On Sequential Simulation-Based Methods for Bayesian Filtering

Arnaud Doucet
Signal Processing Group, Department of Engineering
University of Cambridge CB2 1PZ Cambridge
Email: AD2@ENG.CAM.AC.UK

Technical report CUED/F-INFENG/TR.310 (1998)

Abstract. In this report, we present an overview of sequential simulation-based methods for Bayesian filtering of nonlinear and non-Gaussian dynamic models. It includes in a general framework numerous methods proposed independently in various areas of science and proposes some original developments.

Keywords: Bayesian estimation, optimal filtering, nonlinear non-Gaussian state space models, hidden Markov models, sequential Monte Carlo methods.

1. Introduction

[1] Many problems in statistical signal processing, automatic control, applied statistics or econometrics can be stated as follows. A transition equation describes the prior distribution of the Markovian hidden signal of interest $\{\mathbf{x}_k; k \in \mathbb{N}\}$, the so-called hidden state process, and an observation equation describes the likelihood of the observations $\{\mathbf{y}_k; k \in \mathbb{N}\}$, k being the discrete time index. The aim is to estimate the hidden state process using the observations. In the Bayesian framework, all relevant information on $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$ at time k is included in the posterior distribution $p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k | \mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_k)$. In many applications in signal processing, we are interested in *estimating recursively in time* this distribution and especially one of its marginals, the so-called filtering distribution $p(\mathbf{x}_k | \mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_k)$. This problem is known as the Bayesian filtering problem, also called the optimal filtering or stochastic filtering problem. Except for a few cases including linear Gaussian state space models (Kalman filter) and hidden finite-state space Markov chains (Wohnam filter), it is impossible to evaluate analytically these distributions. From the mid 60's, a huge number of papers and books have been devoted to obtaining approximations of these distributions, see [32] for example. The most popular algorithms, the extended Kalman filter and the Gaussian sum filter, rely on analytical approximations [5, 6] but early well-known work relying on deterministic numerical integration methods was also performed by Bucy and co-workers, see [13] for example. Other interesting work in automatic control was done during the 60's and 70's based on sequential Monte Carlo integration methods, see [1, 2, 4, 25, 26, 47, 61]. Most likely because of the primitive computers available at the time, these last algorithms were overlooked and forgotten [2]. In the late 80's, the great increase of computational power allowed the rebirth of numerical integration methods for Bayesian filtering [33]. Current research has now focused on MC (Monte Carlo) integration methods which have the great advantage of not being subject to any linearity or Gaussianity hypotheses on the model.

The main objective of this report is to include in a unified framework many old and recent algorithms developed independently in various fields of applied science. Some original developments are also presented. The closest work to this report is the work of Liu and Chen [41], developed independently, which underlines similarly the central rôle of sequential importance sampling (SIS) in sequential simulation-based methods for Bayesian filtering.

[1] This technical report is a translation of chapter 3 of [19] in abbreviated form.

[2] To the best of my knowledge, these important works are cited neither in any standard article and book on optimal estimation nor in any current work on the subject.

This report is organized as follows. In section 2, we briefly review the Bayesian filtering problem. A classical MC method, Bayesian importance sampling, is proposed to solve it. We then present a sequential version of this method which allows us to obtain a general recursive MC filter. This algorithm is based on the introduction of a probability distribution known as the importance function. Under a given criterion, we obtain the optimal importance function. Unfortunately, for numerous models, one cannot use this importance function, which is why we propose several suboptimal distributions of practical interest and retrieve as particular cases many algorithms presented independently in the literature. In Section 3, a resampling scheme is used to limit practically the degeneracy of the algorithm. In Section 4, we apply the Rao-Blackwellisation method to SIS and obtain efficient hybrid analytical/MC filters. In Section 5, we show how to use the MC filter to compute the prediction and fixed-interval smoothing distributions as well as the likelihood. Finally, a few simulations are presented in Section 6.

