Optimized Auxiliary Particle Filters:
adapting mixture proposals via convex optimization.
Supplementary Material

Nicola Branchini[1] **Víctor Elvira**[1]

[1]University of Edinburgh

A SUPPLEMENTARY MATERIAL

A.1 THEORETICAL PROPERTIES OF THE OAPF ESTIMATORS

The theoretical properties of the estimators OAPF are analized from the importance sampling perspective. In the case of the mixture proposals ψ_t, we assume that each time t, the support of ψ_t is a superset of the support of $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, i.e., that $\psi_t(\mathbf{x}_t) > 0$ for all $\mathbf{x}_t$ where $p(\mathbf{x}_t|\mathbf{y}_{1:t}) > 0$. Let us define the *partial* normalizing constants as $Z_t \triangleq p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$, the *joint* normalizing constant as $Z_{1:t} \triangleq p(\mathbf{y}_{1:t})$, and also $Z_{t-h:t} \triangleq p(\mathbf{y}_{t-h:t}|\mathbf{y}_{1:t-h-1})$. In the OAPF framework, we can build estimator of those quantities, e.g., the partial estimator $\widehat{Z}_\tau \triangleq \frac{1}{M}\sum_{m=1}^M \widetilde{w}_\tau^{(m)}$, the joint estimator $\widehat{Z}_{1:t} = \prod_{\tau=1}^t \widehat{Z}_\tau$, and also the estimator $\widehat{Z}_{t-h:t} = \prod_{\tau=t-h}^t \widehat{Z}_\tau$, with and the estimator $\widehat{Z}_t \triangleq \frac{1}{M}\sum_{m=1}^M \widetilde{w}_t^{(m)}$. We also assume that the estimators of all the partial normalizing constants have finite variance (see for instance [Owen, 2013, Elvira et al., 2019]). We define the set of weighted samples at time t as $\mathcal{A}_t \triangleq \{\mathbf{x}_t^{(m}, \widetilde{w}_t^{(m}\}_{m=1}^M$. In order to avoid ambiguities when evaluating pdfs, we define the functions $g(\mathbf{y}_t|\mathbf{x}_t) \triangleq p(\mathbf{y}_t|\mathbf{x}_t)$, $g(\mathbf{y}_t|\mathbf{x}_{t-1}) \triangleq p(\mathbf{y}_t|\mathbf{x}_{t-1})$, $g(\mathbf{y}_t, \mathbf{x}_t|\mathbf{x}_{t-1}) \triangleq p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{x}_{t-1})$ and $g(\mathbf{y}_{t-h:t}, \mathbf{x}_t|\mathbf{x}_{t-1}) \triangleq p(\mathbf{y}_{t-h:t}, \mathbf{x}_t|\mathbf{x}_{t-1})$.

In the following, we show that OAPF provides an unbiased estimator of the normalizing constant $p(\mathbf{y}_{1:t})$, which follows a proof by induction, in a similar spirit as in [Pitt et al., 2012], but with more generic results. In particular, here the (approximate) filtering distribution is the marginalized version of the one in [Pitt et al., 2012] and is constituted by a mixture in the numerator of the importance weights (see [Klaas et al., 2005] for an explanation). In OAPF the proposal density can be any mixture $\psi_t(\mathbf{x}_t)$ fulfilling the standard regularity conditions described above, hence in the denominator of the importance weights, a second mixture appears. Theorem 1 is here the main result, and is supported by Lemmas 1 and 2 which we present first.

Lemma 1 *We have that*

$$\mathbb{E}\Big[\widehat{Z}_t|\mathcal{A}_{t-1}\Big] = \sum_{m=1}^M w_{t-1}^{(m)} g(\mathbf{y}_t|\mathbf{x}_{t-1}^{(m)}). \tag{1}$$

Proof:

$$\mathbb{E}\Big[\widehat{Z}_t|\mathcal{A}_{t-1}\Big] = \mathbb{E}\Big[\frac{1}{M}\sum_{m=1}^M \widetilde{w}_t^{(m)}|\mathcal{A}_{t-1}\Big] \tag{2}$$

$$= \mathbb{E}\Big[\frac{1}{M}\sum_{m=1}^M \frac{g(\mathbf{y}_t|\mathbf{x}_t^{(m)})\sum_{j=1}^M w_{t-1}^{(j)} f(\mathbf{x}_t^{(m)}|\mathbf{x}_{t-1}^{(j)})}{\psi(\mathbf{x}_t^{(m)})}|\mathcal{A}_{t-1}\Big] \tag{3}$$

$$= \frac{1}{M}\sum_{m=1}^M \mathbb{E}\Big[\frac{g(\mathbf{y}_t|\mathbf{x}_t^{(m)})\sum_{j=1}^M w_{t-1}^{(j)} f(\mathbf{x}_t^{(m)}|\mathbf{x}_{t-1}^{(j)})}{\psi(\mathbf{x}_t^{(m)})}|\mathcal{A}_{t-1}\Big]. \tag{4}$$

