A. Proof of Proposition 4.1

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

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

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

The ELBO is given by

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

We now compute $\nabla_\theta \ell^{\text{ELBO}}(\theta, \phi)$. We assume from now on that the regularity conditions allowing us to swap the expectation and differentiation operators are satisfied as in (Maddison et al., 2017; Le et al., 2018; Naesseth et al., 2018). We can split the gradient using the product rule and apply the log-derivative trick:

$$\begin{aligned}
\nabla_\theta \ell^{\text{ELBO}}(\theta, \phi) =& \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\nabla_\theta \log \widehat{p}_\theta(y_{1:T})\right] + \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T}) \nabla_\theta \log \overline{q}_{\theta,\phi}(X_{1:T}^{1:N}, A_{1:T-1}^{1:N})\right] \\
=& \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^{N}\omega_{\theta,\phi}(X_1^i, y_1)\right) + \sum_{t=2}^{T}\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^{N}\omega_{\theta,\phi}(X_{t-1}^{A_{t-1}^i}, X_t^i, y_t)\right)\right] \quad (17) \\
&+ \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\left\{\sum_{t=2}^{T}\sum_{i=1}^{N}\nabla_\theta \log w_{t-1}^{A_{t-1}^i}\right\}\right] \quad (18)
\end{aligned}$$

For the first part of the ELBO gradient (17), we have

$$\nabla_\theta \log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(X_1^i), y_1\right) = \sum_{i=1}^{N} w_1^i \nabla_\theta \log w_{\theta,\phi}(X_1^i, y_1) = \sum_{i=1}^{N} \omega_1^i \nabla_\theta \log p_\theta(X_1^i, y_1)$$

and

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

This gives

$$\begin{aligned}
\nabla_\theta \ell^{\text{ELBO}}(\theta, \phi) &= \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\sum_{i=1}^{N} w_1^i \nabla_\theta \log p_\theta(X_1^i, y_1) + \sum_{t=2}^{T}\sum_{i=1}^{N} w_t^i \nabla_\theta \log p_\theta(X_t^i, y_t | X_{t-1}^{A_{t-1}^i})\right] \quad (19) \\
&+ \mathbb{E}_{\overline{q}_{\theta,\phi}}\left[\log \widehat{p}_\theta(y_{1:T})\left\{\sum_{t=2}^{T}\sum_{i=1}^{N}\nabla_\theta \log w_{t-1}^{A_{t-1}^i}\right\}\right]. \quad (20)
\end{aligned}$$

When we ignore the gradient terms due to resampling corresponding to (20) as proposed in (Naesseth et al., 2018; Le et al., 2018; Maddison et al., 2017; Hirt & Dellaportas, 2019), we only use an unbiased estimate of the first term (19), i.e.

$$\hat{\nabla}_\theta \ell^{\text{ELBO}}(\theta, \phi) \coloneqq \sum_{i=1}^{N} w_1^i \nabla_\theta \log p_\theta(X_1^i, y_1) + \sum_{t=2}^{T}\sum_{i=1}^{N} w_t^i \nabla_\theta \log p_\theta(X_t^i, y_t | X_{t-1}^{A_{t-1}^i}), \quad \text{where } (X_{1:T}^{1:N}, A_{1:T-1}^{1:N}) \sim \overline{q}_{\theta,\phi}(\cdot). \quad (21)$$

Now we assume that the mild assumptions ensuring almost sure convergence of the PF estimates are satisfied (see e.g. (Del Moral, 2004)). Under these assumptions, the estimator (21) converges almost surely as $N \to \infty$ towards

$$\int \nabla_\theta \log p_\theta(x_1, y_1) p_\theta(x_1|y_1) \mathrm{d}x_1 + \sum_{t=2}^{T} \int \nabla_\theta \log p_\theta(x_t, y_t|x_{t-1}) p_\theta(x_{t-1:t}|y_{1:t-1}) \mathrm{d}x_{t-1:t}. \tag{22}$$

Under an additional uniform integrability condition on $\hat{\nabla}_\theta \ell^{\mathrm{ELBO}}(\theta, \phi)$, we thus have that $\mathbb{E}_{\overline{q}_{\theta,\phi}}[\hat{\nabla}_\theta \ell^{\mathrm{ELBO}}(\theta, \phi)]$ converges towards (22). We recall that the true score is given by Fisher's identity and satisfies

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

This concludes the proof of Proposition 4.1.

B. Notation and Assumptions

B.1. Filtering Notation

Recall $\mathcal{X} = \mathbb{R}^{d_x}$, denote the Borel sets of $\mathcal{X}$ by $\mathcal{B}(\mathcal{X})$ 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 ψ. In the interest of notational clarity, we will remove subscript θ, ϕ where unnecessary in further workings.

We denote $\{\alpha^{(t)}\}_{t\geq 0}$ the predictive distributions where $\alpha^{(t)}(x_t) = p(x_t|y_{1:t-1})$ for $t > 1$ and $\alpha^{(1)}(x_1) = \mu(x_1)$ while $\{\beta^{(t)}\}_{t\geq 1}$ denotes the filtering distributions; i.e. $\beta^{(t)}(x_t) = p(x_t|y_{1:t})$ for $t \geq 1$.