2. Bayesian Estimation for Hidden Markov Models using Importance Sampling

The signal $\{\mathbf{x}_k; k \in \mathbb{N}\}$, $\mathbf{x}_k \in \mathbb{R}^{n_{\mathbf{x}}}$, is an unobserved (hidden) Markov process of initial distribution $p(\mathbf{x}_0)$ and transition equation $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. The observations $\{\mathbf{y}_k; k \in \mathbb{N}\}$, $\mathbf{y}_k \in \mathbb{R}^{n_{\mathbf{y}}}$, are conditionally independent given the process $\{\mathbf{x}_k; k \in \mathbb{N}\}$ of marginal distribution $p(\mathbf{y}_k|\mathbf{x}_k)$. To sum up, the model is a hidden Markov model (HMM) described by

$$p(\mathbf{x}_0) \text{ and } p(\mathbf{x}_k|\mathbf{x}_{k-1}) \text{ for } k \geq 1 \tag{1}$$

$$p(\mathbf{y}_k|\mathbf{x}_k) \text{ for } k \geq 0 \tag{2}$$

We denote by $\mathbf{x}_{0:n} \triangleq \{\mathbf{x}_0, ..., \mathbf{x}_n\}$ and $\mathbf{y}_{0:n} \triangleq \{\mathbf{y}_0, ..., \mathbf{y}_n\}$, respectively, the signal and the observations up to time n. Our aim is to estimate recursively in time the distribution $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ and its associated features including $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ and the expectation

$$I(f_n) = \mathbb{E}_{p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})}(f_n(\mathbf{x}_{0:n})) = \int f_n(\mathbf{x}_{0:n})\, p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})\, d\mathbf{x}_{0:n} \tag{3}$$

for any $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$-integrable $f_n : \mathbb{R}^{(n+1)\times n_{\mathbf{x}}} \to \mathbb{R}$. We obtain straightforwardly a recursive formula for $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$:

$$\begin{aligned} p(\mathbf{x}_{0:n+1}|\mathbf{y}_{0:n+1}) &= \frac{p(\mathbf{y}_{0:n+1}|\mathbf{x}_{0:n+1})\, p(\mathbf{x}_{0:n+1})}{p(\mathbf{y}_{0:n+1})} \\ &= p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) \times \frac{p(\mathbf{y}_{n+1}|\mathbf{x}_{n+1})\, p(\mathbf{x}_{n+1}|\mathbf{x}_n)}{p(\mathbf{y}_{n+1}|\mathbf{y}_{0:n})} \end{aligned} \tag{4}$$

This recursion is only academic in the sense that one cannot typically compute the normalizing constant $p(\mathbf{y}_{0:n+1})$, the marginals of $p(\mathbf{x}_{0:n+1}|\mathbf{y}_{0:n+1})$ (in particular $p(\mathbf{x}_{n+1}|\mathbf{y}_{n+1})$) and $I(f_{k+1})$ because it requires the ability to evaluate complex high-dimensional integrals. A numerical solution consists of using a Monte Carlo integration method.

Later, we will assume that we know how to sample according to $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and that we can evaluate $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and $p(\mathbf{y}_k|\mathbf{x}_k)$ pointwise.

2.1. Perfect Monte Carlo sampling. Let us assume that we are able to simulate N i.i.d. random samples $\left\{\mathbf{x}_{0:n}^{(i)}; i = 1, ..., N\right\}$ according to $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$. An empirical estimate of this distribution is given by:

$$\widehat{P}(d\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}_{0:n}^{(i)}}(d\mathbf{x}_{0:n}) \tag{5}$$

and one obtains the following estimate:

$$\overline{I_N}(f_n) = \int f_n(\mathbf{x}_{0:n})\,\widehat{P}(d\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) = \frac{1}{N}\sum_{i=1}^{N} f_n\left(\mathbf{x}_{0:n}^{(i)}\right) \tag{6}$$