Now, since given $\mathcal{A}_{t-1}$ the particles at time t are conditionally independent with pdf $\psi_t(\mathbf{x}_t)$, then we have that the integrals within (4) are identical:

$$\mathbb{E}\Big[\widehat{Z}_t|\mathcal{A}_{t-1}\Big] = \int \frac{g(\mathbf{y}_t|\mathbf{x}_t)\sum_{j=1}^M w_{t-1}^{(j)} f(\mathbf{x}_t|\mathbf{x}_{t-1}^{(j)})}{\psi(\mathbf{x}_t)}\psi(\mathbf{x}_t)d\mathbf{x}_t \tag{5}$$

$$= \int g(\mathbf{y}_t|\mathbf{x}_t)\sum_{j=1}^M w_{t-1}^{(j)} f(\mathbf{x}_t|\mathbf{x}_{t-1}^{(j)})d\mathbf{x}_t \tag{6}$$

$$= \sum_{j=1}^M w_{t-1}^{(j)} \int g(\mathbf{y}_t,\mathbf{x}_t|\mathbf{x}_{t-1}^{(j)})d\mathbf{x}_t \tag{7}$$

$$= \sum_{j=1}^M w_{t-1}^{(j)} g(\mathbf{y}_t|\mathbf{x}_{t-1}^{(j)}). \tag{8}$$

□

Lemma 2 *For any $h \in \{1,...,t-1\}$ we have that*

$$\mathbb{E}\Big[\widehat{Z}_{t-h:t}|\mathcal{A}_{t-h-1}\Big] = \sum_{m=1}^M w_{t-h-1}^{(m)} g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1}^{(m)}). \tag{9}$$

Proof: We follow a proof by induction. First, note that (9) is true for $h=0$ due to Lemma 1. Then, we assume that (9) holds for a given h and we will prove that it then holds for $h+1$. Let us start developing the left-hand side of (9) for $h+1$ by first noting that $\widehat{Z}_{t-h-1:t} = \widehat{Z}_{t-h:t}\widehat{Z}_{t-h-1}$. Then,

$$\mathbb{E}\Big[\widehat{Z}_{t-h-1:t}|\mathcal{A}_{t-h-2}\Big] = \mathbb{E}\Bigg[\mathbb{E}\Big[\widehat{Z}_{t-h:t}|\mathcal{A}_{t-h-1}\Big]\widehat{Z}_{t-h-1}|\mathcal{A}_{t-h-2}\Bigg] \tag{10}$$

$$= \mathbb{E}\Bigg[\Big[\sum_{m=1}^M g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1}^{(m)})w_{t-h-1}^{(m)}\Big]\widehat{Z}_{t-h-1}|\mathcal{A}_{t-h-2}\Bigg] \tag{11}$$

$$\tag{12}$$

where we have simply substituted Eq. (9) that we assume to hold for h. Next,

$$\mathbb{E}\Big[\widehat{Z}_{t-h-1:t}|\mathcal{A}_{t-h-2}\Big] = \mathbb{E}\Bigg[\Big[\sum_{m=1}^M g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1}^{(m)})\frac{\widetilde{w}_{t-h-1}^{(m)}}{\sum_{j=1}^M \widetilde{w}_{t-h-1}^{(j)}}\Big]\frac{1}{M}\sum_{j=1}^M \widetilde{w}_{t-h-1}^{(j)}|\mathcal{A}_{t-h-2}\Bigg] \tag{13}$$

$$= \mathbb{E}\Bigg[\frac{1}{M}\sum_{m=1}^M g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1}^{(m)})\frac{g(\mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1}^{(m)})\sum_{j=1}^M w_{t-h-2}^{(j)} f(\mathbf{x}_{t-h-1}^{(m)}|\mathbf{x}_{t-h-2}^{(j)})}{\psi_{t-h-1}(\mathbf{x}_{t-h-1}^{(m)})}|\mathcal{A}_{t-h-2}\Bigg] \tag{14}$$

$$= \frac{1}{M}\sum_{m=1}^M \mathbb{E}\Bigg[g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1}^{(m)})\frac{g(\mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1}^{(m)})\sum_{j=1}^M w_{t-h-2}^{(j)} f(\mathbf{x}_{t-h-1}^{(m)}|\mathbf{x}_{t-h-2}^{(j)})}{\psi(\mathbf{x}_{t-h-1}^{(m)})}|\mathcal{A}_{t-h-2}\Bigg] \tag{15}$$

where we have substituted with the importance weights $\widetilde{w}_{t-h-1}^{(m)}$ of Eq. 7 of the manuscript. Since, given $\mathcal{A}_{t-h-2}$, the particles at time t are conditionally independent with pdf $\psi_{t-h-1}(\mathbf{x}_{t-h-1})$, all M expectations are identical:

$$\mathbb{E}\Big[\widehat{Z}_{t-h-1:t}|\mathcal{A}_{t-h-2}\Big] = \int g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1})\frac{g(\mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1})\sum_{j=1}^{M} w_{t-h-2}^{(j)} f(\mathbf{x}_{t-h-1}|\mathbf{x}_{t-h-2}^{(j)})}{\psi(\mathbf{x}_{t-h-1})}\psi(\mathbf{x}_{t-h-1})d\mathbf{x}_{t-h-1} \tag{16}$$

$$= \int g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1})g(\mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1})\sum_{j=1}^{M} w_{t-h-2}^{(j)} f(\mathbf{x}_{t-h-1}|\mathbf{x}_{t-h-2}^{(j)})d\mathbf{x}_{t-h-1} \tag{17}$$

$$= \int g(\mathbf{y}_{t-h-1:t}|\mathbf{x}_{t-h-1})\sum_{j=1}^{M} w_{t-h-2}^{(j)} f(\mathbf{x}_{t-h-1}|\mathbf{x}_{t-h-2}^{(j)})d\mathbf{x}_{t-h-1} \tag{18}$$

