operations above have well defined and computable gradients with respect to their inputs, as a result, one may perform backpropagation through the operations.

Putting together these results yields Algorithm 4.

## 6.3 Variance Adjusted Ensemble Transform (V-DET)

While the covariance adjusted DET is attractive, it presents the disadvantage of having complexity $O(N^3)$ and an unstable gradient due to the complex successive operations necessary for computing $\mathbf{\Delta}$. To alleviate this we propose here a simplified operation which, whilst not correcting for the full covariance matrix, will correct for the individual dimensions variance and modify the cross-terms as a by-product.

Let $\tilde{\mathbf{X}} = N\mathbf{P}_\epsilon\mathbf{X}$ be the transformed sample, and let $\overline{\mathbf{Z}}^{\mathbf{w}}$ and $\sigma^{\mathbf{w}}(\mathbf{Z}) \in \mathbb{R}^{d_x}$ be respectively the component-wise weighted mean and standard deviation of $\mathbf{Z} \in \mathbb{R}^{N\times d_x}$ (where we drop $\mathbf{w}$ for the equally weighted case). Here only we define $\odot$, $\oslash$ to be respectively the element-wise multiplication and division. Let $\mathbf{a} \in \mathbb{R}_+^{d_x}, \mathbf{b} \in \mathbb{R}^{d_x}$ be such that

$$\sigma^{\mathbf{w}}(\mathbf{X}) = \sigma(a \odot \ \tilde{\mathbf{X}} + \mathbf{b}) \quad \text{and} \quad \overline{\mathbf{X}}^{\mathbf{w}} = \overline{a \odot \ \tilde{\mathbf{X}} + \mathbf{b}} \tag{28}$$

then $\mathbf{a}$ and $\mathbf{b}$ are uniquely defined by

$$\mathbf{a} = \sigma^{\mathbf{w}}(\mathbf{X}) \oslash \sigma(\tilde{\mathbf{X}}) \quad \text{and} \quad \mathbf{b} = (\mathbf{1} - \mathbf{a}) \odot \overline{\mathbf{X}}^{\mathbf{w}} \tag{29}$$

This presents the benefit of correcting for the collapsing of the point cloud while being more stable and requiring only $O(N)$ operations.

**Algorithm 4** Covariance Adjusted Differentiable Ensemble Transform (C-DET)

```
function CORRECTION($\mathbf{A}, \mathbf{B}, N, \epsilon$)
    $\mathbf{H} \leftarrow \begin{bmatrix} -\mathbf{B}^T & -\mathbf{I}_N \\ -\mathbf{A} & \mathbf{B} \end{bmatrix}$
    $\mathbf{S} \leftarrow \mathbf{H}$
    $\tilde{\epsilon} \leftarrow \epsilon$
    while $\tilde{\epsilon} \geq \epsilon$ do
        $\mathbf{S} \leftarrow \frac{1}{2}(\mathbf{S} + \mathbf{S}^{-1})$
        $\mathbf{Y} \leftarrow \frac{1}{2}(\mathbf{S}^2 + \mathbf{S})$
        $\tilde{\epsilon} \leftarrow ||\mathbf{Y}^2 - \mathbf{Y}||_2$
    $(\mathbf{Q}, \mathbf{R}) \leftarrow QR(\frac{1}{2}(\mathbf{I}_{2N} - \mathbf{S}))$
    $\mathbf{U} \leftarrow \mathbf{Q}[1 : N, 1 : N]$
    $\mathbf{V} \leftarrow \mathbf{Q}[N + 1 : 2N, 1 : N]$
    Solve the system $\mathbf{U\Delta} = \mathbf{V}^T$
    return $\mathbf{\Delta}$
```

### 6.4 Particle Cloud Optimization (O-DET)

Instead of transporting the degenerate particles using the coupling $\mathbf{P}_\epsilon^{\mathrm{OT}}$, we could directly minimize $\mathrm{OT}_\epsilon(\alpha, \beta)$ with respect to new particle locations $\tilde{\mathbf{X}}_t$ at time $t$ between $\alpha = \sum w_t^i \delta_{X_t^i}$ and $\beta = \frac{1}{N} \sum \delta_{\tilde{X}_t^i}$. This idea was first proposed in [11] in the context of minimizing $\mathrm{OT}_\epsilon$. However contrarily to $\mathrm{OT}(\alpha, \beta)$, this does not define a positive definite loss function. As a consequence we propose here instead to minimize the Sinkhorn divergence $\mathrm{SD}_\epsilon(\alpha, \beta)$ defined in (14). This divergence was first introduced by [22] and was later proven to be a positive-definite symmetric loss function [20]. A gradient of this divergence with respect to $\tilde{\mathbf{X}}_t$ can be computed using Algorithm 2 as [20, Section 3.2] shows that the derivatives of $\mathrm{OT}_\epsilon(\alpha, \beta)$ satisfy

$$\frac{\partial \mathrm{OT}_\epsilon}{\partial a_i} = f_i^*, \quad \frac{\partial \mathrm{OT}_\epsilon}{\partial b_j} = g_j^*, \quad \frac{\partial \mathrm{OT}_\epsilon}{\partial c_{i,j}} = \frac{\partial \Phi}{\partial c_{i,j}}, \tag{30}$$

where

$$\Phi = \sum_{i=1}^N a_i \mathcal{T}_\epsilon(\mathbf{b}, \mathbf{g}, \mathbf{C}_{i:}) + \sum_{j=1}^N b_j \mathcal{T}_\epsilon(\mathbf{a}, \mathbf{f}, \mathbf{C}_{:j}).$$

One may perform gradient descent w.r.t. the particles locations. While the gradient of the particle cloud at convergence (with respect to the original locations) is not accessible analytically, it can be obtained by performing automatic differentiation through each step of the unrolled gradient descent. Because the operations involved during the gradient descent are differentiable, this procedure forms a differentiable resampling algorithm. A few different choices may be considered as the starting point for the gradient descent, the most obvious one being the original particle locations $\mathbf{X}_t$. However, the original point cloud $\mathbf{X}_t$, corresponding to the degenerate sample we want to modify, may be far from the optimal one, instead we start from the differentiable ET proposed in Algorithm 3. As already discussed, this is a differentiable step, therefore the whole procedure is end-to-end differentiable. The algorithm is summarized in Algorithm 5.

**Algorithm 5** Particle Cloud Optimisation (O-DET)

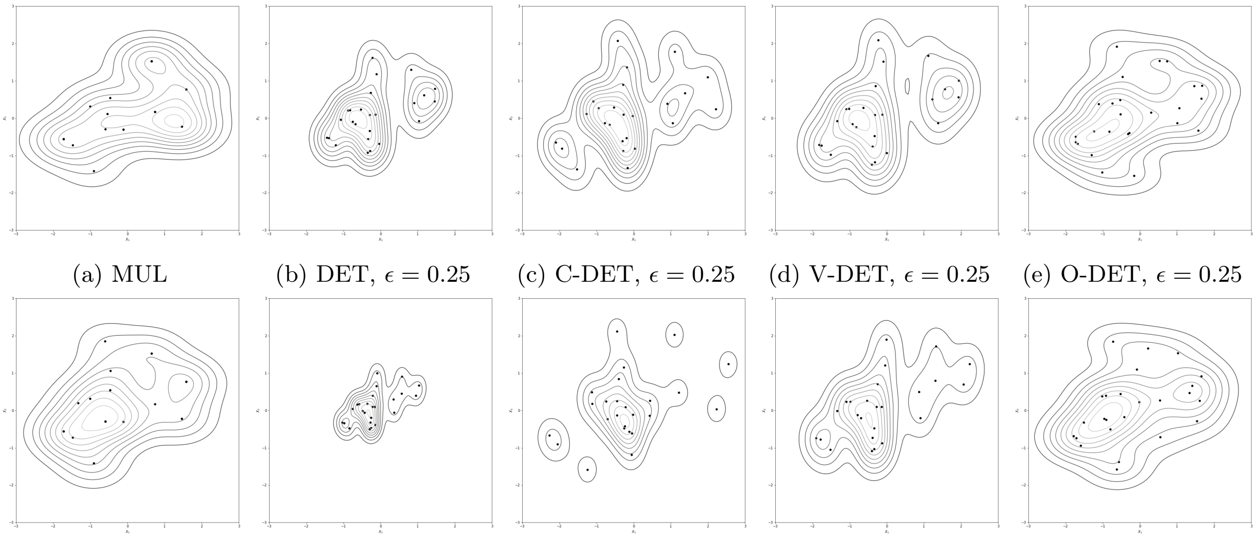
```
function POINTCLOUD($\mathbf{X}, \mathbf{w}, N$)
    $\tilde{\mathbf{X}} \leftarrow$ ENSTRANSFORM($\mathbf{X}, \mathbf{w}, N$)
    for $i \in [n_{\text{steps}}]$ do
        $\tilde{\mathbf{X}} \leftarrow \tilde{\mathbf{X}} - \lambda \nabla_{\tilde{\mathbf{X}}} \mathrm{SD}_\epsilon((\mathbf{X}, \mathbf{w}), (\tilde{\mathbf{X}}, \frac{1}{N}))$
    return $\tilde{\mathbf{X}}$
```

Such an algorithm executes $n_{\text{steps}}$ steps of gradient descent, each with $O(N^2)$ complexity. The resulting run time will therefore be $n_{\text{steps}+1}$ times higher than the direct differentiable ET.

### 6.5 Illustrations

We now illustrate how these different methods compare on a 2D example: we generate $N = 25$ weighted particles $\mathbf{w} \sim \mathcal{U}(0, 1)$, $\mathbf{X} \sim \mathcal{N}(0, I_2)$ and normalize $\mathbf{w}$ prior to resampling the degenerate particles. The resulting applications of our different resampling techniques for different regularization parameters are illustrated in Figure 5. It can be observed that increasing $\epsilon$ collapses the resulting DET particles while leaving the other schemes somewhat unchanged. In particular the Point Cloud Optimized particles exhibit robustness to the choice of $\epsilon$ albeit at the cost of increased computational complexity.

While the C-DET and V-DET seem attractive in principle, they are only useful when the covariance of the weighted sample is meaningful. This is only the case when enough particles are non-degenerate, otherwise, the resulting correction carries only noise forward. This is why we limited our experiments to the use of DET.