Using this notation, we have

$$\alpha^{(t)}(\psi) = \int \psi(x_t) f(x_t|x_{t-1}) \beta^{(t-1)}(x_{t-1}) \mathrm{d}x_{t-1} \mathrm{d}x_t \coloneqq \beta^{(t-1)} f(\psi), \tag{23}$$

$$\beta^{(t)}(\psi) = \frac{\alpha^{(t)}(g(y_t|\cdot)\psi)}{\alpha^{(t)}(g(y_t|\cdot))} = \frac{\beta^{(t-1)}(f(g(y_t|\cdot)\psi))}{\beta^{(t-1)}(f(g(y_t|\cdot)))}. \tag{24}$$

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

$$\beta^{(t)}(\psi) = \frac{\beta^{(t-1)}(q(\omega_t\, \psi))}{\beta^{(t-1)}(q(\omega_t))} \tag{25}$$

$$\omega_t(x_{t-1}, x_t) \coloneqq \omega(x_{t-1}, x_t, y_t) = \frac{g(y_t|x_t) f(x_t|x_{t-1})}{q(x_t|x_{t-1}, y_t)}. \tag{26}$$

To simplify the presentation, we will present the analysis in the scenario where $\phi = \theta$ and $q(x_t|x_{t-1}, y_t) = f(x_t|x_{t-1})$ so we will analyze (23) for which $\omega_t(x_{t-1}, x_t) = g(y_t|x_t)$. In this case, the particle approximations of μ is denoted μ_N and for $t > 1$, $\alpha^{(t)}$ and $\beta^{(t)}$ are given by the random measures

$$\alpha_N^{(t)}(\psi) = \frac{1}{N}\sum_{i=1}^{N} \psi(X_t^i), \quad \beta_N^{(t)}(\psi) = \sum_{i=1}^{N} w_t^i \psi(X_t^i), \quad \tilde{\beta}_N^{(t)}(\psi) = \frac{1}{N}\sum_{i=1}^{N} \psi(\tilde{X}_t^i), \tag{27}$$

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

Here $\beta_N^{(t)}$ denotes the weighted particle approximation of $\beta^{(t)}$ while $\tilde{\beta}_N^{(t)}$ is the uniformly weighted approximation obtained after the DET transformation described in Section 3.2.

B.2. Optimal Transport Notation

Recall from Section 2.1, $\mathcal{P}_t^{\mathrm{OT}}$ denotes a transport between $\alpha^{(t)}$ and $\beta^{(t)}$ with accompanying map $\mathbf{T}^{(t)}$. $\mathcal{P}_t^{\mathrm{OT},N}$ denotes an optimal transport between particle approximations $\alpha_N^{(t)}$ and $\beta_N^{(t)}$ with corresponding transport matrix, $\mathbf{P}^{\mathrm{OT}}$ with i,j entry $p_{i,j}^{\mathrm{OT}}$. To simplify notation, we remove script t when not needed.

Similarly from Section 3.1, $\mathcal{P}_\epsilon^{\mathrm{OT,N}}$ denotes the regularized transport between $\alpha_N^{(t)}$ and $\beta_N^{(t)}$ with accompanying matrix $\mathbf{P}_\epsilon^{\mathrm{OT}}$ with i,j entry $p_{\epsilon,i,j}^{\mathrm{OT}}$. Recall $\tilde{\beta}_N^{(t)} = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}^i}$ is the uniformly weighted particle approximation for $\beta^{(t)}$ under the DET, i.e. $\tilde{X}^i = \mathbf{T}_{N,\epsilon}^{(t)}(X^i) = \int y \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}y|x^i)$. Note that $\tilde{X}_{N,\epsilon}^i$ will be used where necessary to avoid ambiguity when comparing to other resampling schemes.

Recall also for $p > 0$:

$$\mathcal{W}_p^p(\alpha,\beta) = \min_{\mathcal{P}\in\mathcal{U}(\alpha,\beta)} \mathbb{E}_{(U,V)\sim\mathcal{P}}\left[||U-V||^p\right] \tag{28}$$

where $\mathcal{U}(\alpha,\beta)$ is the collection of couplings with marginals α and β.

B.3. Assumptions

Our results will rely on the following four assumptions.

Assumption B.1. $\mathcal{X} \subset \mathbb{R}^d$ is a compact subset with diameter

$$\mathfrak{d} := \sup_{x,y\in\mathcal{X}} |x-y|.$$

Assumption B.2. There exists $\kappa \in (0,1)$ such that for any two probability measures π, ρ on $\mathcal{X}$

$$\mathcal{W}_k(\pi f, \rho f) \leq \kappa \mathcal{W}_k(\pi,\rho), \qquad k = 1,2.$$

Assumption B.3. The weight function $\omega^{(t)} : \mathcal{X} \to [\Delta, \Delta^{-1}]$ is 1-Lipschitz for all t.

Assumption B.4. There exists a $\lambda > 0$, such that for all $t \geq 0$ the unique optimal transport plan between $\alpha^{(t)}$ and $\beta^{(t)}$ is given by a deterministic, λ-Lipschitz map $\mathbf{T}^{(t)}$.

C. Auxiliary Results and Proof of Proposition 4.2

We start by establishing a couple of key auxiliary results which will be then used subsequently to establish Proposition 4.2.

C.1. Auxiliary Results

As per section 2.1, let $\mathcal{S}(\alpha_N, \beta_N)$ denote the collection of coupling matrices between $\alpha_N = \sum_{i=1}^N a_i \delta_{Y^i}$ with $a_i > 0$ and $\beta_N = \sum_{i=1}^N b_i \delta_{X^i}$. We also denote entropy by H where $H(\mathbf{P}) = \sum_{i,j} p_{i,j}\log(1/p_{i,j})$ for $\mathbf{P} = (p_{i,j})_{i,j} \in \mathcal{S}(\alpha_N,\beta_N)$.