Step (17) to (18) is justified since $\mathbf{y}_{t-h:t} \perp\!\!\!\perp \mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1}$, so we can replace $g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1})$ in 17 with $g(\mathbf{y}_{t-h:t}|\mathbf{y}_{t-h-1}, \mathbf{x}_{t-h-1})$ and then $g(\mathbf{y}_{t-h-1:t}|\mathbf{x}_{t-h-1}) = g(\mathbf{y}_{t-h:t}|\mathbf{x}_{t-h-1})g(\mathbf{y}_{t-h-1}|\mathbf{x}_{t-h-1})$ follows by the chain rule. Next,

$$= \sum_{j=1}^{M} w_{t-h-2}^{(j)} \int g(\mathbf{y}_{t-h-1:t}, \mathbf{x}_{t-h-1}|\mathbf{x}_{t-h-2}^{(j)})d\mathbf{x}_{t-h-1} \tag{19}$$

$$= \sum_{j=1}^{M} w_{t-h-2}^{(j)} g(\mathbf{y}_{t-h-1:t}|\mathbf{x}_{t-h-2}^{(j)}) \tag{20}$$

$$\tag{21}$$

which is the right-hand side of (9). □

Theorem 1 *The OAPF estimator of the normalizing constant is unbiased, i.e.,* $\mathbb{E}[\widehat{Z}_{1:t}] = p(\mathbf{y}_{1:t})$.

Proof: The unbiasedness is a consequence of Lemma 2 with $h = t - 1$. □

Now we look at the variance of the normalizing constant estimators. First, we establish a superiority in performance (i.e., equal or less variance) of the OAPF importance weights. This result is also used below to prove the convergence of the estimators by standard results in particle filtering.

Let us particularize importance weights in OAPF for the case with $K = M$ as

$$\widetilde{w}_t^{(m)} = \frac{g(\mathbf{y}_t|\mathbf{x}_t^{(m)})\sum_{i=1}^{M} w_{t-1}^{(i)} f(\mathbf{x}_t^{(m)}|\mathbf{x}_{t-1}^{(i)})}{\sum_{i=1}^{M} \lambda_t^{(i)} q_t^{(i)}(\mathbf{x}_t^{(m)}|\bar{\mathbf{x}}_{t-1}^{(i)})}. \tag{22}$$

We also consider the generalized APF weights given by

$$\widetilde{v}_t^{(m)} = \frac{g(\mathbf{y}_t|\mathbf{x}_t^{(m)})w_{t-1}^{(m)} f(\mathbf{x}_t^{(m)}|\mathbf{x}_{t-1}^{(m)})}{\lambda_t^{(m)} q_t^{(m)}(\mathbf{x}_t^{(m)}|\bar{\mathbf{x}}_{t-1}^{(m)})}. \tag{23}$$

These are generalized in the sense that the concrete APF described in the main paper is obtained by setting $\lambda_t^{(m)} \propto w_{t-1}^{(m)} g(\mathbf{y}_t|\boldsymbol{\mu}_t^{(m)})$ and propagating particles with transition kernels $f(\cdot)$, thus our following discussion holds for any choice of $\lambda_t^{(m)}$.

Lemma 3 *The conditional variance of* $\widehat{Z}_t^{OAPF}$ *using the OAPF weights in* (22) *is always less or equal than the same estimator* $\widehat{Z}_t^{APF}$ *using the APF weights in* (23).

Proof: First, note that $\widetilde{v}_t^{(m)}$ can be interpreted as an importance weight in an extended space on $\mathbf{x}_t$ and the auxiliary variable m (see for instance [Klaas et al., 2005, Section 3.1] and [Pitt and Shephard, 1999, Godsill, 2019]). Next, $\widetilde{w}_t^{(m)}$ can be interpreted as a version of $\widetilde{v}_t^{(m)}$ where both in the numerator (approximate filtering pdf) and denominator (proposal pdf), the auxiliary variable has been marginalized. Then, the variance inequality for each importance weight holds from the application of the variance decomposition lemma (also known as law of total variance). This proof generalizes the result in [Klaas et al., 2005] for any set of mixture weights $\{\lambda_t^{(m)}\}_{m=1}^{M}$, with $\sum_{j=1}^{M} \lambda_t^{(j)}$ and $\lambda_t^{(m)} \geq 0$, for all m. Finally, since both $\hat{Z}_t^{\text{OAPF}}$ and $\hat{Z}_t^{\text{APF}}$ are constructed as the average of the OAPF and APF weights, respectively, the conditional variance of $\hat{Z}_t^{\text{OAPF}}$ is necessarily upper-bounded by that of $\hat{Z}_t^{\text{APF}}$. □

We now address the consistency of the normalizing constant, $\widehat{Z}_{1:t}$, and the self-normalized IS (SNIS) estimator $\widehat{I}(h_t) = \sum_{m=1}^{M} w_t^{(m)} h_t(\mathbf{x}_t^{(m)})$.