(a) MUL (b) DET, $\epsilon = 0.25$ (c) C-DET, $\epsilon = 0.25$ (d) V-DET, $\epsilon = 0.25$ (e) O-DET, $\epsilon = 0.25$

(f) Systematic (g) DET, $\epsilon = 0.75$ (h) C-DET, $\epsilon = 0.75$ (i) V-DET, $\epsilon = 0.75$ (j) O-DET, $\epsilon = 0.75$

Figure 5: Resulting equally weighted point clouds for the proposed resampling schemes with different $\epsilon$ regularization parameters

# 7 Appendix B: Proof of Proposition 3.1

## 7.1 Notation and definition

Recall that we consider $\mathcal{X} = \mathbb{R}^{d_x}$, we denote by $\mathcal{B}(\mathcal{X})$ the Borel sets of $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

In an abuse of notation, we shall use the same notation for a probability measure and its density w.r.t. Lebesgue measure; i.e. $\nu(\mathrm{d}x) = \nu(x)\mathrm{d}x$. We also use the standard notation $\nu(\psi) = \int \psi(x)\nu(x)\mathrm{d}x$ for any test function $\psi$.

We denote

$$||\psi||_\infty = \sup_{x \in \mathcal{X}} |\psi(x)|, \qquad ||\psi||_{\text{Lip}} = \sup_{x,y \in \mathcal{X}: x \neq y} \frac{|\psi(x) - \psi(y)|}{||x - y||}. \tag{31}$$

We then introduce the following set of bounded 1-Lipschitz functions

$$\mathrm{BL} = \Big\{\psi : ||\psi||_\infty < 1, ||\psi||_{\mathrm{Lip}} < 1\}. \tag{32}$$

We will need to work with random probability measures. We recall here briefly some definitions and results. Consider a probability space $(\Lambda, \mathcal{F}, \mathbb{P})$. The product space $\Lambda \times \mathcal{X}$ is equipped with the product $\sigma$-algebra, $\mathcal{F} \otimes \mathcal{B}(\mathcal{X})$.

**Definition 7.1** (*Random probability measure*)**.** A random probability measure is a map $\mu\colon \Lambda \times \mathcal{B}(\mathcal{X}) \to [0,1]$ such that $\mu(\lambda, \cdot) \in \mathcal{P}(\mathcal{X})$ for almost every $\lambda \in \Lambda$ and for every $B \in \mathcal{B}(\mathcal{X})$ the map $\lambda \mapsto \mu(\lambda, B) = \mu^\lambda(B)$ is measurable.

**Definition 7.2** (*Weak convergence of random measures*)**.** A sequence of random probability measures $(\mu_n^\lambda)_{n\geq 1}$ converges weakly almost surely to a probability measure $\mu$, denoted $\mu_n^\lambda \underset{a.s.}{\rightsquigarrow} \mu$, if

$$\mathbb{P}\left(\lambda \in \Lambda : \ \mu_n^\lambda \rightsquigarrow \mu\right) = 1, \tag{33}$$

where $\rightsquigarrow$ denotes standard weak convergence.

As $\mathcal{X}$ is a Polish space, it is a standard result that almost sure weak convergence of $(\mu_n^\lambda)_{n\geq 1}$ is equivalent to

$$\mu_n^\lambda(\psi) \xrightarrow{\text{a.s.}} \mu(\psi) \ \text{ for all } \ \psi \in \mathrm{BL}; \tag{34}$$

see e.g. [47] for a detailed discussion.

For two random measures $\rho_1^\lambda, \rho_2^\lambda$ on $\mathcal{X}$, we will also introduce

$$|||\rho_1^\lambda - \rho_2^\lambda||| = \sup_{\psi\in\mathrm{BL}} |\mathbb{E}[\rho_1^\lambda(\psi) - \rho_2^\lambda(\psi)]|. \tag{35}$$

To simplify presentation, we will omit $\lambda$ notationally.

## 7.2 Assumptions

Consider the random probability measures $\beta_N = \frac{1}{N}\sum_{i=1}^N \delta_{X^i}$ and $\alpha_N(\psi) = \beta_N(\omega\psi)/\beta_N(\omega)$ for some positive bounded function $\omega : \mathcal{X} \mapsto \mathbb{R}^{+*}$; i.e. $\alpha_N = \sum_{i=1}^N w_i \delta_{X^i}$ for $w_i \propto \omega(X^i)$. We will consider such sequences of random measures, each sequence being implicitly indexed by $\lambda \in \Lambda$ and defined over an underlying probability space; see Subsection 7.1. In our DPF context, we can think of $\lambda$ as the random seed $\mathbf{u}$ used. We will also denote $\mathcal{P}^{\mathrm{OT},N}$ an optimal transport between $\alpha_N$ and $\beta_N$. We will make the following assumptions, Assumptions 7.1 and 7.2 are in the spirit of those made in [40, Theorem 3.5]. Below, when we write a.s. we mean $\lambda$-almost surely.

**Assumption 7.1.** *Random Measures and Optimal Transport Regularity.*
**A**. There exist probability measures $\breve{\alpha}$ and $\breve{\beta}$ with $\breve{\alpha}(\psi) = \breve{\beta}(\omega\psi)/\breve{\beta}(\omega)$ such that $\alpha_N \underset{a.s.}{\rightsquigarrow} \breve{\alpha}$ and $\beta_N \underset{a.s.}{\rightsquigarrow} \breve{\beta}$.

**B**. The OT problem

$$\mathrm{OT}(\breve{\alpha}, \breve{\beta}) = \min_{\mathcal{P}\in\mathcal{U}(\breve{\alpha},\breve{\beta})} \mathbb{E}_{(U,V)\sim\mathcal{P}}\left[||U - V||^2\right]$$

admits a unique solution $\mathcal{P}^{\mathrm{OT}}$ which is given by a deterministic map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$.

**C**. For any $\psi \in \mathrm{BL}$, we have $\beta_N(\psi \odot \mathbf{T}) \underset{a.s.}{\to} \breve{\beta}(\psi \odot \mathbf{T}) := \breve{\beta}(\psi(\mathbf{T}(x))) = \breve{\alpha}(\psi)$.

**Assumption 7.2.** *Growth.* We have

$$\lim \sup_{N\to\infty} \mathbb{E}[\beta_N(x \to ||x||^2)] < \infty,$$

and there exists some $p > 1$ such that

$$\lim \sup_{N\to\infty} \mathbb{E}[\beta_N(x \to ||\mathbf{T}(x)||^p) + \alpha_N(x \to ||x||^p)] < \infty.$$

**Assumption 7.3.** There exists an increasing sequence $(N_k)_{k\geq 1}$ such that $\mathcal{P}^{\mathrm{OT},N_k} \underset{a.s.}{\rightsquigarrow} \mathcal{P}^{\mathrm{OT}}$.

Assumption 7.1-A will be typically easy to check while Assumption 7.1-B is verified under mild assumptions.

## 7.3 Proofs

The precise statement of Proposition 3.1 is given in Proposition 7.1.

**Proposition 7.1.** Let Assumptions 7.1, 7.2 and 7.3 hold. Let $\bar{\mathbf{X}} = N\mathbf{P}^{\mathrm{OT},N}\mathbf{X}$ and $\tilde{\mathbf{X}} = N\mathbf{P}_\epsilon^{\mathrm{OT}}\mathbf{X}$. Here $\mathbf{P}^{\mathrm{OT},N}$, resp. $\mathbf{P}_\epsilon^{\mathrm{OT}}$, is the OT, resp. regularized OT, between $\alpha_N$ and $\beta_N$. Denote $\bar{\alpha}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{X}^i}$ and $\tilde{\alpha}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}^i}$. Then, for any $\psi \in \mathrm{BL}$, there exists an increasing sequence $(N_k)_{k\geq 1}$ such that

$$\Big|\mathbb{E}[\bar{\alpha}_{N_k}(\psi) - \alpha_{N_k}(\psi)]\Big| \underset{k\to\infty}{\longrightarrow} 0, \tag{36}$$

while

$$\Big|\mathbb{E}\Big[\tilde{\alpha}_{N_k}(\psi) - \alpha_{N_k}(\psi)\Big]\Big| \leq \kappa_{N_k} + \mathcal{E}_{N_k,\epsilon}.$$

Here $\kappa_{N_k}$ denotes the error from the ET, which is such that $\kappa_{N_k} \underset{k\to\infty}{\longrightarrow} 0$, and $\mathcal{E}_{N_k,\epsilon}$ denotes the additional error introduced by the DET, which is such that $\mathcal{E}_{N_k,\epsilon} \underset{\epsilon\to 0}{\longrightarrow} 0$.

To prove Proposition 7.1, we first prove (36) in the next proposition and corollary using a slightly modified version of [40, Theorem 3.5] and elements of [48, Theorem 2], with particular care to recognize that we are working here with random measures, and only have convergence for an extracted subsequence.

**Proposition 7.2.** Let Assumptions 7.1 and 7.2 hold. Let $\bar{\mathbf{X}} = N\mathbf{P}^{\mathrm{OT},N}\mathbf{X}$ where $\mathbf{P}^{\mathrm{OT},N}$ is the optimal transport between $\alpha_N$ and $\beta_N$[6] and denote $\bar{\alpha}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{X}^i}$. Then, for any $\psi \in \mathrm{BL}$ and for almost all $\lambda$, there exists an increasing $\lambda$-dependent sequence $(N_k)_{k\geq 1}$ such that

$$\Big|\bar{\alpha}_{N_k}(\psi) - \alpha_{N_k}(\psi)\Big| \underset{k\to\infty}{\longrightarrow} 0. \tag{37}$$

*Proof.* Let $\mathbf{T}$ be defined per Assumption 7.1-B as the unique transport map between $\breve{\alpha}$ and $\breve{\beta}$.

$$\begin{aligned}\Big|\frac{1}{N}\sum_{i=1}^N \psi(\bar{X}^i) - \sum_{i=1}^N w^i\psi(X^i)\Big| \\ \leq \Big|\frac{1}{N}\sum_{i=1}^N \psi(\mathbf{T}(X^i)) - \sum_{i=1}^N w^i\psi(X^i)\Big| + \Big|\frac{1}{N}\sum_{i=1}^N \psi(\bar{X}^i) - \frac{1}{N}\sum_{i=1}^N \psi(\mathbf{T}(X^i))\Big|\end{aligned} \tag{38}$$

By Assumption 7.1-A, we have $\alpha_N(\psi) \underset{a.s.}{\to} \breve{\alpha}(\psi)$. By Assumption 7.1-C, $\beta_N(\psi \odot \mathbf{T}) \to \breve{\beta}(\psi \odot \mathbf{T}) = \breve{\alpha}(\psi)$ almost surely. Hence, for the first term of the r.h.s. of (38), we have $\left|\frac{1}{N}\sum_{i=1}^N \psi(\mathbf{T}(X^i)) - \sum_{i=1}^N w^i\psi(X^i)\right| \underset{a.s.}{\to} 0$.

[6] If $\mathbf{P}^{\mathrm{OT},N}$ is not unique, select the OT with maximal entropy.