Lemma C.1. *The entropic radius, R_H, of simplex $\mathcal{U}(\alpha_N,\beta_N)$ may be bounded above as follows*

$$R_H := \max_{\mathbf{P}_1,\mathbf{P}_2\in\mathcal{S}(\alpha_N,\beta_N)} H(\mathbf{P}_1) - H(\mathbf{P}_2) \leq 2\log(N)$$

Proof. Notice that $-H(\mathbf{P})$ is convex, so that $H(\mathbf{P})$ is concave.

$$\begin{aligned}\sum_{i,j} p_{i,j}\log\left(\frac{1}{p_{i,j}}\right) &= N^2 \sum_{i,j}\frac{1}{N^2} p_{i,j}\log\left(\frac{1}{p_{i,j}}\right) \\ &\leq N^2 H\left(\frac{1}{N^2}\sum_{i,j} p_{i,j}\right) = N^2 H(1/N^2) = N^2\frac{1}{N^2}\log\big(N^2\big) = 2\log(N).\end{aligned}$$

In addition since $p_{i,j} \leq 1$ for all i, j, we have that $H(\mathbf{P}) \geq 0$ and therefore we can bound

$$R_H = \max_{P_1, P_2 \in \mathcal{P}} H(\mathbf{P}_1) - H(\mathbf{P}_2) \leq \max_{\mathbf{P}_1 \in \mathcal{S}(\alpha_N, \beta_N)} H(\mathbf{P}_1) \leq 2\log(N).$$

□

Lemma C.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact with diameter $\mathfrak{d} > 0$. Suppose we are given two probability measures α, β on $\mathcal{X}$ with a unique deterministic, λ-Lipschitz optimal transport map $\mathbf{T}$ while $\alpha_N = \sum_{i=1}^N a_i \delta_{Y^i}$ with $a_i > 0$ and $\beta_N = \sum_{i=1}^N b_i \delta_{X^i}$. We write $\mathcal{P}^{\mathrm{OT},N}$, resp. $\mathcal{P}_\epsilon^{\mathrm{OT},N}$, for an optimal coupling between α_N and β_N, resp. the ϵ-regularized optimal transport plan, between α_N and β_N. Then*

$$\left[\int ||y - \mathbf{T}(x)||^2 \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}x, \mathrm{d}y)\right]^{\frac{1}{2}} \leq 2\lambda^{1/2}\mathcal{E}^{1/2}\left[\mathfrak{d}^{1/2} + \mathcal{E}\right]^{1/2} + \max\{\lambda, 1\}\left[\mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta)\right],$$

where

$$\mathcal{E} \coloneqq \mathcal{E}(N, \epsilon, \alpha, \beta) \coloneqq \mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) + \sqrt{2\epsilon\log(N)}.$$

Proof. From Corollary 3.8 from (Li & Nochetto, 2021)

$$\left[\int \|\mathbf{T}(x) - y\|^2 \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}x, \mathrm{d}y)\right]^{1/2} \leq 2\lambda^{1/2}\sqrt{\tilde{e}_{N,\epsilon}}\left[\mathcal{W}_2(\alpha, \beta) + \tilde{e}_{N,\epsilon}\right]^{1/2} + \lambda\mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta),$$

where λ is the Lipschitz constant of the optimal transport map $\mathbf{T}$ sending α to β, and

$$\tilde{e}_{N,\epsilon} \coloneqq \mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) + \left[\int \|x - y\|^2 \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}x, \mathrm{d}y)\right]^{1/2} - \mathcal{W}_2(\alpha_N, \beta_N). \tag{29}$$

From Proposition 4 of (Weed, 2018),

$$\sum_{i,j=1,\ldots,N} p_{\epsilon,i,j}^{\mathrm{OT}} |Y_i - X_j|^2 - \mathcal{W}_2^2(\alpha_N, \beta_N) \leq \epsilon R_H,$$

where R_H is the entropic radius as defined in Lemma C.1.

By Lemma C.1 we therefore have that

$$\int \|x - y\|^2 \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}x, \mathrm{d}y) - \mathcal{W}_2^2(\alpha_N, \beta_N) \leq 2\epsilon\log(N).$$

Since $x \mapsto \sqrt{x}$ is sub-additive, for $r, s > 0$ we have that $\sqrt{r} - \sqrt{s} \leq \sqrt{r - s}$, whence

$$\left[\int \|x - y\|^2 \mathcal{P}_\epsilon^{\mathrm{OT},N}(\mathrm{d}x, \mathrm{d}y)\right]^{1/2} - \mathcal{W}_2(\alpha_N, \beta_N) \leq \sqrt{2\epsilon\log N}.$$

We thus have

$$\tilde{e}_{N,\epsilon} \leq \mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) + \sqrt{2\epsilon\log(N)}.$$

In addition, by Assumption B.1 we have that $\mathcal{W}_2(\alpha, \beta) \leq \mathfrak{d}^{1/2}$ and the result follows. □

C.2. Proof of Proposition 4.2

Proof of Proposition 4.2. By definition, we have $\tilde{\beta}_N(\mathrm{d}\tilde{x}) = \int \alpha_N(\mathrm{d}x)\delta_{\mathbf{T}_{N,\epsilon}(x)}(\mathrm{d}\tilde{x})$ with $\mathbf{T}_{N,\epsilon}(x) \coloneqq \int \tilde{x}\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}\tilde{x}|x)$ while, as $\mathcal{P}_\epsilon^{\mathrm{OT},N}$ belongs to $\mathcal{U}(\alpha_N, \beta_N)$, we also have $\beta_N(\mathrm{d}\tilde{x}) = \int \alpha_N(\mathrm{d}x)\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}\tilde{x}|x)$. We then have for any 1-Lipschitz

function

$$\begin{aligned}\left|\beta_N(\psi) - \tilde{\beta}_N(\psi)\right| &= \left|\int \left[\int (\psi(\tilde{x}) - \psi(\mathbf{T}_{N,\epsilon}(x)))\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}\tilde{x}|x)\right] \alpha_N(\mathrm{d}x)\right| \\ &\leq \iint |\psi(\tilde{x}) - \psi(\mathbf{T}_{N,\epsilon}(x))| \, \alpha_N(\mathrm{d}x)\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}\tilde{x}|x) \\ &\leq \iint ||\tilde{x} - \mathbf{T}_{N,\epsilon}(x)||\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x, \mathrm{d}\tilde{x}) \\ &\leq \left(\iint ||\tilde{x} - \mathbf{T}_{N,\epsilon}(x)||^2\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x, \mathrm{d}\tilde{x})\right)^{\frac{1}{2}} \\ &\leq \left(\iint ||\tilde{x} - \mathbf{T}(x)||^2\mathcal{P}_\epsilon^{\mathrm{OT,N}}(\mathrm{d}x, \mathrm{d}\tilde{x})\right)^{\frac{1}{2}},\end{aligned}$$

where the final inequality follows from the fact that for any random vector V the mapping $v \mapsto \mathbb{E}[\|V - v\|^2]$ is minimized at $v = \mathbb{E}[V]$. The stated result is then obtained using Lemma C.2. □