Corollary 1 *The OAPF estimator of the normalizing constant $\widehat{Z}_{1:t}$ and the SNIS estimator $\widehat{I}(h_t)$ are consistent, i.e., $\lim_{M\to\infty} \widehat{Z}_{1:t} = p(\mathbf{y}_{1:t})$ and $\lim_{M\to\infty} \widehat{I}(h_t) = I(h_t)$ a.s. (almost surely) for a finite t.*

Proof: The consistency of $\widehat{Z}_{1:t}$ is a consequence of its unbiasedness, proved in Theorem 1, and the variance inequality in Lemma 3, which ensures the variance convergence to zero a.s. when $N \to \infty$ since the APF, which upper-bounds its variance, is also consistent [Doucet and Johansen, 2009, Section 3.6]. A similar argumentation can be done for the SNIS estimator $\widehat{I}(h_t)$. Note that the SNIS estimator can be re-expressed as $\widehat{I}(h_t) = \sum_{m=1}^{M} \frac{\widetilde{w}_t^{(m)}}{M\widehat{Z}_t} h_t(\mathbf{x}_t^{(m)}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\widetilde{w}_t^{(m)}}{\widehat{Z}_t} h_t(\mathbf{x}_t^{(m)})$. Since $\widehat{Z}_t$ is a consistent estimator of $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$, the denominator converges to $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ while the numerator converges to $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})I(h_t)$, when $N \to \infty$. Therefore, the ratio converges to $I(h_t)$ a.s. □

A.2 ADDITIONAL EXPERIMENTS AND RESULTS

A.2.1 Experiment 1

We provide all necessary parameters to reproduce Figure 1 in the main paper. We recall that in this toy example we do the Bayesian recursion from $t-1$ to t with $M = 4$ particles. In Figure 1(a), we have set the particles $\{\bar{\mathbf{x}}_{t-1}^{(m)}\}_{m=1}^{M=4} = \{2, 2.5, 3, 3.5\}$, the normalized weights $\{3/10, 3/10, 1/5, 1/5\}$, likelihood centered at 3, and $\sigma_{\text{lik}} = 0.8$, and $\sigma_{\text{kern}} = 0.5$. In Figure 1(b), $\{\bar{\mathbf{x}}_{t-1}^{(m)}\}_{m=1}^{M=4} = \{2, 2.5, 5, 5.5\}$, the normalized weights are $\{7/22, 1/11, 1/2, 1/11\}$, the likelihood is centered at 3.5, and $\sigma_{\text{lik}} = 1.2$, and $\sigma_{\text{kern}} = 0.5$. The proposals of all algorithms are then calculated as:

$$\sum_{m=1}^{4} \lambda_t^{(m)} f(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(m)}), \tag{24}$$

where the mixture weights $\lambda_t^{(m)}$ for BPF are $w_{t-1}^{(m)}$, for APF are $\propto$ $w_{t-1}^{(m)} g(\mathbf{y}_t|\boldsymbol{\mu}_t^{(m)})$, for IAPF $\propto$ $g(\mathbf{y}_t|\boldsymbol{\mu}_t^{(m)}) \sum_{m=1}^{M} w_{t-1}^{(m)} f(\boldsymbol{\mu}_t^{(m)}|\overline{\mathbf{x}}_{t-1}^{(m)}) / \sum_{m=1}^{M} f(\boldsymbol{\mu}_t^{(m)}|\overline{\mathbf{x}}_{t-1}^{(m)})$ and finally for OAPF they are the solution to the NNLS optimization problem.
As specified in the main paper and can be seen from (24), we used transition kernels as proposal kernels for OAPF. Moreover, we used the centers of the transition kernels $\boldsymbol{\mu}_t^{(m)}$ as evaluation points, which in this case they correspond to the resampled particles $\{\bar{\mathbf{x}}_{t-1}^{(m)}\}_{m=1}^{M=4}$.

A.2.2 Experiment 2

In this Section we provide results for estimation of the marginal likelihood, additional results in the estimation of the posterior mean and relevant equations for the linear dynamical model (LDM) (Experiment 2 in the main paper). The model is given by

$$p(\mathbf{x}_0) = \mathcal{N}_{\mathbf{x}_0}(\mathbf{m}_0, \boldsymbol{\Sigma}_0) \tag{25}$$

$$f\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}_{\mathbf{x}_t}\left(\mathbf{A}\mathbf{x}_{t-1} + \mathbf{c}, \mathbf{R}\right) \tag{26}$$

$$g\left(\mathbf{y}_t \mid \mathbf{x}_t\right) = \mathcal{N}_{\mathbf{y}_t}\left(\mathbf{C}\mathbf{x}_t + \mathbf{g}, \mathbf{Q}\right). \tag{27}$$