For the second term on the r.h.s. of (38), we have

$$\begin{aligned}\Big|\frac{1}{N}\sum_{i=1}^{N}\psi(\bar{X}^i)-\frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{T}(X^i))\Big| &\leq \frac{1}{N}\sum_{i=1}^{N}\Big|\Big|\bar{X}^i-\mathbf{T}(X^i)\Big|\Big|\\ &=\frac{1}{N}\sum_{i=1}^{N}\Big|\Big|N\sum_{j=1}^{N}p_{i,j}^{\mathrm{OT},N}X^j-\mathbf{T}(X^i)\Big|\Big|\\ &=\frac{1}{N}\sum_{i=1}^{N}\Big|\Big|N\sum_{j=1}^{N}p_{i,j}^{\mathrm{OT},N}X^j-N\sum_{j=1}^{N}p_{i,j}^{\mathrm{OT},N}\mathbf{T}(X^i)\Big|\Big|\\ &\leq\sum_{i=1}^{N}\sum_{j=1}^{N}p_{i,j}^{\mathrm{OT},N}\Big|\Big|X^j-\mathbf{T}(X^i)\Big|\Big|\\ &=\mathcal{P}^{\mathrm{OT},N}(F)\end{aligned}$$

where $F(x,y)=\Big|\Big|x-\mathbf{T}(y)\Big|\Big|$.

By [54, Theorem 5.20], we can deduce that for almost all $\lambda$ there exists an increasing $\lambda$-dependent sequence $(N_k)_{k\geq 1}$ such that $\mathcal{P}^{\mathrm{OT},N_k}\rightsquigarrow\mathcal{P}^{\mathrm{OT}}$. Using Assumption 7.2 and Minkowski inequality, $\limsup_N\mathcal{P}^{\mathrm{OT},N_k}(F^p)<\infty$, hence $\mathcal{P}^{\mathrm{OT},N_k}(F)\to\mathcal{P}^{\mathrm{OT}}(F)=0$. Thus this means that for almost all $\lambda$, there exists an increasing $\lambda$-dependent sequence $(N_k)_{k\geq 1}$ such that $\left|\frac{1}{N_k}\sum_{i=1}^{N_k}\psi(\bar{X}^i)-\sum_{i=1}^{N_k}w^i\psi(X^i)\right|\to 0$. □

**Corollary 7.1.** Let Assumptions 7.1, 7.2 and 7.3 hold. Let $\bar{\mathbf{X}}=N\mathbf{P}^{\mathrm{OT},N}\mathbf{X}$ where $\mathbf{P}^{\mathrm{OT},N}$ is the optimal transport between $\alpha_N$ and $\beta_N$. Then, for any $\psi\in\mathrm{BL}$, there exists an increasing sequence $(N_k)_{k\geq 1}$

$$\Big|\mathbb{E}[\bar{\alpha}_{N_k}(\psi)-\alpha_{N_k}(\psi)]\Big|\xrightarrow[k\to\infty]{}0. \tag{39}$$

*Proof.* Similar to Proposition 7.2, we have

$$\begin{aligned}\Big|\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N}\psi(\bar{X}^i)-\sum_{i=1}^{N}w^i\psi(X^i)\Big]\Big| &\leq\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=1}^{N}\psi(\bar{X}^i)-\frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{T}(X^i))\Big|\Big]\\ &+\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{T}(X^i))-\sum_{i=1}^{N}w^i\psi(X^i)\Big|\Big]\end{aligned}$$

Recall from Proposition 7.2 that the second term inside the expectation converges to 0 almost surely. Hence as $\psi$ is bounded, the expectation will also converge to 0 by bounded convergence. Again, using the same steps from Proposition 7.2 on the first term

$$\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=1}^{N}\psi(\bar{X}^i)-\frac{1}{N}\sum_{i=1}^{N}\psi(\mathbf{T}(X^i))\Big|\Big]\leq\mathbb{E}[\mathcal{P}^{\mathrm{OT},N}(F)].$$

Following again the proof of Proposition 7.2, under Assumption 7.3, there exists an increasing sequence $(N_k)_{k\geq 1}$ independent of $\lambda$ such that $\mathcal{P}^{\mathrm{OT},N_k}\underset{a.s.}{\rightsquigarrow}\mathcal{P}^{\mathrm{OT}}$. Under Assumption 7.2, $\limsup\mathbb{E}[\mathcal{P}^{\mathrm{OT},N_k}(F^p)]<\infty$ so $\mathbb{E}[\mathcal{P}^{\mathrm{OT},N_k}(F)]\underset{k\to\infty}{\to}\mathbb{E}[\mathcal{P}^{\mathrm{OT}}(F)]=0$ thus $\left|\mathbb{E}[\frac{1}{N_k}\sum_{i=1}^{N_k}\psi(\bar{X}^i)]-\mathbb{E}[\sum_{i=1}^{N_k}w^i\psi(X^i)]\right|\to 0$. □

We are now in a position to prove Proposition 7.1.

*Proof.* (Proof of Proposition 7.1). The first part of the result (36) has been established in Corollary 7.1. One may decompose the DET error using the non-regularized ET as follows:

$$\Big|\mathbb{E}\Big[\tilde{\alpha}_N(\psi)-\alpha_N(\psi)\Big]\Big|\leq\Big|\mathbb{E}\Big[\tilde{\alpha}_N(\psi)-\bar{\alpha}_N(\psi)\Big]\Big|+\Big|\mathbb{E}\Big[\bar{\alpha}_N(\psi)-\alpha_N(\psi)\Big]\Big|.$$

By Lipschitz property of $\psi$, definition of the ET and DET and Cauchy Schwarz's inequality, we have a.s.:

$$\begin{aligned}\Big|\tilde{\alpha}_N(\psi) - \bar{\alpha}_N(\psi)\Big| &\leq \frac{1}{N}\sum_{i=1}^{N}\Big|\psi(\tilde{X}^i) - \psi(\bar{X}^i)\Big| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}||\tilde{X}^i - \bar{X}^i|| \\ &= \sum_{i=1}^{N}\Big|\Big|\sum_{j=1}^{N}(p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j})X^j\Big|\Big| \\ &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\Big|\Big|(p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j})X^j\Big|\Big| \\ &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\Big[|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|^{\frac{1}{2}}\Big]\Big[|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|^{\frac{1}{2}}||X^j||\Big] \\ &\leq \Big[\sum_{i=1}^{N}\sum_{j=1}^{N}|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|\Big]^{\frac{1}{2}}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|||X^j||^2\Big]^{\frac{1}{2}}. \end{aligned} \tag{40}$$

A similar argument appears in the proof of [48, Theorem 2]. Now by Cauchy Schwartz's inequality, we have

$$\mathbb{E}\Big[\Big|\tilde{\alpha}_N(\psi) - \bar{\alpha}_N(\psi)\Big|\Big] \leq \mathbb{E}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|\Big]^{1/2}\mathbb{E}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|||X^j||^2\Big]^{1/2}. \tag{41}$$

The first product term on the r.h.s. tends to 0 as $\epsilon \to 0$ using [12, Proposition 4.1]. The second product term may be bounded by the second moment:

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}|p^{\mathrm{OT},N}_{\epsilon,i,j} - p^{\mathrm{OT},N}_{i,j}|||X^j||^2 &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}(p^{\mathrm{OT},N}_{\epsilon,i,j} + p^{\mathrm{OT},N}_{\epsilon,i,j})||X^j||^2 \\ &= \sum_{j=1}^{N}\Big[\sum_{i=1}^{N}p^{\mathrm{OT},N}_{\epsilon,i,j} + \sum_{i=1}^{N}p^{\mathrm{OT},N}_{i,j}\Big]||X^j||^2 \\ &= \frac{2}{N}\sum_{j=1}^{N}||X^j||^2\end{aligned}$$

It follows that

$$\begin{aligned}\Big|\mathbb{E}\Big[\tilde{\alpha}_N(\psi) - \bar{\alpha}_N(\psi)\Big]\Big| &\leq \mathbb{E}\Big[\Big|\tilde{\alpha}_N(\psi) - \bar{\alpha}_N(\psi)\Big|\Big] \\ &\leq 2\mathbb{E}\Big[\beta_N(x \to ||x||^2)\Big].\end{aligned}$$

So we can conclude by Assumption 7.2. □

*Remark.* Exponential convergence of $\mathbf{P}^{\mathrm{OT},N}_{\epsilon}$ to $\mathbf{P}^{\mathrm{OT},N}$ as $\epsilon \to 0$ has been established in [55]. However, these results do not appear useful in our setting as it is unclear how one could quantify the "sub-optimality gap".

*Remark.* An alternative promising approach to bound the bias introduced by the DET is in terms of the row wise variance of the transport matrix. Let $X'|(X = X_i) \sim \mathcal{P}^{\mathrm{OT,N}}_{\epsilon}(\cdot|X = X_i) = N\sum_j p^{\mathrm{OT},N}_{\epsilon,i,j}\delta_{X_j}(\cdot)$ be samples from the transport map and $\alpha'_N = \frac{1}{N}\sum_i \delta_{X'_i}$ where $X'_i \sim \mathcal{P}^{\mathrm{OT,N}}_{\epsilon}(\cdot|X = X_i)$. It is clear that resampling from the transport map is unbiased, i.e. $\mathbb{E}[\alpha'_N(\psi)] = \mathbb{E}[\alpha_N(\psi)]$, so for any $\psi \in \mathrm{BL}$

$$
\begin{aligned}
|\mathbb{E}[\alpha_N(\psi)] - \mathbb{E}[\tilde{\alpha}_N(\psi)]| &= |\mathbb{E}\left[\alpha'_N(\psi)\right] - \mathbb{E}[\tilde{\alpha}_N(\psi)]\,| \\
&= \left|\mathbb{E}\left[\iint ||x' - \mathbb{E}[X'|x]||\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x'|x)\beta_N(\mathrm{d}x)\right]\,\right| \\
&\leq \mathbb{E}\left[\left(\iint ||x' - \mathbb{E}[X'|x]||^2\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x'|x)\beta_N(\mathrm{d}x)\right)^{\frac{1}{2}}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_i \int ||x' - \mathbb{E}[X'|X_i]||^2\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x'|X_i)\right)^{\frac{1}{2}}\right]
\end{aligned}
$$

This expression shows that the DET error is directly related to how diffuse the conditional distributions $\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x'|X_i)$ are. The benefit of this approach is that it bypasses having to introduce $\mathcal{P}^{\mathrm{OT}}$. However, although seemingly more elegant, it is not clear how one could control this error in terms of $\epsilon$ and $N$.