D. Proof of Proposition 4.3

For technical reasons, we analyse here a slightly modified PF algorithm where

$$\alpha_N^{(t)} = \frac{1}{N}\sum_{j=1}^N \delta_{X_t^j}, \qquad X_t^j \overset{\text{i.i.d.}}{\sim} \tilde{\beta}_N^{(t-1)} f = \frac{1}{N}\sum_{j=1}^N f\left(\cdot\middle|\tilde{X}_{t-1}^j\right). \tag{30}$$

instead of the standard version where one has

$$\alpha_N^{(t)} = \frac{1}{N}\sum_{j=1}^N \delta_{X_t^j}, \qquad X_t^j \sim f\left(\cdot\middle|\tilde{X}_{t-1}^j\right).$$

This slightly modified version of the bootstrap PF was analyzed for example in (Del Moral & Guionnet, 2001). The analysis does capture the additional error arising from the use of DET instead of resampling. Similar results should hold for the standard PF algorithm. The main technical reason for analysing this modified algorithm is our reliance on Theorem 2 of (Fournier & Guillin, 2015); analysing the standard PF algorithm requires a version of (Fournier & Guillin, 2015) for stratified sampling and will be done in future work.

Proposition D.1. *Suppose that Assumptions B.1, B.2 and B.3 hold. Suppose also that given $\tilde{\beta}_N^{(t-1)}$, $\alpha_N^{(t)}$ is defined through (30). Define the functions*

$$\begin{aligned}\mathcal{F}(x) &\coloneqq x + \sqrt{\mathfrak{d}K_1(\Delta, \mathfrak{d})x} \\ f_d(x) &\coloneqq \begin{cases} x, & d < 4 \\ \frac{x}{\log(2+1/x)}, & d = 4 \\ x^{d/2}, & d > 4. \end{cases} \\ \mathcal{F}_{N,\epsilon,\delta,d}(x) &\coloneqq \mathcal{F}\left(\kappa x + \sqrt{f_d^{-1}\left(\frac{\log(C/\delta)}{cN}\right)}\right), \\ \frac{1}{\mathfrak{d}}\mathfrak{G}_{\epsilon,\delta,N,d}^2(x) &\coloneqq 2\lambda^{1/2}\left[\mathcal{F}_{N,\epsilon,\delta,d}(x) + \sqrt{2\epsilon\log N}\right]^{1/2}\left[\mathfrak{d}^{1/2} + \mathcal{F}_{N,\epsilon,\delta,d}(x) + \sqrt{2\epsilon\log N}\right]^{1/2} \\ &\quad + \lambda\kappa\mathcal{F}_{N,\epsilon,\delta,d}(x) + \max\{\lambda, 1\}\mathcal{F}_{N,\epsilon,\delta,d}(x).\end{aligned} \tag{31}$$

Then for any $\epsilon, \delta > 0$ we have with probability at least $1 - \delta$, over the sampling step in (30), that

$$\mathcal{W}_2\left(\tilde{\beta}_N^{(t)}, \beta^{(t)}\right) \leq \mathfrak{G}_{\epsilon,\delta,N,d}\left[\mathcal{W}_2\left(\tilde{\beta}_N^{(t-1)}, \beta^{(t-1)}\right)\right] \tag{32}$$

In particular if $\mathcal{W}_2(\tilde{\beta}_N^{(t-1)}, \beta^{(t-1)}) \to 0$ and $\epsilon_N = o(1/\log(N))$ as $N \to \infty$ we have that

$$\mathcal{W}_2(\tilde{\beta}_N^{(t)}, \beta^{(t)}) \to 0,$$

in probability.

Proof of Proposition D.1. To keep notation concise we write for $N \geq 1$

$$\alpha_N \coloneqq \alpha_N^{(t)}, \quad \alpha'_N \coloneqq \tilde{\beta}_N^{(t-1)} f, \quad \beta_N \coloneqq \beta_N^{(t)}, \quad \tilde{\beta}_N \coloneqq \tilde{\beta}_N^{(t-1)}.$$

Controlling $\mathcal{W}_1(\beta_N, \beta)$. Let ψ be 1-Lipschitz. Without loss of generality we may assume that $\psi(0) = 0$ since otherwise we can remove a constant.

$$\begin{aligned}
|\beta_N(\psi) - \beta(\psi)| &= \left| \frac{\alpha_N(\omega\psi)}{\alpha_N(\omega)} - \frac{\alpha(\omega\psi)}{\alpha(\omega)} \right| \\
&\leq \left| \frac{\alpha_N(\omega\psi)}{\alpha_N(\omega)} - \frac{\alpha(\omega\psi)}{\alpha_N(\omega)} \right| + \left| \frac{\alpha(\omega\psi)}{\alpha_N(\omega)} - \frac{\alpha(\omega\psi)}{\alpha(\omega)} \right| \\
&\leq \Delta^{-1} \left| \alpha_N(\omega\psi) - \alpha(\omega\psi) \right| + \Delta^{-2} \alpha(\omega\psi) |\alpha_N(\omega) - \alpha(\omega)|.
\end{aligned}$$

At this stage notice that

$$|(\omega\psi)'| \leq |\omega'\psi| + |\omega\psi'| \leq \|\psi\|_\infty + \|\omega\|_\infty.$$