The posterior filtering distribution can be computed in closed form via the Kalman filter:

$$\begin{aligned} p(\mathbf{x}_t \mid \mathbf{y}_{1:t}) &= \mathcal{N}_{\mathbf{x}_t}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) && (28)\\ \boldsymbol{\mu}_t &= \overline{\boldsymbol{\mu}}_t + \mathbf{K}\left(\mathbf{y}_t - \mathbf{C}\overline{\boldsymbol{\mu}}_t - \mathbf{g}\right) && (29)\\ \boldsymbol{\Sigma}_t &= \left(\mathbf{I} - \mathbf{K}\mathbf{C}\right)\overline{\boldsymbol{\Sigma}}_t && (30)\\ \overline{\boldsymbol{\mu}}_t &= \mathbf{A}\boldsymbol{\mu}_{t-1} + \mathbf{c} && (31)\\ \overline{\boldsymbol{\Sigma}}_t &= \mathbf{A}\boldsymbol{\Sigma}_{t-1}\mathbf{A}^\top + \mathbf{R} && (32)\\ \mathbf{K} &= \overline{\boldsymbol{\Sigma}}_t\mathbf{C}^\top\left(\mathbf{C}\overline{\boldsymbol{\Sigma}}_t\mathbf{C}^\top + \mathbf{Q}\right)^{-1}. && (33) \end{aligned}$$

Moreover, $p(\mathbf{y}_{1:t})$ can also be computed in closed form from $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$. For numerical stability, one computes $\log p(\mathbf{y}_{1:t})$ and $\log p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$, which are given by:

$$\log p(\mathbf{y}_{1:t}) = \log p(\mathbf{y}_1) + \sum_{\tau=2}^{t} \log p(\mathbf{y}_\tau|\mathbf{y}_{1:\tau-1}) \tag{34}$$

$$\log p(\mathbf{y}_\tau|\mathbf{y}_{1:\tau-1}) = -\frac{1}{2}\left[\log(|\mathbf{C}\overline{\boldsymbol{\Sigma}}_\tau\mathbf{C}^\top + \mathbf{Q}|) + (\mathbf{y}_\tau - \mathbf{C}\overline{\boldsymbol{\mu}}_\tau - \mathbf{g})^{\mathrm{T}}(\mathbf{C}\overline{\boldsymbol{\Sigma}}_\tau\mathbf{C}^\top + \mathbf{Q})^{-1}(\mathbf{y}_\tau - \mathbf{C}\overline{\boldsymbol{\mu}}_\tau - \mathbf{g}) + d_{\mathbf{y}_\tau}\ln(2\pi)\right] \tag{35}$$

$$\log p(\mathbf{y}_1) = -\frac{1}{2}\left[\log(|\mathbf{C}\boldsymbol{\Sigma}_0\mathbf{C}^\top + \mathbf{Q}|) + (\mathbf{y}_\tau - \mathbf{C}\overline{\boldsymbol{\mu}}_1 - \mathbf{g})^{\mathrm{T}}(\mathbf{C}\boldsymbol{\Sigma}_0\mathbf{C}^\top + \mathbf{Q})^{-1}(\mathbf{y}_\tau - \mathbf{C}\overline{\boldsymbol{\mu}}_1 - \mathbf{g}) + d_{\mathbf{y}_1}\ln(2\pi)\right]. \tag{36}$$

We set $\mathbf{R} = 5\mathbf{I}$ and $\mathbf{Q} = 2.5\mathbf{I}$. This setting is of particular interest, as the kernels overlap and the observations are very informative. Therefore, the setting is particularly advantageous for IAPF, and hence it is more difficult to beat its performance. Moreover, $\mathbf{A} = \frac{1}{2}\mathbf{I}$ and $\mathbf{C} = \frac{1}{2}\mathbf{I}$. For $d_{\mathbf{x}} = 2$, then $\mathbf{c} = \mathbf{g} = (-2, 2)^\top$; for $d_{\mathbf{x}} = 5$, then $\mathbf{c} = \mathbf{g} = (-2, 2, -2, 2, -2)^\top$; similarly defined for $d_{\mathbf{x}} = 10$. The results for all $d_{\mathbf{x}}$ in the estimation of the marginal likelihood are shown in Table 1. We recall that OAPF ran with $\boldsymbol{K = 5, E = 5}$. This implies large computational savings with respect to the IAPF (or similarly to any other algorithm which uses the full mixture in the denominator of the importance weights). Finally, we also show additional results in the estimation of the posterior mean in Figure 1.

Table 1: Additional results for Experiment 2 in main paper. The Table shows normalized MSE to the true marginal likelihood $p(\mathbf{y}_{1:t})$, with standard errors over 100 Monte Carlo runs. Recall that whenever $d_{\mathbf{x}} \in \{2, 5\}$ then $M = 100$ and when $d_{\mathbf{x}} = 10$ then $M = 1000$.