# 8 Appendix C: Proof of Proposition 3.2

A particle filter with multinomial resampling is defined by the following joint distribution

$$
\overline{q}_{\theta,\phi}(x_{1:T}^{1:N}, a_{1:T-1}^{1:N}) = \prod_{i=1}^{N} q_\phi\left(x_1^i\right) \prod_{t=2}^{T}\prod_{i=1}^{N} w_{t-1}^{a_{t-1}^i} q_\phi\left(x_t^i | x_{t-1}^{a_{t-1}^i}\right)
$$

where $a_{t-1}^i \in \{1, ..., N\}$ is the ancestral index of particle $x_t^i$ and

$$
\omega_{\theta,\phi}(x_1) = \frac{p_\theta(x_1, y_1)}{q_\phi(x_1)}, \quad \omega_{\theta,\phi}(x_{t-1}, x_t) = \frac{p_\theta(x_t, y_t|x_{t-1})}{q_\phi\left(x_t|x_{t-1}\right)}.
$$

Finally, we have $w_t^i \propto \omega_{\theta,\phi}(x_{t-1}^{a_{t-1}^i}, x_t^i)$, $\sum_{i=1}^N w_t^i = 1$. We do not emphasize notationally that the weights $w_{t-1}^{a_{t-1}^i}$ are $\theta, \phi$ and observations dependent.

*Proof.* The ELBO is given by

$$
\ell^{\mathrm{ELBO}}(\theta,\phi) =\mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\right] = \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_1^i)\right) + \sum_{t=2}^T \log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_{t-1}^{A_{t-1}^i}, X_t^i)\right)\right].
$$

We now compute $\nabla_\theta \ell^{\mathrm{ELBO}}(\theta,\phi)$. We can split the gradient using the product rule and apply the log-derivative trick:

$$
\begin{aligned}
\nabla_\theta \ell^{\mathrm{ELBO}}(\theta,\phi) =&\mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\nabla_\theta \log \widehat{p}_\theta(y_{1:T})\right] + \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\nabla_\theta \log \overline{q}_{\theta,\phi}(X_{1:T}^{1:N}, A_{1:T-1}^{1:N})\right] && (42)\\
=&\mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_1^i)\right) + \sum_{t=2}^T \nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_{t-1}^{A_{t-1}^i}, X_t^i)\right)\right] && (43)\\
&+ \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\left\{\sum_{t=2}^T\sum_{i=1}^N \nabla_\theta \log w_{t-1}^{A_{t-1}^i}\right\}\right] && (44)
\end{aligned}
$$

We have for the first part of the ELBO gradient (43) that

$$
\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_1^i)\right) = \sum_{i=1}^N \omega_1^i \nabla_\theta \log w_{\theta,\phi}(X_1^i) = \sum_{i=1}^N \omega_1^i \nabla_\theta \log p_\theta(X_1^i, y_1)
$$

and for the second part

$$\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^N \omega_{\theta,\phi}(X_{t-1}^{A_{t-1}^i}, X_t^i)\right) = \sum_{i=1}^N w_t^i \nabla_\theta \log \omega_{\theta,\phi}(X_{t-1}^{A_{t-1}^i}, X_t^i) = \sum_{i=1}^N w_t^i \nabla_\theta \log p_\theta(X_t^i, y_t | X_{t-1}^{A_{t-1}^i})$$

This gives

$$\nabla_\theta \ell^{\mathrm{ELBO}}(\theta,\phi) = \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\sum_{i=1}^N w_1^i \nabla_\theta \log p_\theta(X_1^i, y_1) + \sum_{t=2}^T \sum_{i=1}^N w_t^i \nabla_\theta \log p_\theta(X_t^i, y_t | X_{t-1}^{A_{t-1}^i})\right] \\ + \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\left\{\sum_{t=2}^T \sum_{i=1}^N \nabla_\theta \log w_{t-1}^{A_{t-1}^i}\right\}\right].$$

When we ignore the gradient terms due to resampling as proposed in [41, 33, 38], we only use an unbiased estimate of the first term on the r.h.s., i.e.

$$\sum_{i=1}^N w_1^i \nabla_\theta \log p_\theta(X_1^i, y_1) + \sum_{t=2}^T \sum_{i=1}^N w_t^i \nabla_\theta \log p_\theta(X_t^i, y_t | X_{t-1}^{A_{t-1}^i}), \quad \text{where } (X_{1:T}^{1:N}, A_{1:T-1}^{1:N}) \sim \overline{q}_{\theta,\phi}(\cdot). \tag{45}$$

Under the very mild assumptions given in [16], this quantity will converge in probability as $N \to \infty$ towards

$$\int \nabla_\theta \log p_\theta(x_1, y_1) p_\theta(x_1|y_1) dx_1 + \sum_{t=2}^T \int \nabla_\theta \log p_\theta(x_t, y_t|x_{t-1}) p_\theta(x_{t-1:t}|y_{1:t-1}) dx_{t-1:t}.$$

We recall that the true score is given by Fisher identity and satisfies

$$\int \nabla_\theta \log p_\theta(x_1, y_1) p_\theta(x_1|y_{1:T}) dx_1 + \sum_{t=2}^T \int \nabla_\theta \log p_\theta(x_t, y_t|x_{t-1}) p_\theta(x_{t-1:t}|y_{1:T}) dx_{t-1:t}.$$

This concludes the proof of Proposition 3.2. □

# 9 Appendix D: Proof of Proposition 3.3

In this section, we will prove Proposition 3.3. We introduce some convenient notation in Subsection 9.1 and establish a key auxiliary proposition in Subsection 9.2. Under a mixing assumption, we then analyze the propagation of errors in Subsection 9.3.We thus provide in Subsection 9.4 assumptions which guarantee this bias vanishes as $N \to \infty$ and $\epsilon \to 0$. We then eventually establish Proposition 3.3 in Subsection 9.5.

## 9.1 Filtering and particle filtering

We denote $\{\xi_\theta^{(t)}\}_{t\geq 0}$ the predictive distribution $\xi_\theta^{(t)}(x_t) = p_\theta(x_t|y_{1:t-1})$ for $t \geq 1$ and $\xi_\theta^{(0)}(x_1) = \mu_\theta(x_1)$ while $\{\eta_\theta^{(t)}\}_{t\geq 1}$ denote the filtering distributions; i.e. $\eta_\theta^{(t)}(x_t) = p_\theta(x_t|y_{1:t})$ for $t \geq 1$.

Using this notation, we have

$$\xi_\theta^{(t)}(\psi) = \int \psi(x_t) f_\theta(x_t|x_{t-1}) \eta_\theta^{(t-1)}(x_{t-1}) \mathrm{d}x_{t-1}\mathrm{d}x_t := \eta_\theta^{(t-1)} f_\theta(\psi), \tag{46}$$

$$\eta_\theta^{(t)}(\psi) = \frac{\xi_\theta^{(t)}(g_\theta(y_t|\cdot)\psi)}{\xi_\theta^{(t)}(g_\theta(y_t|\cdot))} = \frac{\eta_\theta^{(t-1)}(f_\theta(g_\theta(y_t|\cdot)\psi))}{\eta_\theta^{(t-1)}(f_\theta(g_\theta(y_t|\cdot)))}. \tag{47}$$

More generally, for a proposal distribution $q_\phi(x_t|x_{t-1}, y_t) \neq f_\theta(x_t|x_{t-1})$, the following recursion holds

$$\eta_\theta^{(t)}(\psi) = \frac{\eta_\theta^{(t-1)}(q_\phi(\omega_{\theta,\phi,t}\,\psi))}{\eta_\theta^{(t-1)}(q_\phi(\omega_{\theta,\phi,t}))} = \mathcal{F}_{\theta,\phi,t}\eta_\theta^{(t-1)}(\psi) \tag{48}$$

where $\mathcal{F}_{\theta,\phi,t}$ will be called the filtering operator and

$$\omega_{\theta,\phi,t}(x_{t-1}, x_t) := \omega_{\theta,\phi}(x_{t-1}, x_t, y_t) = \frac{g_\theta(y_t|x_t) f_\theta(x_t|x_{t-1})}{q_\phi(x_t|x_{t-1}, y_t)}. \tag{49}$$

To simplify the presentation, we will present the analysis in the scenario where $\phi = \theta$ and $q_\phi(x_t|x_{t-1}, y_t) = f_\theta(x_t|x_{t-1})$ so we will analyze (46) for which $\omega_{\theta,\phi,t}(x_{t-1}, x_t) = g_\theta(y_t|x_t)$. All results presented can be extended to the general case. In the interest of notational clarity, we will remove subscript $\theta, \phi$ in further workings. The particle approximations of $\xi^{(t)}$ and $\eta^{(t)}$ are given by the random measures

$$\xi_N^{(t)}(\psi) = \frac{1}{N}\sum_{i=1}^N \psi(X_t^i), \quad \eta_N^{(t)}(\psi) = \sum_{i=1}^N w_t^i \psi(X_t^i), \quad \tilde{\eta}_N^{(t)}(\psi) = \frac{1}{N}\sum_{i=1}^N \psi(\tilde{X}_t^i), \tag{50}$$

where $w_t^i \propto \omega_t(\tilde{X}_{t-1}^i, X_t^i)$ with $\sum_{i=1}^N w_t^i = 1$ and particles are drawn from $X_t^i \sim q(\cdot|\tilde{X}_{t-1}^i, y_t)$.

Here $\eta_N^{(t)}$ denotes the weighted particle approximation of $\eta^{(t)}$ while $\tilde{\eta}_N^{(t)}$ is the uniformly weighted approximation obtained after resampling.

We will use the notation

$$\tilde{\eta}_N^{(t)} = \hat{\mathcal{F}}_t \tilde{\eta}_N^{(t-1)}. \tag{51}$$

## 9.2 Auxiliary results

The aim of this subsection is to prove the following proposition which will be needed to control the error introduced by the particle filter. This proposition upper bounds the bias introduced by a normalized importance sampling approximation.

**Proposition 9.1.** Let $\xi$ and $\{\xi^i\}_{i\in[N]}$ be probability distributions on $\mathcal{X}$ such that the empirical measure $\xi_N(\cdot) = \frac{1}{N}\sum_{i=1}^N \delta_{X^i}(\cdot)$, where $X^i \sim \xi^i$ independently for $i \in [N]$, satisfies $\mathbb{E}[\xi_N(\psi)] = \xi(\psi)$ for any test function $\psi$.
Let $u : \mathcal{X} \to \mathbb{R}$, $v : \mathcal{X} \to (0, \infty)$ denote Borel measurable functions such that: $\sup_{x\in\mathcal{X}} |u(x)| < \infty$, $\sup_{x\in\mathcal{X}} v(x) < \infty$ and $\inf_{x\in\mathcal{X}} v(x) > 0$ then

$$\left|\mathbb{E}\left[\frac{\xi_N(u)}{\xi_N(v)}\right] - \frac{\xi(u)}{\xi(v)}\right| \leq \frac{3ab^2}{N}$$

where $a = \sup_{x',x''\in\mathcal{X}} \left|\frac{u(x')}{v(x')} - \frac{u(x'')}{v(x'')}\right|$ and $b = \sup_{x',x''\in\mathcal{X}} \frac{v(x')}{v(x'')}$.