Notice that

$$|\psi(x)| = |\psi(x) - \psi(0)| \leq |x - 0| \leq \mathfrak{d}.$$

Therefore we have that

$$|(\omega\psi)'| \leq \mathfrak{d} + \Delta^{-1},$$

and thus $\omega\psi$ is $(\mathfrak{d} + \Delta^{-1})$-Lipschitz. It follows that

$$\begin{aligned}
|\beta_N(\psi) - \beta(\psi)| &\leq \Delta^{-1} \left| \alpha_N(\omega\psi) - \alpha(\omega\psi) \right| + \Delta^{-2} \alpha(\omega\psi) |\alpha_N(\omega) - \alpha(\omega)| \\
&\leq \Delta^{-1} (\mathfrak{d} + \Delta^{-1}) \mathcal{W}_1(\alpha_N, \alpha) + \Delta^{-3} \mathfrak{d} \mathcal{W}_1(\alpha_N, \alpha) \\
&=: K_1(\Delta, \mathfrak{d}) \mathcal{W}_1(\alpha_N, \alpha).
\end{aligned}$$

Therefore we have that

$$\mathcal{W}_1(\beta_N, \beta) \leq K_1(\Delta, \mathfrak{d}) \mathcal{W}_1(\alpha_N, \alpha). \tag{33}$$

Notice that using the compactness of the state space we easily get also that

$$\mathcal{W}_2(\beta_N, \beta) \leq \sqrt{\mathfrak{d} \mathcal{W}_1(\beta_N, \beta)} \leq \sqrt{\mathfrak{d} K_1(\Delta, \mathfrak{d}) \mathcal{W}_1(\alpha_N, \alpha)} \leq \sqrt{\mathfrak{d} K_1(\Delta, \mathfrak{d}) \mathcal{W}_2(\alpha_N, \alpha)}, \tag{34}$$

since clearly $\mathcal{W}_1(\rho, \sigma) \leq \mathcal{W}_2(\rho, \sigma)$ for any two probability measures ρ, σ.

Controlling $\mathcal{W}_1(\tilde{\beta}_{N,\epsilon}, \beta)$. Again supposing ψ is 1-Lipschitz, and $\psi(0) = 0$, consider

$$\begin{aligned}
\left| \tilde{\beta}_N(\psi) - \tilde{\beta}(\psi) \right| &= \left| \int \psi(\mathbf{T}_{N,\epsilon}(x)) \alpha_N(\mathrm{d}x) - \int \psi(\mathbf{T}(x)) \alpha(\mathrm{d}x) \right| \\
&\leq \left| \int \psi(\mathbf{T}_{N,\epsilon}(x)) \alpha_N(\mathrm{d}x) - \int \psi(\mathbf{T}(x)) \alpha_N(\mathrm{d}x) \right| \\
&\quad + \left| \int \psi(\mathbf{T}(x)) \alpha_N(\mathrm{d}x) - \int \psi(\mathbf{T}(x)) \alpha(\mathrm{d}x) \right|
\end{aligned}$$

For the second term, using the fact that $\mathbf{T}$ and ψ are λ- and 1-Lipschitz respectively, we have that $\psi \circ \mathbf{T}$ is λ-Lipschitz and therefore

$$\left| \int \psi(\mathbf{T}(x)) \alpha_N(\mathrm{d}x) - \int \psi(\mathbf{T}(x)) \alpha(\mathrm{d}x) \right| \leq \lambda \mathcal{W}_1(\alpha_N, \alpha) \leq \lambda \mathcal{W}_2(\alpha_N, \alpha),$$

where we used Assumption B.2 for that last inequality For the first term recall that using Cauchy-Schwarz and Jensen we get

$$\begin{aligned}
&\left|\int \psi(\mathbf{T}_{N,\epsilon}(x))\alpha_N(\mathrm{d}x) - \int \psi(\mathbf{T}(x))\alpha_N(\mathrm{d}x)\right| \\
&\leq \int |\mathbf{T}_{N,\epsilon}(x) - \mathbf{T}(x)|\, \alpha_N(\mathrm{d}x) \\
&\leq \int \left|\int y\mathcal{P}_{N,\epsilon}(x,\mathrm{d}y) - \mathbf{T}(x)\right| \alpha_N(\mathrm{d}x) \\
&\leq \iint |y - \mathbf{T}(x)|\, \alpha_N(\mathrm{d}x)\mathcal{P}_{N,\epsilon}(x,\mathrm{d}y) \\
&\leq \left[\iint |y - \mathbf{T}(x)|^2\, \alpha_N(\mathrm{d}x)\mathcal{P}_{N,\epsilon}(x,\mathrm{d}y)\right]^{1/2}.
\end{aligned}$$

Here we can directly apply Lemma C.2 to obtain

$$\begin{aligned}
&\left[\iint |y - \mathbf{T}(x)|^2\, \alpha_N(\mathrm{d}x)\mathcal{P}_{N,\epsilon}(x,\mathrm{d}y)\right]^{1/2} \\
&\leq 2\lambda^{1/2}\mathcal{E}^{1/2}\left[\mathfrak{d}^{1/2} + \mathcal{E}\right]^{1/2} + \max\{\lambda, 1\}\left[\mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta)\right],
\end{aligned}$$

where

$$\mathcal{E} \coloneqq \mathcal{E}(n, \epsilon, \alpha, \beta) \coloneqq \mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) + \sqrt{2\epsilon\log(N)}.$$

From (34) we have that

$$\mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) \leq \mathcal{W}_2(\alpha_N, \alpha) + \sqrt{\mathfrak{d}K_1(\Delta, \mathfrak{d})\mathcal{W}_2(\alpha_N, \alpha)}.$$

Next we want to bound $\mathcal{W}_2(\alpha_N, \alpha)$. Notice first that

$$\mathcal{W}_2(\alpha_N, \alpha) \leq \mathcal{W}_2(\alpha_N, \alpha'_N) + \mathcal{W}_2(\alpha'_N, \alpha) \leq \mathcal{W}_2(\alpha_N, \alpha'_N) + \kappa\mathcal{W}_2\left(\tilde{\beta}_N^{(t-1)}, \beta^{(t-1)}\right),$$

by Assumption B.2.