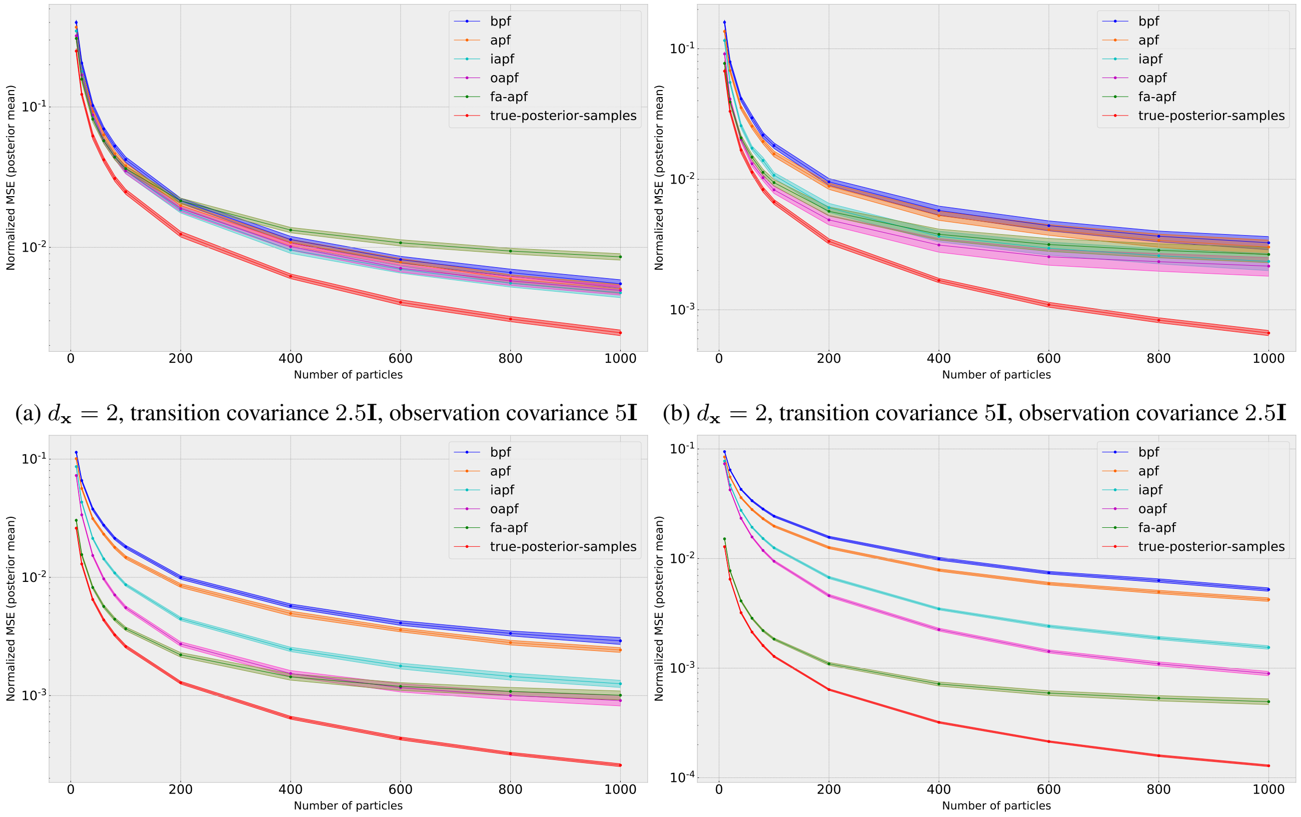
Method	$d_{\mathbf{x}} = 2$	$d_{\mathbf{x}} = 5$	$d_{\mathbf{x}} = 10$
BPF	$3.19 \cdot 10^{-7} \pm 3.14 \cdot 10^{-8}$	$5.09 \cdot 10^{-7} \pm 4.40 \cdot 10^{-8}$	$1.539 \cdot 10^{-7} \pm 1.267 \cdot 10^{-8}$
APF	$3.51 \cdot 10^{-7} \pm 3.88 \cdot 10^{-8}$	$4.68 \cdot 10^{-7} \pm 4.49 \cdot 10^{-8}$	$1.333 \cdot 10^{-7} \pm 1.224 \cdot 10^{-8}$
IAPF	$2.15 \cdot 10^{-7} \pm 2.25 \cdot 10^{-8}$	$1.63 \cdot 10^{-7} \pm 1.58 \cdot 10^{-8}$	$6.330 \cdot 10^{-8} \pm 6.204 \cdot 10^{-9}$
OAPF	$\mathbf{1.35 \cdot 10^{-7} \pm 1.23 \cdot 10^{-8}}$	$\mathbf{9.67 \cdot 10^{-8} \pm 9.03 \cdot 10^{-9}}$	$\mathbf{4.771 \cdot 10^{-8} \pm 5.329 \cdot 10^{-9}}$

A.2.3 Experiment 3

In this experiment, we have used the standard parameters for the Lorenz model given by $(s, r, b) = (10, 28, 2.667)$. We set transition and observation noise as independent standard normally distributed random variables. In Figure 3, we show visually, as stated in the main paper, how a small change in Δt can lead to very different trajectories of $\mathbf{x}_t$. The sensitivity to Δt, to the initialization, and even to the parameters (s, r, b), jointly with the strong non-linearity of the generated trajectories, make the Lorenz model particularly challenging.