The proof relies on the following auxiliary Lemma which can be found in [50]. We include it here for the convenience of the reader.

**Lemma 9.1.** For any probability distributions $\xi'$, $\xi''$ on $\mathcal{X}$ and Borel measurable functions $u : \mathcal{X} \to \mathbb{R}$, $v : \mathcal{X} \to (0, \infty)$ such that $\sup_{x\in\mathcal{X}} |u(x)| < \infty$, $\sup_{x\in\mathcal{X}} v(x) < \infty$ and $\inf_{x\in\mathcal{X}} v(x) > 0$, we have

$$\frac{\xi'(v)}{\xi''(v)} \leq b, \qquad\qquad \frac{\xi'(u)}{\xi'(v)} \leq \frac{\xi''(u)}{\xi''(v)} + a,$$

$$\Big|\frac{\xi'(v)}{\xi''(v)} - 1\Big| \leq \frac{\xi'(v)}{\xi''(v)} + 1 \leq 2b, \qquad \Big|\frac{\xi'(u)}{\xi'(v)} - \frac{\xi''(u)}{\xi''(v)}\Big| \leq a,$$

where $a = \sup_{x',x''\in\mathcal{X}} \left|\frac{u(x')}{v(x')} - \frac{u(x'')}{v(x'')}\right|$, $b = \sup_{x',x''\in\mathcal{X}} \frac{v(x')}{v(x'')}$.

*Proof.* From the definition of $b$, we have

$$v(x') \leq bv(x'').$$

Hence, we have

$$\begin{aligned}\xi'(v) &= \iint v(x')\xi'(\mathrm{d}x')\xi''(\mathrm{d}x'')\\ &\leq b\iint v(x'')\xi'(\mathrm{d}x')\xi''(\mathrm{d}x'')\\ &= b\xi''(v).\end{aligned}$$

From the definition of $a$, we have

$$u(x')v(x'') \leq (u(x'') + av(x''))v(x').$$

So it follows that

$$\begin{aligned}\xi'(u)\xi''(v) &= \iint u(x')v(x'')\xi'(\mathrm{d}x')\xi''(\mathrm{d}x'')\\ &\leq \iint (u(x'') + av(x''))v(x')\xi'(\mathrm{d}x')\xi''(\mathrm{d}x'')\\ &= (\xi''(u) + a\xi''(v))\xi'(v).\end{aligned}$$

Therefore, we have

$$\frac{\xi'(v)}{\xi''(v)} \leq b, \qquad \frac{\xi'(u)}{\xi'(v)} \leq \frac{\xi''(u)}{\xi''(v)} + a, \tag{52}$$

and, as $b \geq 1$, it follows that

$$\Big|\frac{\xi'(v)}{\xi''(v)} - 1\Big| \leq \frac{\xi'(v)}{\xi''(v)} + 1 \leq 2b, \tag{53}$$

$$\Big|\frac{\xi'(u)}{\xi'(v)} - \frac{\xi''(u)}{\xi''(v)}\Big| \leq a. \tag{54}$$

□

We are now in position to prove Proposition 9.1.

*Proof.* (Proof of Proposition 9.1) This proof is a generalisation of the proof of [50, Proposition 3.1] which is restricted to the case where the random variables $X^i$ are i.i.d. It relies on the key identity introduced therein

$$\begin{aligned}\mathbb{E}\Big[\frac{\xi_N(u)}{\xi_N(v)}\Big] - \frac{\xi(u)}{\xi(v)} =& \mathbb{E}\Big[\frac{\xi(u)}{\xi(v)}\Big(\frac{\xi_N(v)}{\xi(v)} - 1\Big)^2\Big] - \mathbb{E}\Big[\frac{\xi_N(u)}{\xi(v)}\Big(\frac{\xi_N(v)}{\xi(v)} - 1\Big)\Big]\\ &+ \mathbb{E}\Big[\Big(\frac{\xi_N(u)}{\xi_N(v)} - \frac{\xi(u)}{\xi(v)}\Big)\Big(\frac{\xi_N(v)}{\xi(v)} - 1\Big)^2\Big].\end{aligned} \tag{55}$$

This can be easily checked by expanding the r.h.s. and rearrangement. Applying the triangle inequality and using the inequality (54) from Lemma 9.1 for the third term gives

$$\Big|\mathbb{E}\Big[\frac{\xi_N(u)}{\xi_N(v)}\Big]-\frac{\xi(u)}{\xi(v)}\Big|\leq\Big|\underbrace{\mathbb{E}\Big[\frac{\xi_N(u)}{\xi(v)}\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)\Big]}_{\text{Term 1}}-\underbrace{\frac{\xi(u)}{\xi(v)}\mathbb{E}\Big[\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)^2\Big]}_{\text{Term 2}}\Big| \\ +\underbrace{a\mathbb{E}\Big[\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)^2\Big]}_{\text{Term 3}} \tag{56}$$

We are now going to study those three terms. We will make use repeatedly of the fact that $\mathbb{E}\Big[\xi_N(v)\Big]=\xi(v)$ in the form of $\mathbb{E}\Big[\frac{\xi_N(v)}{\xi(v)}-1\Big]=0$ which follows from the unbiased condition given in the lemma statement.

**Term 1**
By independence, we have

$$\mathbb{E}\Big[\frac{\xi_N(u)}{\xi(v)}\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)\Big]=\underbrace{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]}_{\text{Term 1A}} \\ +\underbrace{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big]\mathbb{E}\Big[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)\Big]}_{\text{Term 1B}}. \tag{57}$$

The first term, **Term 1A**, simplifies as follows

$$\begin{aligned}\mathbb{E}\Big[\frac{\xi_N(u)}{\xi(v)}\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)\Big]&=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]\\&=\frac{1}{N}\mathbb{E}\Big[\int\frac{u(x)}{\xi(v)}\Big(\frac{v(x)}{\xi(v)}-1\Big)\xi_N(\mathrm{d}x)\Big]\\&=\frac{1}{N}\int\frac{u(x)}{\xi(v)}\Big(\frac{v(x)}{\xi(v)}-1\Big)\xi(\mathrm{d}x).\end{aligned} \tag{58}$$

The second term, **Term 1B**, is given by

$$\begin{aligned}&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big]\mathbb{E}\Big[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)\Big]\\=&\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big]\Big[\Big(\sum_{j=1}^{N}\mathbb{E}\Big[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)\Big]\Big)-\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]\Big]\\=&\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big]\mathbb{E}\Big[\Big(1-\frac{v(X^i)}{\xi(v)}\Big)\Big].\end{aligned} \tag{59}$$

**Term 2**
Similarly, we have

$$\mathbb{E}\Big[\Big(\frac{\xi_N(v)}{\xi(v)}-1\Big)^2\Big]=\underbrace{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)^2]}_{\text{Term 2A}}+\underbrace{\frac{1}{N^2}\sum_{i,j\neq i}^{N}\mathbb{E}[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)]\mathbb{E}[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)]}_{\text{Term 2B}}. \tag{60}$$

Again the first term, **Term 2A** satisfies

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)^2]=\frac{1}{N}\int\Big(\frac{v(x)}{\xi(v)}-1\Big)^2\xi(\mathrm{d}x), \tag{61}$$

while **Term 2B** satisfies

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i,j\neq i}^{N}\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]\mathbb{E}\Big[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)\Big] && (62)\\
&=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]\Big[\Big(\sum_{j=1}^{N}\mathbb{E}\Big[\Big(\frac{v(X^j)}{\xi(v)}-1\Big)\Big]\Big)-\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]\Big]\\
&=-\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]^2. && (63)
\end{aligned}$$

Subtracting **Term 2A** from **Term 1A**, and applying Lemma 9.1

$$\begin{aligned}
&\Big|\frac{1}{N}\int\frac{u(x)}{\xi(v)}\Big(\frac{v(x)}{\xi(v)}-1\Big)\xi(\mathrm{d}x)-\frac{1}{N}\int\frac{\xi(u)}{\xi(v)}\Big(\frac{v(x)}{\xi(v)}-1\Big)^2\xi(\mathrm{d}x)\Big|\\
&=\frac{1}{N}\int\frac{g^2(x)}{\xi^2(v)}\Big|\frac{u(x)}{v(x)}-\frac{\xi(u)}{\xi(v)}\Big|\xi(\mathrm{d}x)\leq\frac{ab^2}{N}
\end{aligned} \tag{64}$$

Similarly applying Lemma 9.1 to **Term 3**

$$a\frac{1}{N}\int\Big(\frac{v(x)}{\xi(v)}-1\Big)^2\xi(\mathrm{d}x)\leq a\frac{1}{N}\int\Big(\frac{v(x)}{\xi(v)}\Big)^2\xi(\mathrm{d}x)\leq\frac{ab^2}{N} \tag{65}$$

Subtracting **Term 2B** from **Term 1B** and applying Lemma 9.1

$$\begin{aligned}
&\Big|\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\frac{u(X^i)}{\xi(v)}\Big]\mathbb{E}\Big[\Big(1-\frac{v(X^i)}{\xi(v)}\Big)\Big]+\frac{1}{N^2}\frac{\xi(u)}{\xi(v)}\sum_{i=1}^{N}\mathbb{E}\Big[\Big(\frac{v(X^i)}{\xi(v)}-1\Big)\Big]^2\Big|\\
=\quad&\Big|\frac{1}{N^2}\sum_{i=1}^{N}\Big[\frac{\mathbb{E}[u(X^i)]}{\xi(v)}-\frac{\mathbb{E}[u(X^i)]}{\xi(v)}\frac{\mathbb{E}[v(X^i)]}{\xi(v)}+\frac{\xi(u)}{\xi(v)}\Big(\frac{\mathbb{E}[v(X^i)]^2}{\xi(v)^2}-2\frac{\mathbb{E}[v(X^i)]}{\xi(v)}+1\Big)\Big]\Big|\\
=\quad&\Big|\frac{1}{N^2}\sum_{i=1}^{N}\Big[\frac{\xi(u)}{\xi(v)}\frac{\mathbb{E}[v(X^i)]^2}{\xi(v)^2}-\frac{\mathbb{E}[u(X^i)]}{\xi(v)}\frac{\mathbb{E}[v(X^i)]}{\xi(v)}\Big]\\
&+\frac{1}{N^2}\sum_{i=1}^{N}\Big[\frac{\mathbb{E}[u(X^i)]}{\xi(v)}-2\frac{\xi(u)}{\xi(v)}\frac{\mathbb{E}[v(X^i)]}{\xi(v)}+\frac{\xi(u)}{\xi(v)}\Big]\Big|\\
=\quad&\Big|\frac{1}{N^2}\sum_{i=1}^{N}\Big[\Big(\frac{\xi(u)}{\xi(v)}-\frac{\mathbb{E}[u(X^i)]}{\mathbb{E}[v(X^i)]}\Big)\frac{\mathbb{E}[v(X^i)]^2}{\xi(v)^2}\Big]+\frac{1}{N}\Big[\frac{\xi(u)}{\xi(v)}-2\frac{\xi(u)}{\xi(v)}+\frac{\xi(u)}{\xi(v)}\Big]\Big|\\
=\quad&\Big|\frac{1}{N^2}\sum_{i=1}^{N}\Big[\Big(\frac{\xi(u)}{\xi(v)}-\frac{\xi^i(u)}{\xi^i(v)}\Big)\frac{\xi^i(v)^2}{\xi(v)^2}\Big]+0\Big|\\
\leq\quad&\frac{1}{N^2}\sum_{i=1}^{N}ab^2=\frac{ab^2}{N}.
\end{aligned} \tag{66}$$