To control the other term we use (Fournier & Guillin, 2015) to obtain a high probability bound on $\mathcal{W}_2(\alpha_N, \alpha'_N)$. In particular, using Theorem 2 from (Fournier & Guillin, 2015), with $\alpha = \infty$ since we are in a compact domain, that for some positive constants C, c we have

$$\mathbb{P}\left[\mathcal{W}_2^2(\alpha_N, \alpha'_N) \geq x\right] \leq C\exp\left[-cNf_d^2(x)\right], \tag{35}$$

where

$$f_d(x) \coloneqq \begin{cases} x, & d < 4 \\ \frac{x}{\log(2+1/x)}, & d = 4 \\ x^{d/2}, & d > 4. \end{cases} \tag{36}$$

In particular, for any $\delta > 0$, with probability at least $1 - \delta$ over the sampling step in F_N we have that

$$\mathcal{W}_2(\alpha_N, \alpha'_N) \leq \sqrt{f_d^{-1}\left(\frac{\log(C/\delta)}{cN}\right)}. \tag{37}$$

Assuming that $d \geq 4$ the rate then is of order $N^{-1/d}$ as expected.

Therefore with probability at least $1 - \delta$ over the sampling step we have that

$$\mathcal{W}_2(\alpha_N, \alpha) + \mathcal{W}_2(\beta_N, \beta) \leq \mathcal{F}_{N,\epsilon,\delta,d}\left(\mathcal{W}_2\left(\tilde{\beta}_N^{(t-1)}, \beta^{(t-1)}\right)\right),$$

where

$$\mathcal{F}_{N,\epsilon,\delta,d}(x) = \mathcal{F}\left(\kappa x + \sqrt{f_d^{-1}\left(\frac{\log(C/\delta)}{cN}\right)}\right), \qquad \mathcal{F}(x) \coloneqq x + \sqrt{\mathfrak{d}K_1(\Delta, \mathfrak{d})x} \tag{38}$$

Thus overall we have with probability at least $1-\delta$ over the sample

$$\mathcal{W}_2(\tilde{\beta}_{N,\epsilon},\tilde{\beta}) \leq \sqrt{\mathfrak{d}\mathcal{W}_1(\tilde{\beta}_{N,\epsilon},\tilde{\beta})} \leq \mathfrak{G}_{\epsilon,\delta,N,d}\left(\mathcal{W}_2\left(\tilde{\beta}_N^{(t-1)},\beta^{(t-1)}\right)\right),$$

where

$$\begin{aligned}\frac{1}{\mathfrak{d}}\mathfrak{G}^2_{\epsilon,\delta,N,d}(x) \coloneqq 2\lambda^{1/2}\left[\mathcal{F}_{N,\epsilon,\delta,d}(x)+\sqrt{2\epsilon\log N}\right]^{1/2}\left[\mathfrak{d}^{1/2}+\mathcal{F}_{N,\epsilon,\delta,d}(x)+\sqrt{2\epsilon\log N}\right]^{1/2}\\ +\lambda\kappa\mathcal{F}_{N,\epsilon,\delta,d}(x)+\max\{\lambda,1\}\mathcal{F}_{N,\epsilon,\delta,d}(x).\end{aligned}$$

In particular notice that if we set $\epsilon_N = o(1/\log N)$ and $x_N = o(1)$ we have

$$\mathfrak{G}_{\epsilon_N,\delta,N,d}(x_N) \to 0.$$

Therefore, notice that if $\epsilon_N = o(1/\log N)$ and $\mathcal{W}_2(\mu_N,\mu)\to 0$, then for any $x>0$ we have that

$$\mathbb{P}\left[\mathcal{W}_2\big(\tilde{\beta}_{N,\epsilon},\tilde{\beta}\big)\geq x\right] \leq \mathbb{P}[\mathcal{W}_2(\alpha'_N,\alpha_N)\geq x'],$$

for some x' that does not depend on N, where the probability is over the sampling step. The convergence in probability follows. □

Proposition D.2. *Let $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_1^i}$ where $X_1^i \overset{\text{i.i.d.}}{\sim} \mu \coloneqq q(\cdot|y_1)$ for $i\in[N]$ and suppose that for $t\geq 1$, $\alpha_N^{(t)}$ is defined through (30). Under Assumptions B.1, B.2, B.3 and B.4, for any $\delta>0$, with probability at least $1-2\delta$ over the sampling steps, for any bounded 1-Lipschitz ψ, for any $t\in[1:T]$, the approximations of the filtering distributions and log-likelihood computed by DPF satisfy*

$$|\tilde{\beta}_N^{(t)}(\psi)-\beta^{(t)}(\psi)| \leq \mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}\left(\sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}\right) \tag{39}$$

$$\left|\log\frac{\hat{p}_N(y_{1:T})}{p(y_{1:T})}\right| \leq \frac{\kappa}{\Delta}\max_{t\in[1:T]}\mathrm{Lip}\left[g(y_t\mid\cdot)\right]\sum_{t=1}^T \mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}\left(\sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}\right) \tag{40}$$

where C is a finite constant independent of T, $\mathfrak{G}_{\epsilon,\delta/T,N,d}$, f_d are defined in (31), and $\mathrm{Lip}[f]$ is the Lipschitz constant of the function f. $\mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}$ denotes the t-repeated composition of function $\mathfrak{G}_{\epsilon,\delta/T,N,d}$. In particular, if we set $\epsilon_N = o(1/\log N)$

$$\left|\log\frac{\hat{p}_N(y_{1:T})}{p(y_{1:T})}\right| \to 0,$$

in probability.