A.2.4 Experiment 4

In this experiment, we have used a challenging multivariate stochastic volatility model, which is common in related works (see for instance [Guarniero, 2017]). Additional results with parameters $\mathbf{m} = \mathbf{0}, \mathrm{U}_0 = \mathbf{I}, \mathbf{U} = \mathbf{I}, \phi = \mathbf{1}$ are shown in Table



(a) $d_{\mathbf{x}} = 2$, transition covariance $2.5\mathbf{I}$, observation covariance $5\mathbf{I}$ (b) $d_{\mathbf{x}} = 2$, transition covariance $5\mathbf{I}$, observation covariance $2.5\mathbf{I}$

(c) $d_{\mathbf{x}} = 5$, transition covariance $5\mathbf{I}$, observation covariance $2.5\mathbf{I}$ (d) $d_{\mathbf{x}} = 10$, transition covariance $5\mathbf{I}$, observation covariance $2.5\mathbf{I}$

Figure 1: Here, we show results for the linear Gaussian model for additional settings (complement to Figure 2(a)). In all settings, OAPF ran with $\boldsymbol{K = 5, E = 5}$. Notice that the closed form FA-APF is often outperformed by OAPF, and sometimes even by BPF.

2.

B DISCUSSION ON NUMBER OF EVALUATION POINTS AND KERNELS

We expand here our intuition for the need of only few kernels/evaluation points in many scenarios (see Figure 2 for explanation).

C DISCUSSION ON THE FULLY ADAPTED PF

Previous works have discussed how the FA-APF described in [Pitt and Shephard, 1999, Pitt et al., 2012] optimality criterion is not optimal in a global sense: the main intuition they provide is that it only minimizes the one-step variance of the importance weights [Johansen and Doucet, 2008, Whiteley and Johansen, 2011, Chopin and Papaspiliopoulos, 2020]. Here, we will provide a perspective inspired by MIS to informally explain how FA-APF can be suboptimal in general.
Let us assume that we have access to M samples simulated exactly from the filtering distribution at time $t-1$:

$$\mathbf{x}_{t-1}^{(m)} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \qquad m = 1, \dots, M. \tag{37}$$

These M samples can be used to form a particle approximation of $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ simply as:

$$p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \approx \frac{1}{M}\sum_{m=1}^{M} \delta_{\mathbf{x}_{t-1}^{(m)}}. \tag{38}$$

Table 2: Results with additional parameters for Experiment 4. Note that when $d_{\mathbf{x}} = 10$ then $M = 1000$. Averaged ESS and standard errors over 100 Monte Carlo runs.

Method	$d_{\mathbf{x}} = 2$	$d_{\mathbf{x}} = 5$	$d_{\mathbf{x}} = 10$
BPF	50.8 ± 0.2	21.2 ± 0.4	46.6 ± 0.5
APF	59.7 ± 0.2	31.9 ± 0.4	83.9 ± 0.6
IAPF	80.5 ± 0.1	49.4 ± 0.5	199.9 ± 1.7
OAPF	$\mathbf{92.6 \pm 0.1}$	$\mathbf{59.5 \pm 0.7}$	$\mathbf{229.5 \pm 2.4}$

This particle approximation can in turn be used to approximate the intractable integral in the definition of the filtering posterior and form an approximation to it:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto g(\mathbf{y}_t|\mathbf{x}_t) \int f(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \tag{39}$$

$$\approx g(\mathbf{y}_t|\mathbf{x}_t) \int f(\mathbf{x}_t|\mathbf{x}_{t-1}) \frac{1}{M} \sum_{m=1}^{M} \delta_{\mathbf{x}_{t-1}^{(m)}} \mathrm{d}\mathbf{x}_{t-1} \qquad \text{substituting } p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \text{ for approximation in Eq. (38)} \tag{40}$$

$$= g(\mathbf{y}_t|\mathbf{x}_t) \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}) \tag{41}$$

Now, we will exploit the identity used by the FA-APF. The identity in question is:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t) = \frac{g(\mathbf{y}_t|\mathbf{x}_t) f(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{y}_t|\mathbf{x}_{t-1})}; \tag{42}$$

often the term $p(\mathbf{y}_t|\mathbf{x}_t - 1)$ is referred to as *predictive likelihood*. The FA-APF propagates each particle m using $p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}, \mathbf{y}_t)$ and resamples with weights $\frac{p(\mathbf{y}_t|\mathbf{x}_{t-1}^{(m)})}{M}$. It is easy to derive that this leads to constant importance weights $w_t^{(m)}$, when these are defined as :

$$w_t^{(m)} = \frac{p(\mathbf{x}_{1:t}^{(m)}|\mathbf{y}_{1:t})}{q(\mathbf{x}_{1:t-1}^{(m)}|\mathbf{y}_{1:t}) q(\mathbf{x}_t^{(m)}|\mathbf{y}_t, \mathbf{x}_{t-1}^{(m)})} \tag{43}$$