Hence, term 1, term 2 and term 3 are each bounded by $\frac{ab^2}{N}$ so the total bound on (56) is

$$\Big|\mathbb{E}\Big[\frac{\xi_N(u)}{\xi_N(v)}\Big]-\frac{\xi(u)}{\xi(v)}\Big|\leq\frac{ab^2}{N}+\frac{ab^2}{N}+\frac{ab^2}{N}=\frac{3ab^2}{N}. \tag{67}$$

□

### 9.3 Particle Approximation Error

We recall here that for two random measures $\rho_1, \rho_2$ on $\mathcal{X}$:

$$|||\rho_1 - \rho_2||| = \sup_{\psi \in \mathrm{BL}} |\mathbb{E}[\rho_1(\psi) - \rho_2(\psi)]|. \tag{68}$$

**Lemma 9.2.** One Step Error.

$$\left|\left|\left|\hat{\mathcal{F}}_k \tilde{\eta}_N^{(k-1)} - \mathcal{F}_k \tilde{\eta}_N^{(k-1)}\right|\right|\right| \leq \frac{6\delta^{-4}}{N} + \left|\left|\left|\eta_N^{(k)} - \tilde{\eta}_N^{(k)}\right|\right|\right|. \tag{69}$$

*Proof.* Recall that $\tilde{\eta}_N^{(k)} = \hat{\mathcal{F}}_k \tilde{\eta}_N^{(k-1)}$ and $\eta^{(k)} = \mathcal{F}_k \eta^{(k-1)}$, hence

$$\mathcal{F}_k \tilde{\eta}_N^{(k-1)}(\psi) = \frac{\tilde{\eta}_N^{(k-1)}(f(\omega_k \psi))}{\tilde{\eta}_N^{(k-1)}(f(\omega_k))}. \tag{70}$$

From (46), the weighted particle approximation, i.e. prior to resampling, may be written

$$\eta_N^{(k)}(\psi) = \frac{\xi_N^{(k)}(\omega_k \psi)}{\xi_N^{(k)}(\omega_k)}. \tag{71}$$

We have the following inequality

$$\begin{aligned}|\mathbb{E}[\mathcal{F}_k \tilde{\eta}_N^{(k-1)}(\psi) - \tilde{\eta}_N^{(k)}(\psi)]| &\leq \underbrace{|\mathbb{E}[\mathcal{F}_k \tilde{\eta}_N^{(k-1)}(\psi) - \eta_N^{(k)}(\psi)]|}_{\text{Normalization Error}} + \underbrace{|\mathbb{E}[\eta_N^{(k)}(\psi) - \tilde{\eta}_N^{(k)}(\psi)]|}_{\text{Resampling Error}} \\ &= \left|\mathbb{E}\Big[\frac{\tilde{\eta}_N^{(k-1)}(f(\omega_k \psi))}{\tilde{\eta}_N^{(k-1)}(f(\omega_k))} - \frac{\xi_N^{(k)}(\omega_k \psi)}{\xi_N^{(k)}(\omega_k)}\Big]\right| + |\mathbb{E}[\eta_N^{(k)}(\psi) - \tilde{\eta}_N^{(k)}(\psi)]|.\end{aligned} \tag{72}$$

To bound the normalization error, recall that the predictive particle approximation is $\xi_N^{(k)}(\cdot) = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^i}(\cdot)$ where $X_k^i \sim f(\cdot|\tilde{X}_{k-1}^i)$, and $\tilde{\eta}_N^{(k-1)} = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}_{k-1}^i}$. We have

$$\mathbb{E}[\xi_N^{(k)}(\psi)] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^N \psi(X_k^i)] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^N \int \psi(x_k^i) f(x_k^i|\tilde{X}_{k-1}^i) \mathrm{d}x_k^i] = \tilde{\eta}_N^{(t-1)}(f(\psi)).$$

One may therefore apply Lemma 9.1 to the first term of equation (72), hence we have for $\psi \in \mathrm{BL}$

$$\left|\mathbb{E}\Big[\frac{\tilde{\eta}_N^{(t-1)}(f(\omega_t \psi))}{\tilde{\eta}_N^{(t-1)}(f(\omega_t))} - \frac{\xi_N^{(t)}(\omega_t \psi)}{\xi_N^{(t)}(\omega_t)}\Big]\right| \leq \frac{6\delta^{-4}}{N}. \tag{73}$$

Combining equations (73) to (72) gives the result. □

**Assumption 9.1.** *Mixing.* There exists $\delta > 0$ and a probability density $\nu$ on $\mathcal{X}$ such that for all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$\delta \nu(x') \leq f(x'|x) \leq \delta^{-1} \nu(x') \quad \text{and} \quad \delta \leq g(y|x) \leq \delta^{-1}.$$

Although Assumption 9.1 is restrictive, it is also standard and widely used in the literature on PFs since its introduction in [15].

**Lemma 9.3.** Particle approximation under filter stability.
Under Assumption 9.1, there exists finite $K, C$ independent of $t, N$ such that

$$\left|\left|\left|\tilde{\eta}_N^{(t)} - \eta^{(t)}\right|\right|\right| \leq K\left(\frac{C}{N} + \sup_{k \leq t}\left|\left|\left|\eta_N^{(k)} - \tilde{\eta}_N^{(k)}\right|\right|\right|\right). \tag{74}$$

*Proof.* Given the mixing Assumption 9.1, one may appeal to the exponential forgetting properties of the (true) optimal filter to obtain "better" error bounds for the PF. This was first proposed by Del Moral; see e.g. [15, Chapter 7]. We follow here a presentation close to the one in [53, Chapter 5]. We can easily extend Lemma 5.2, Theorem 5.4 and Theorem 5.6 of [53] to (35). This allows us to obtain

$$\begin{aligned}
\left|\left|\left|\tilde{\eta}_N^{(t)} - \eta^{(t)}\right|\right|\right| &\le \sum_{k=1}^{t} \delta^{-2}(1-\delta^2)^{t-k} \left|\left|\left|\mathcal{F}_k \tilde{\eta}_N^{(k-1)} - \tilde{\eta}_N^{(k)}\right|\right|\right| \\
&\le \delta^{-4} \sup_{k\le t} \left|\left|\left|\mathcal{F}_k \tilde{\eta}_N^{(k-1)} - \tilde{\eta}_N^{(k)}\right|\right|\right| \\
&\le \delta^{-4} \sup_{k\le t} \left|\left|\left|\mathcal{F}_k \tilde{\eta}_N^{(k-1)} - \hat{\mathcal{F}}_k \tilde{\eta}_N^{(k-1)}\right|\right|\right| \\
&\le \delta^{-4} \left( \frac{6\delta^{-4}}{N} + \sup_{k\le t} \left|\left|\left|\eta_N^{(k)} - \tilde{\eta}_N^{(k)}\right|\right|\right| \right),
\end{aligned}$$

where we recall that, by definition, $\eta^{(k)} = \mathcal{F}_k \eta^{(k-1)}$ and $\hat{\mathcal{F}}_k$ by $\tilde{\eta}_N^{(k)} = \hat{\mathcal{F}}_k \tilde{\eta}_N^{(k-1)}$. We use the convention $\mathcal{F}_0 \tilde{\eta}_N^{(0)} = \mu$. The last inequality results from Lemma 9.2.

□

### 9.4 Control of DET errors

We provide here some assumptions allowing us to control terms of the form $\left|\left|\left|\eta_N^{(k)} - \tilde{\eta}_N^{(k)}\right|\right|\right|$. These assumptions are similar to the ones made in Subsection 7.1 where $\eta_N^{(t)}$ and $\xi_N^{(t)}$ play the role of $\alpha_N$ and $\beta_N$ respectively. We will also denote $\mathcal{P}_t^{\mathrm{OT},N}$ an optimal transport between $\eta_N^{(t)}$ and $\xi_N^{(t)}$.

**Assumption 9.2.** *Random Measures and Optimal Transport Regularity.*
**A**. For $t \in [T]$, there exist probability measures $\breve{\eta}_\epsilon^{(t)}, \breve{\xi}_\epsilon^{(t)}$ with $\breve{\eta}_\epsilon^{(t)}(\psi) = \breve{\xi}_\epsilon^{(t)}(\omega_t \psi)/\breve{\xi}_\epsilon^{(t)}(\omega_t)$ such that $\eta_N^{(t)} \underset{a.s.}{\rightsquigarrow} \breve{\eta}_\epsilon^{(t)}$ and $\xi_N^{(t)} \underset{a.s.}{\rightsquigarrow} \breve{\xi}_\epsilon^{(t)}$.

**B**. For $t \in [T]$, the OT problem

$$\mathrm{OT}(\breve{\eta}_\epsilon^{(t)}, \breve{\xi}_\epsilon^{(t)}) = \min_{\mathcal{P} \in \mathcal{U}(\breve{\eta}_\epsilon^{(t)}, \breve{\xi}_\epsilon^{(t)})} \mathbb{E}_{(U,V)\sim\mathcal{P}} \left[ ||U - V||^2 \right]$$

admits a unique solution $\mathcal{P}_t^{\mathrm{OT}}$ which is given by a deterministic map $\mathbf{T}_t : \mathcal{X} \to \mathcal{X}$.

**C**. For any $\psi \in \mathrm{BL}$, we have $\xi_N^{(t)}(\psi \odot \mathbf{T}_t) \underset{a.s.}{\to} \breve{\xi}_\epsilon^{(t)}(\psi \odot \mathbf{T}_t) := \breve{\eta}_\epsilon^{(t)}(\psi(\mathbf{T}_t(x))) = \breve{\eta}_\epsilon^{(t)}(\psi)$.