Proof of Proposition D.2. Following the proof of Proposition D.1, we define ${\alpha_N^{(t)}}' = \tilde{\beta}_N^{(t-1)}f$ and for $t\in[1:T]$, the events

$$A_t \coloneqq \mathcal{W}_2\left(\alpha_N^{(t)},{\alpha_N^{(t)}}'\right) \leq \sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}.$$

We know from Theorem 2 in (Fournier & Guillin, 2015) that $\mathbb{P}(A_t)\geq 1-\delta/T$, where the probability is over the sampling step. In particular we have that

$$\mathbb{P}\left[\bigcap_{t=1}^T A_t\right] = 1-\mathbb{P}\left[\bigcup_{t=1}^T A_t^{\mathsf{c}}\right] \geq 1-\sum_{t=1}^N \mathbb{P}\left[A_t^{\mathsf{c}}\right] \geq 1-T\frac{\delta}{T} = 1-\delta.$$

Notice that on the event $\cap_{t=1}^T A_t$, iterating the bound (32) we have

$$\mathcal{W}_2\left(\tilde{\beta}_N^{(t)},\beta^{(t)}\right) \leq \mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}\left(\mathcal{W}_2\left(\mu_N,\mu\right)\right),$$

with probability at least $1-\delta$. Again by Theorem 2 in (Fournier & Guillin, 2015) we have that with probability at least $1-\delta$

$$\mathcal{W}_2(\mu_N, \mu) \leq \sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}.$$

Therefore with probability at least $1-2\delta$ we have

$$\mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}\left(\sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}\right).$$

It remains to prove (40). Note that $|\log(x) - \log(y)| \leq \frac{|x-y|}{\min\{x,y\}}$ for any $x, y > 0$ so

$$\begin{aligned}
\left|\log \hat{p}(y_{1:T}) - \log p(y_{1:T})\right| &\leq \sum_{t=1}^{T} \left|\log \hat{p}(y_t|y_{1:t-1}) - \log p(y_t|y_{1:t-1})\right| \\
&\leq \sum_{t=1}^{T} \left|\frac{\hat{p}(y_t|y_{1:t-1}) - p(y_t|y_{1:t-1})}{\min(\hat{p}(y_t|y_{1:t-1}), p(y_t|y_{1:t-1}))}\right| \\
&\leq \Delta^{-1} \sum_{t=1}^{T} \left|\hat{p}(y_t|y_{1:t-1}) - p(y_t|y_{1:t-1})\right| \qquad (41)
\end{aligned}$$

where Δ is defined in Assumption B.3.

The term in line (41) may be written as follows

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

for $\Delta^2 \leq h(x) := \int g(y_t|x') f(x'|x) \mathrm{d}x' \leq \Delta^{-2}$. At this point notice also that

$$\begin{aligned}
h(x) - h(x') &= \int f(\mathrm{d}w|x) g(y_t \mid w) - \int f(\mathrm{d}w|x') g(y_t \mid w) \\
&= \int \delta_x(\mathrm{d}z) \int f(\mathrm{d}w|z) g(y_t \mid w) - \int \delta_{x'}(\mathrm{d}z) \int f(\mathrm{d}w|z) g(y_t \mid w) \\
&= [\delta_x f][g(y_t \mid \cdot)] - [\delta_{x'} f][g(y_t \mid \cdot)] \\
&\leq \mathrm{Lip}\left[g(y_t \mid \cdot)\right] \mathcal{W}_1\left(\delta_x f, \delta_{x'} f\right) \leq \kappa \mathrm{Lip}\left[g(y_t \mid \cdot)\right] \mathcal{W}_1\left(\delta_x, \delta_{x'}\right) = \kappa \mathrm{Lip}\left[g(y_t \mid \cdot)\right] |x - x'|,
\end{aligned}$$

by Assumption B.2. It follows therefore that h is Lipschitz and therefore that

$$\hat{p}(y_t|y_{1:t-1}) - p(y_t|y_{1:t-1}) = \tilde{\beta}_N^{(t-1)}(h) - \beta^{(t-1)}(h) \leq \kappa \mathrm{Lip}\left[g(y_t \mid \cdot)\right] \mathcal{W}_1\left(\beta_N^{(t-1)}, \beta^{(t-1)}\right).$$

Combining (39) and (41), and using the fact that $\mathcal{W}_1 \leq \mathcal{W}_2$, we thus get

$$\begin{aligned}
\left|\log \hat{p}(y_{1:T}) - \log p(y_{1:T})\right| &\leq \Delta^{-1} \kappa \sum_{t=1}^{T} \mathrm{Lip}\left[g(y_t \mid \cdot)\right] \mathcal{W}_1\left(\tilde{\beta}_N^{(t-1)}, \beta^{(t)}\right) \\
&\leq \Delta^{-1} \kappa \max_{t \in [1:T]} \mathrm{Lip}\left[g(y_t \mid \cdot)\right] \sum_{t=1}^{T} \mathfrak{G}^{(t)}_{\epsilon,\delta/T,N,d}\left(\sqrt{f_d^{-1}\left(\frac{\log(CT/\delta)}{cN}\right)}\right),
\end{aligned}$$

where the last inequality holds with probability at least $1-\delta$ over the sampling steps.