using a joint proposal and target as common in SMC.
Our observation is that the choices made by FA-APF can be viewed as sampling from the mixture in Eq. (41), when rearranged using (42):

$$\frac{1}{M} g(\mathbf{y}_t|\mathbf{x}_t) \sum_{m=1}^{M} f(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}) = \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_t|\mathbf{x}_{t-1}^{(m)}) p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}, \mathbf{y}_t), \tag{44}$$

since $g(\mathbf{y}_t|\mathbf{x}_t) f(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)}, \mathbf{y}_t) p(\mathbf{y}_t|\mathbf{x}_{t-1}^{(m)})$, where indeed i.i.d. sampling from this mixture is equivalent to resampling and propagating in FA-APF.
This observation highlights some of the assumptions behind FA-APF from a different perspective: firstly, we assumed i.i.d. samples from the true filtering distribution at $t - 1$ were available; secondly, we formed a particle approximation to $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ in Eq. (39) which may be more or less accurate depending on the situation. Therefore, the FA-APF choices of resampling weights and kernels, even when analytically available, can still lead to very poor performance if (1) the previous set of samples is a bad approximation of $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, and consequently if the approximation to the predictive distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ is poor.

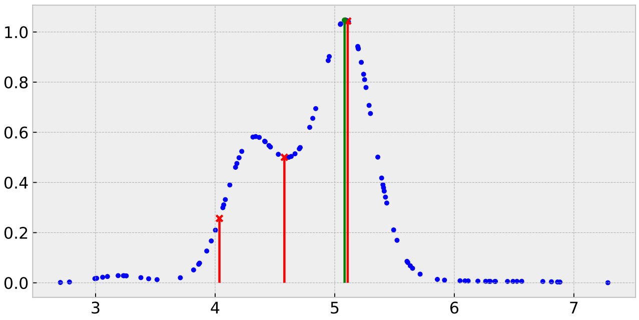
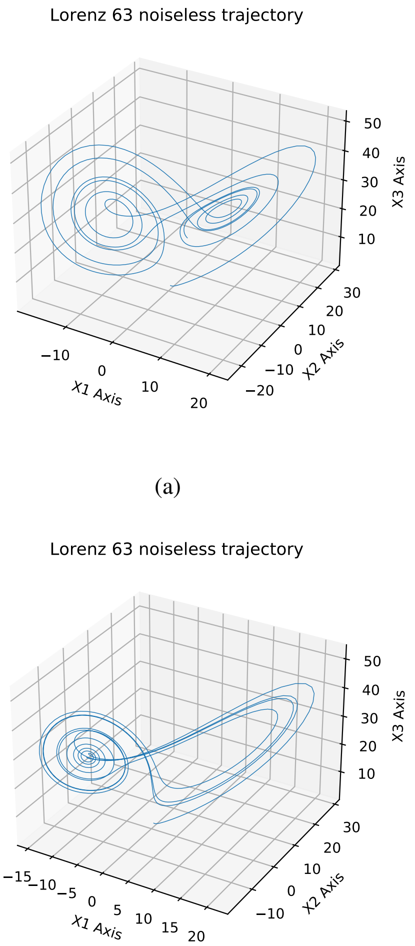


Figure 2: In blue, evaluations of a posterior generated by multiplying a mixture of 100 Gaussian kernels with a Gaussian likelihood. Building the OAPF proposal would take approximately 100^3 computation steps. However, one can see that it seems highly wasteful to adapt the mixture of Gaussian kernels proposal by evaluating at *all blue points*: e.g., if the proposal matches the posterior at the rightmost red point, it will probably go through the green point too: we do not need to include that point in the optimization. Moreover, it may also be wasteful to match proposal and target at points where the target has little probability mass. From these considerations, one may consider that trying to only match the three highlighted red points (and perhaps a couple more), would likely result in a proposal that is closely as good as the one we would get by using all blue points.



(a)

(b)

Figure 3: In this figure, we show the noiseless versions of a trajectory for $\Delta t = 0.01$ (a) and $\Delta t = 0.008$ (b) to emphasize how different trajectories can be in a Lorenz 63 model even with small parameter changes. Note that particle filters will have to deal with both transition noise and observation noise.

References

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.

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

Víctor Elvira, Luca Martino, David Luengo, Mónica F Bugallo, et al. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

Simon Godsill. Particle filtering: the first 25 years and beyond. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7760–7764. IEEE, 2019.

Pieralberto Guarniero. *The Iterated Auxiliary Particle Filter and Applications to State Space Models and Diffusion Processes*. PhD thesis, University of Warwick, 2017.

Adam M Johansen and Arnaud Doucet. A note on auxiliary particle filters. *Statistics & Probability Letters*, 78(12): 1498–1504, 2008.

Mike Klaas, Nando de Freitas, and Arnaud Doucet. Toward practical n2 monte carlo: the marginal particle filter. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 308–315, Arlington, Virginia, 2005. AUAI Press.

Art B Owen. Monte carlo theory, methods and examples. 2013.

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

Michael K Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of markov chain monte carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

Nick Whiteley and Adam M Johansen. Auxiliary particle filtering: recent developments. *Bayesian time series models. Cambridge University Press, Cambridge*, 2011.