**Assumption 9.3.** *Growth.* For $t \in [T]$, we have

$$\lim_{N\to\infty} \sup \mathbb{E}[\xi_N^{(t)}(x \to ||x||^2)] < \infty,$$

and there exists some $p > 1$ such that

$$\lim_{N\to\infty} \sup \mathbb{E}[\xi_N^{(t)}(x \to ||\mathbf{T}_t(x)||^p) + \xi_N^{(t)}(x \to ||x||^p)] < \infty.$$

**Assumption 9.4.** For $t \in [T]$, there exists a subsequence of $N$ such that $\mathcal{P}_t^{\mathrm{OT},N} \underset{a.s.}{\rightsquigarrow} \mathcal{P}_t^{\mathrm{OT}}$ along this subsequence.

It is beyond the scope of this paper to give sufficient conditions for Assumption 9.2-A to be satisfied. However, we note that the limiting measures $\breve{\xi}_\epsilon^{(t)}$ and $\breve{\eta}_\epsilon^{(t)}$ appearing therein are not the true predictive and filtering distributions $\xi^{(t)}$ and $\eta^{(t)}$. They are instead, under additional regularity conditions, defined by the following recursion

$$\eta_\epsilon^{(t)}(\psi) = \frac{\xi_\epsilon^{(t)}(\omega_t\psi)}{\xi_\epsilon^{(t)}(\omega_t)}, \quad \eta_\epsilon^{\mathrm{DET},(t)}(\psi) = \xi_\epsilon^{(t)}(\psi \odot \mathbf{T}_t^{\mathrm{DET},\epsilon}), \quad \xi_\epsilon^{(t+1)}(\psi) = \eta_\epsilon^{\mathrm{DET},(t)} f(\psi), \tag{75}$$

initialized at $\xi_1 = \mu$ where $\mathbf{T}_t^{\mathrm{DET},\epsilon}$ is the DET map defined as follows. Given $\xi_\epsilon^{(t)}$ and $\eta_\epsilon^{(t)}$, we denote the entropy-regularized OT $\mathcal{P}_t^{\mathrm{OT},\epsilon}$ between these two measures and set

$$\mathbf{T}_t^{\mathrm{DET},\epsilon}(x) = \int x' \mathcal{P}_t^{\mathrm{OT},\epsilon}(\mathrm{d}x'|x), \tag{76}$$

where $\eta_\epsilon^{(t)}(\mathrm{d}x') = \int \mathcal{P}_t^{\mathrm{OT},\epsilon}(\mathrm{d}x'|x)\xi_\epsilon^{(t)}(\mathrm{d}x)$.

**Proposition 9.2.** Let Assumptions 9.2, 9.3 and 9.4 hold. Let $\tilde{\mathbf{X}}_t = N\mathbf{P}_\epsilon^{\mathrm{OT}}\mathbf{X}_t$ be the DET at time $t \in [T]$. Here $\mathbf{P}_\epsilon^{\mathrm{OT}}$ is the regularized OT between $\eta_N^{(t)}$ and $\xi_N^{(t)}$ and recall that $\tilde{\eta}_N^{(t)} = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}_t^i}$. Then, for any $\psi \in \mathrm{BL}$, there exists an increasing sequence $(N_{t,k})_{k\geq 1}$ such that

$$\left|\mathbb{E}\left[\tilde{\eta}_{N_{t,k}}^{(t)}(\psi) - \eta_{N_{t,k}}^{(t)}(\psi)\right]\right| \leq \kappa_{t,N_{t,k}}(\psi) + \mathcal{E}_{t,N_{t,k},\epsilon}.$$

Here $\kappa_{t,N_{t,k}}(\psi)$ denotes the error from the ET, which is such that $\kappa_{t,N_{t,k}}(\psi) \underset{k\to\infty}{\longrightarrow} 0$, and $\mathcal{E}_{t,N_{t,k},\epsilon}$ denotes the additional error introduced by the DET, which is such that $\mathcal{E}_{t,N_{t,k},\epsilon} \underset{\epsilon\to 0}{\longrightarrow} 0$.

*Proof.* This is a direct application of Proposition 7.1 to the case where $\alpha_N = \eta_N^{(t)}$, $\tilde{\alpha}_N = \tilde{\eta}_N^{(t)}$, $\beta_N = \xi_N^{(t)}$. The statement emphasizes that, from inspection of the proof of Proposition 7.1, $\kappa_{t,N_{t,k}}$ is a function of $\psi$ while $\mathcal{E}_{t,N_{t,k},\epsilon}$ is not. □

**Assumption 9.5.** Let Assumptions 9.2, 9.3 and 9.4 hold then $\sup_{\psi\in\mathrm{BL}} \kappa_{t,N_{t,k}}(\psi) \underset{k\to\infty}{\longrightarrow} 0$.

**Proposition 9.3.** Let Assumptions 9.2, 9.3, 9.4 and 9.5 hold then there exists an increasing sequence $(N_k)_{k\geq 1}$ such that

$$\sup_{t\leq T} \left|\left|\left|\eta_{N_k}^{(t)} - \tilde{\eta}_{N_k}^{(t)}\right|\right|\right| \leq \bar{\kappa}_{N_k} + \bar{\mathcal{E}}_{N_k,\epsilon}, \tag{77}$$

where $\bar{\kappa}_{N_k} \underset{k\to\infty}{\longrightarrow} 0$ and $\bar{\mathcal{E}}_{N_k,\epsilon} \underset{\epsilon\to 0}{\longrightarrow} 0$.

*Proof.* Assume $T = 1$ for the time being, we can simply apply Proposition 9.2 directly as Assumptions 9.2, 9.3, 9.4 hold and then using Assumption 9.5, one may set $N_k = N_{1,k}$, $\bar{\kappa}_{N_k} = \sup_{\psi\in\mathrm{BL}} \kappa_{1,N_{1,k}}(\psi)$ and $\bar{\mathcal{E}}_{N_k,\epsilon} = \bar{\mathcal{E}}_{1,N_k,\epsilon}$. If $T = 2$, we can extract a subsequence of $(N_{1,k})_{k\geq 1}$ called abusively $N_k$ again for which $\sup_{\psi\in\mathrm{BL}} \kappa_{2,N_k}(\psi) \to 0$ holds and define $\bar{\kappa}_{N_k} = \max\{\sup_{\psi\in\mathrm{BL}} \kappa_{1,N_k}(\psi), \sup_{\psi\in\mathrm{BL}} \kappa_{2,N_k}(\psi)\}$ and $\bar{\mathcal{E}}_{N_k,\epsilon} = \max\{\bar{\mathcal{E}}_{1,N_k,\epsilon}, \bar{\mathcal{E}}_{2,N_k,\epsilon}\}$. By induction, we can prove the result for any $T$.

□

## 9.5 Bound on the likelihood bias

We provide here a precise statement of Proposition 3.3. We emphasize the dependence on $N$ of the log-likelihood estimator, whose expectation gives $\ell_\epsilon^{\mathrm{ELBO}}(\theta, \phi)$. Recall that here $\theta, \phi$ are fixed.

**Proposition 9.4.** Under Assumption 9.1, we have

$$\frac{1}{T}\left|\mathbb{E}[\log \hat{p}_N(y_{1:T})] - \log p(y_{1:T})\right| \leq K\left[\frac{C}{N} + \sup_{t\leq T}\left|\left|\left|\eta_N^{(t)} - \tilde{\eta}_N^{(t)}\right|\right|\right|\right] \tag{78}$$

for some finite constants $C$ and $K$ independent of $T$.
Additionally, if Assumptions 9.2 to 9.5 hold, then there exists an increasing sequence $(N_k)_{k\geq 1}$ such that

$$\frac{1}{T}|\mathbb{E}[\log \hat{p}_{N_k}(y_{1:T})] - \log p(y_{1:T})| \leq K\Big[\frac{C}{N_k} + \bar{\kappa}_{N_k} + \bar{\mathcal{E}}_{N_k,\epsilon}\Big], \tag{79}$$

where $\bar{\kappa}_{N_k} \xrightarrow[k\to\infty]{} 0$ and $\bar{\mathcal{E}}_{N_k,\epsilon} \xrightarrow[\epsilon\to 0]{} 0$.

*Proof.* We first establish (78). Let $\mathcal{A}_t = \left\{\{X_t^i\}_{i=1}^N, \{w_t^i\}_{i=1}^N\right\}$ be the particle locations and weights at time $t$. Note that $|\log(x) - \log(y)| \leq \frac{|x-y|}{\min\{x,y\}}$ for any $x, y > 0$ so

$$\begin{aligned}
|\mathbb{E}[\log \hat{p}(y_{1:T})] - \log p(y_{1:T})| &\leq \sum_{t=1}^{T} |\mathbb{E}[\log \hat{p}(y_t|y_{1:t-1}) - \log p(y_t|y_{1:t-1}]| \\
&\leq \sum_{t=1}^{T} |\mathbb{E}\left[\frac{\hat{p}(y_t|y_{1:t-1}) - p(y_t|y_{1:t-1})}{\min(\hat{p}(y_t|y_{1:t-1}), p(y_t|y_{1:t-1}))}\right]| \\
&\leq \delta \sum_{t=1}^{T} |\mathbb{E}\big[\mathbb{E}[\hat{p}(y_t|y_{1:t-1})|\mathcal{A}_{t-1}] - p(y_t|y_{1:t-1})\big]| \qquad (80)
\end{aligned}$$

where $\delta$ is defined in Assumption 9.1.

The inner expectation in line (80) may be written as follows

$$\begin{aligned}
&\mathbb{E}[\hat{p}(y_t|y_{1:t-1})|\mathcal{A}_{t-1}] - p(y_t|y_{1:t-1}) \\
&= \iint g(y_t|x_t) f(\mathrm{d}x_t|\tilde{x}_{t-1})\tilde{\eta}_N^{(t-1)}(\mathrm{d}\tilde{x}_{t-1}) - \iint g(y_t|x_t) f(\mathrm{d}x_t|\tilde{x}_{t-1})\tilde{\eta}^{(t-1)}(\mathrm{d}\tilde{x}_{t-1}) \\
&= \tilde{\eta}_N^{(t-1)}(h) - \eta^{(t-1)}(h)
\end{aligned}$$

for $\delta^2 \leq h(x) := \int g(y_t|x')f(x'|x)\mathrm{d}x' \leq \delta^{-2}$.