The convergence in probability follows from the corresponding statement of Proposition D.1. □

E. Additional Experiments and Details

E.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:

$$X_t|\{X_{t-1}=x\} \sim \mathcal{N}\left(\text{diag}(\theta_1\ \theta_2)x, 0.5\mathbf{I}_2\right), \quad Y_t|\{X_t=x\} \sim \mathcal{N}(x, 0.1\cdot\mathbf{I}_2). \tag{42}$$

We simulate $T=150$ observations using $\theta=(\theta_1,\theta_2)=(0.5,0.5)$. As a result, we expect in these scenarios that the filtering distribution $p_\theta(x_t|y_{1:t})$ is not too distinct from the smoothing distribution $p_\theta(x_t|y_{1:T})$ as the latent process is mixing quickly. From Proposition 4.1, this is thus a favourable scenario for methods ignoring resampling terms in the gradient as the bias should not be very large. Figure 1, displayed earlier, shows $\ell(\theta)$ obtained by Kalman and $\hat{\ell}(\theta;\mathbf{u})$ computed regular PF and DPF for the same number $N=25$ of 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 1, where the gradient is computed using the biased gradient from (Maddison et al., 2017; Naesseth et al., 2018; Le et al., 2018) for regular PF.

We now compare the performance of the estimators $\hat{\theta}_{\text{SMLE}}$ (for DPF) and $\hat{\theta}_{\text{ELBO}}$ (for both regular PF and DPF) learned using gradient with learning rate 10^{-4} on 100 steps, using $N=25$ for DPF and $N=500$ for regular PF, to $\hat{\theta}_{\text{MLE}}$ 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 $\ell_{\text{ELBO}}(\theta)$ ignoring the contributions of resampling steps as in (Maddison et al., 2017; Naesseth et al., 2018; Le et al., 2018) (we recall that unbiased estimates suffer from very high variance) and unbiased gradients of $\ell^{\text{ELBO}}(\theta)$ using DPF. 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 $\hat{\ell}_{\text{SMLE}(\theta;\mathbf{u}_{1:B})} := \frac{1}{B}\sum_{b=1}^{B}\hat{\ell}(\theta;\mathbf{u}_b)$. The results are given in Table 4. For this example, $\hat{\theta}^{\text{DPF}}_{\text{ELBO}}$ maximizing $\ell^{\text{ELBO}}_{\text{DPF}}(\theta)$ outperforms $\hat{\theta}^{\text{PF}}_{\text{ELBO}}$ and $\hat{\theta}_{\text{SMLE}}$. However, as B increases, $\hat{\theta}_{\text{SMLE}}$ gets closer to $\hat{\theta}^{\text{DPF}}_{\text{ELBO}}$ which is to be expected as $\hat{\ell}_{\text{SMLE}}(\theta;\mathbf{u}_{1:B}) \longrightarrow \ell^{\text{ELBO}}(\theta)$. In Table 4, the Root Mean Square Error (RMSE) is defined as $\sqrt{\frac{1}{M}\sum_{i=1}^{2}\sum_{k=1}^{M}(\hat{\theta}_i^k-\hat{\theta}^k_{\text{MLE},i})^2}$.

Table 4. $10^3\times$ RMSE[4] over 50 datasets - lower is better

B	$\hat{\theta}^{\text{PF}}_{\text{ELBO}}$	$\hat{\theta}^{\text{DPF}}_{\text{ELBO}}$	$\hat{\theta}_{\text{SMLE}}$
1	1.94	1.30	7.94
4	2.40	1.35	3.28
10	2.80	1.37	2.18

E.2. Variational Recurrent Neural Network

$N=32$ particles were used for training, with a regularization parameter of $\epsilon=0.5$. The ELBO (scaled by sequence length) was used as the training objective to maximise for each resampling/ DET procedure. The ELBO evaluated on test data using $N=500$ particles and multinomial resampling. Resampling / DET operations were carried out when effective sample (ESS) size fell below $N/2$. Learning rate 0.001 was used with the Adam optimizer.

Recall the state-space model is given by

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

Network architectures and data preprocessing steps were based loosely on (Maddison et al., 2017). Given the low volume of data and sparsity of the observations, relatively small neural networks were considered to prevent overfitting, larger neural

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

networks are considered in the more complex robotics experiments. R_t is of dimension $d_r = 16$, Z_t is of dimension $d_z = 8$. E_θ is a single layer fully connected network with hidden layer of width 16, output of dimension 16 and RELU activation.

μ_θ and σ_θ are both fully connected neural networks with two hidden layers, each of 16 units and RELU activation, the activation function is not applied to the final output of μ_θ but the softplus is applied to the output of σ_θ, which is the diagonal entries of the covariance matrix of the normal distribution that is used to sample Z_t.

h_θ is a single layer fully connected network with two hidden layers, each of width 16 and RELU activation. The final output is not put through the RELU and is instead used as the logits for the Bernoulli distribution of observations.

E.3. Robot Localization

Similar to the VRNN example, $N = 32$ particles were used for training, with a regularization parameter of $\epsilon = 0.5$ and resampling / DET operations were carried out when ESS size fell below $N/2$. Learning rate 0.001 was used with the Adam optimizer.

Network architectures and data preprocessing were based loosely on (Jonschkowski et al., 2018). There are 3 neural networks being considered:

- Encoder E_θ maps RBG 24×24 pixel images, hence dimension $3 \times 24 \times 24$, to encoding of size $d_E = 128$. This network consists of a convolutional network (CNN) of kernel size 3 and a single layer fully connected network of hidden width 128 and RELU activation.
- Decoder D_θ maps encoding back to original image. This consists of a fully connected neural network with three hidden layers of width 128 and RELU activation function. This is followed by a transposed convolution network with matching specification to the CNN in the encoder, to return an output with the same dimension as observation images, $3 \times 24 \times 24$.
- Network G_θ maps the state $S_t = (X_t^{(1)}, X_t^{(2)}, \gamma_t)$ to encoding of dimension 128. First angle γ_t was converted to $\sin(\gamma_t), \cos(\gamma_t)$. Then the augmented state $(X_t^{(1)}, X_t^{(2)}, \sin(\gamma_t), \cos(\gamma_t))$ was passed to a 3 layer fully connected network with hidden layers of dimensions $16, 32, 64$ and RELU activation function, with final output of dimension 128.