Applying Lemma 9.3 with Assumption 9.1

$$\left|\mathbb{E}[\tilde{\eta}_N^{(t-1)}(h)] - \eta^{(t-1)}(h)\right| \leq C_1\Big(\frac{C_2}{N} + \sup_{k\leq t}\left|\left|\left|\eta_N^{(k)} - \tilde{\eta}_N^{(k)}\right|\right|\right|\Big) \tag{81}$$

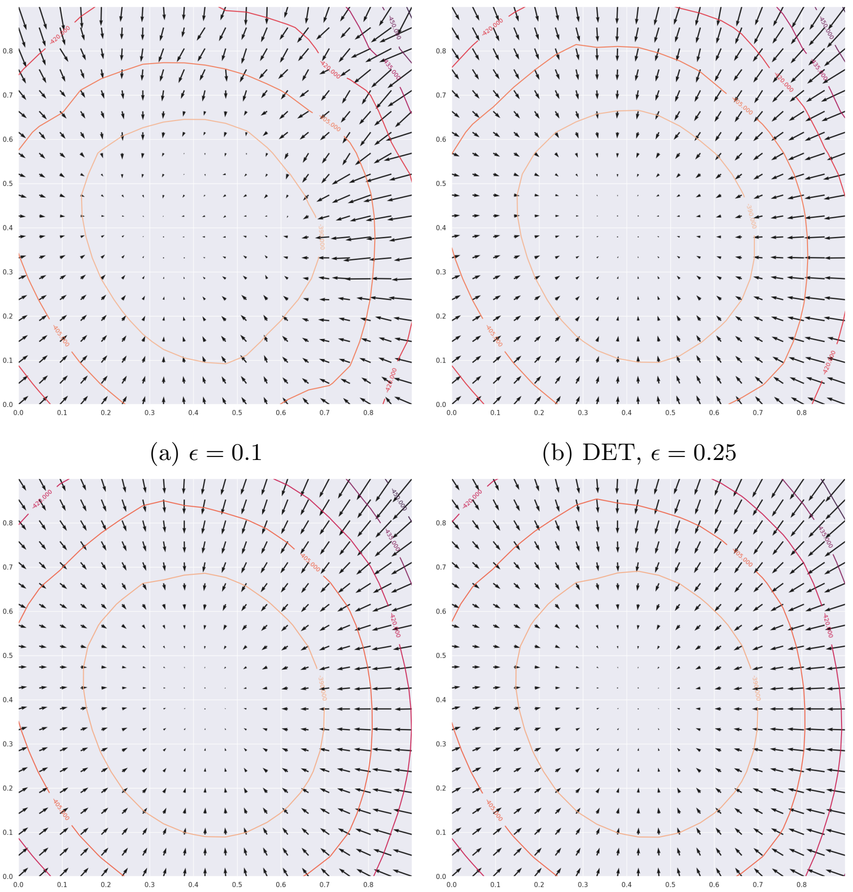
for some constants $C, K$. From there, we obtain (78).

The bound (79) follows directly from (78) and Proposition 9.3. □

# 10 Appendix E: Further Experimental Results

## 10.1 Linear state-space model

We now reproduce the results of Figure 2 for different regularization parameters $\epsilon$ to assess its effect on the estimated log-likelihood $\hat{\ell}_\epsilon(\theta, \mathbf{u})$: in practice anything above 0.25 provides numerically stable gradients estimates. However, Figure 5 shows that taking a too high regularization parameter will collapse the distribution to its weighted average. As a tradeoff, we picked $\epsilon = 0.5$ which provides stable results in all our experiments.



(a) $\epsilon = 0.1$

(b) DET, $\epsilon = 0.25$

(c) DET, $\epsilon = 0.75$

(d) DET, $\epsilon = 1.0$

Figure 6: DET vector field for $\hat{\ell}_\epsilon(\theta; \mathbf{u})$ for different values of $\epsilon$

## 10.2 VRNN Details

**Model Details**
Recall the VRNN is given by

$$\begin{aligned} (R_{t+1}, O_{t+1}) &= \text{RNN}_\theta(R_t, Y_{1:t-1}, E_\theta(Z_t)) & \hat{p}_t &= h_\theta(E_\theta(Z_t), O_t) \\ Z_{t+1} &\sim \mathcal{N}(\mu_\theta(O_{t+1}), \sigma_\theta(O_{t+1})) & Y_t | X_t &\sim \text{Ber}(\hat{p}_t) \end{aligned}$$

where $E_\theta$, $h_\theta$, $\mu_\theta$, $\sigma_\theta$ single layer neural networks. The corresponding graphical model is given in Figure 7.

Recall $Z_t, R_t, Y_t$ are of dimension 10, 5 and 88 respectively. $E_\theta$ has a hidden layer of size 32 and output of dimension 32. Networks $\mu_\theta$ and $\sigma_\theta$ share a hidden layer of dimension 10, each with output of dimension $d_z = 10$ which parametrize the mean and diagonal matrix of an isotropic Gaussian distribution. The *softplus* function is applied to the output of $\sigma_\theta$. $\text{RNN}_\theta$ is an LSTM cell with hidden size 5. $h_\theta$ has hidden layer of size 88 and output dimension 88.

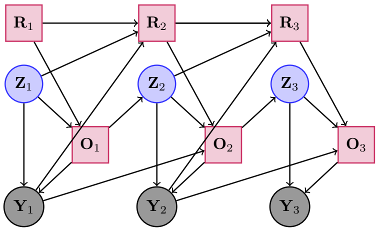


Figure 7: VRNN Graphical Model where observations are grey nodes, deterministic state associated to the RNN are red and stochastic latent states are blue

**Data Sources**

Similar to [38] and [6], we assessed our algorithm on the following polyphonic music datasets. At each time point in the sequence $t$, observations $Y_t$ is of length 88 for each dataset.

- JSB chorales, sequence length 129
- MuseData library of classical piano music, length 127
- Nottingham folk music, sequnce length 192
- Piano-Midi.de[7], sequence length 347

**Further Results**

$\ell^{\text{ELBO}}$ was used to train the VRNN using multinomial sampling (MUL) with biased gradients, ignoring the resampling terms in the gradient. $\ell^{\text{ELBO}}_{\epsilon}$ was the objective used to train the VRNN with the differentiable ensemble transform (DET) resampling, where $\epsilon = 0.5$. 25 particles were used and trained for $10,000$ iterations.

We present $\frac{1}{T}\ell^{\text{ELBO}}$ and $\frac{1}{T}\ell^{\text{ELBO}}_{\epsilon}$ in Table 3 after training and present the cross entropy loss for the one-step prediction of the observation at time $T$ in Table 4, which is equivalent to the negative observational log-likelihood. Although ELBO is marginally higher for the DET method, the predictive loss is clearly lower.

Table 3: ELBO per unit time

| | **MUL** | **DET** |
|---:|---|---|
| **jsb** | -2.86 | -2.61 |
| **musedata** | -1.76 | -0.72 |
| **nottingham** | -2.98 | -2.47 |
| **piano-midi.de** | -5.44 | -5.22 |

Table 4: Cross Entropy Loss

| | **MUL** | **DET** |
|---:|---|---|
| **jsb** | 1.77 | 0.53 |
| **musedata** | 2.43 | 0.16 |
| **nottingham** | 1.04 | 0.73 |
| **piano-midi.de** | 0.96 | 0.37 |

## 10.3 Multivariate stochastic volatility model

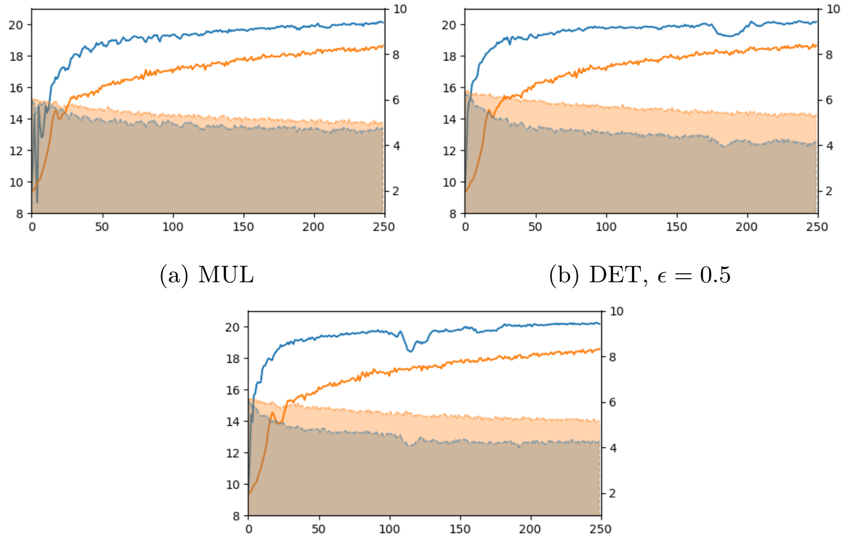
We now consider a multivariate Stochastic Volatility (SV) model (see [21] for a comprehensive introduction). The observed security (index price, Forex level, etc.) log-returns $Y_t$ are noisy observations of a latent Gaussian process $X_t$:

$$X_t|X_{t-1} \sim \mathcal{N}(\mu + \mathbf{\Phi}(X_{t-1} - \mu), \mathbf{\Sigma_X}), \quad Y_t|X_t \sim \exp(X_t/2)\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_Y}) \tag{82}$$

[7] http://www.piano-midi.de/

Here we fix $\mathbf{\Sigma_Y} = I_5$, set $\theta = (\mu, \mathbf{\Phi}, \mathbf{\Sigma_X})$ and we use $q_\phi(x_t|x_{t-1}, y_t) = f_\theta(x_t|x_{t-1})$, and we use $N = 10$ particles. We consider the 5 dimensional time series of log-return of the currency pairs EUR/AUD-GBP-CAD-CHF-USD and learn $\theta$ using stochastic gradient ascent using Adam with learning rate $10^{-3}$ and $10^{-2}$ on data from 2019-06-01 to 2020-01-01 which corresponds to $T = 150$ observations.

The results are shown in Figure 8. While the end result of the calibration was the same for the three methods used, PF MUL exhibited empirically more instability and, contrarily to PF DET, did not converge with the same learning rates when using a higher number of particles or when changing the initial seed $\mathbf{u}$ to optimize $\ell^{\text{ELBO}}(\theta)$.



(a) MUL

(b) DET, $\epsilon = 0.5$

(c) Variance Adjusted DET, $\epsilon = 0.5$

Figure 8: ELBO maximization for the stochastic volatility model (82), blue is with learning rate $10^{-2}$, orange $10^{-3}$. Left axis and line is $\frac{1}{T}\ell^{\text{ELBO}}(\theta)$, right axis and shade is ESS. MUL ran in approx. 1mn while DET and V-DET ran in approx. 2mn30s.