From the strong law of large numbers (SLLN)

$$\overline{I_N}(f_n) \underset{N\to+\infty}{\overset{a.s.}{\to}} I(f_n) \tag{7}$$

where $\overset{a.s.}{\to}$ denotes almost sure convergence. If the posterior variance of $f_n(\mathbf{x}_{0:n})$ satisfies

$$\begin{aligned} \overline{\sigma}_{f_n}^2 &\triangleq var_{p(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})] \\ &= \mathbb{E}_{p(\cdot|\mathbf{y}_{0:n})}\left[f_n^2(\mathbf{x}_{0:n})\right] - \mathbb{E}^2_{p(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})] < +\infty \end{aligned} \tag{8}$$

then a central limit theorem holds:

$$\sqrt{N}[\overline{I_N}(f_n) - I(f_n)] \underset{N\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \overline{\sigma}_{f_n}^2\right) \tag{9}$$

where $\Longrightarrow$ denotes convergence in distribution. The advantage of this perfect MC method is clear. From the set of random samples $\left\{\mathbf{x}_{0:n}^{(i)}; i = 1, ..., N\right\}$, one can easily estimate any quantity $\mathbb{E}_{p(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})]$ and the speed of convergence of this estimate neither depends on $n_{\mathbf{x}} \times n$ nor on $f_n(\mathbf{x}_{0:n})$ but only on N. Unfortunately, it is usually impossible to sample efficiently from the posterior distribution $p(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$ at any time n, $p(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$ being multivariate, non standard and only known up to a proportionality constant.

2.2. Bayesian Importance Sampling. An alternative solution consists of using the importance sampling (IS) method. The basic idea of this method is the following. We choose a so-called importance function, that is a probability distribution $\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$ (which depends here on the observations until time n) from which one can easily sample. The IS method is based on the following simple remark. If $p(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) > 0$ implies $\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) > 0$ then one can write:

$$\begin{aligned} I(f_n) &= \int f_n(\mathbf{x}_{0:n}) \frac{p(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})}{\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})} \pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})\, d\mathbf{x}_{0:n} &(10)\\ &= \mathbb{E}_{\pi(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})\, w^*(\mathbf{x}_{0:n})] &(11) \end{aligned}$$

where

$$w^*(\mathbf{x}_{0:n}) = p(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) / \pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) \tag{12}$$

Thus if one can simulate N i.i.d. samples $\{\mathbf{x}_{0:n}, i = 1, ..., N\}$ according to $\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$, a possible estimate of $I(f_n)$ is:

$$\widehat{I_N^*}(f_n) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_n\left(\mathbf{x}_{0:n}^{(i)}\right) w_n^{*(i)} \tag{13}$$

where the *importance weights* $\left\{w_n^{*(i)}, i = 1, ..., N\right\}$ are equal to:

$$w_n^{*(i)} = w^*\left(\mathbf{x}_{0:n}^{(i)}\right) = \frac{p(\mathbf{x}_{0:n}^{(i)}\,|\,\mathbf{y}_{0:n})}{\pi\left(\mathbf{x}_{0:n}^{(i)}\,|\,\mathbf{y}_{0:n}\right)} = \frac{p(\mathbf{y}_{0:n}|\,\mathbf{x}_{0:n}^{(i)}) p(\mathbf{x}_{0:n}^{(i)})}{p(\mathbf{y}_{0:n})\pi\left(\mathbf{x}_{0:n}^{(i)}\,|\,\mathbf{y}_{0:n}\right)} \tag{14}$$

The estimate (13) is unbiased and converges a.s. according to the SLLN toward $I(f_n)$ when $N \to +\infty$.

In a Bayesian framework, this estimate cannot generally be used as it requires the knowledge of the normalizing constant $p(\mathbf{y}_{0:n})$:

$$p(\mathbf{y}_{0:n}) = \int p(\mathbf{y}_{0:n}|\mathbf{x}_{0:n})\, p(\mathbf{x}_{0:n})\, d\mathbf{x}_{0:n} \tag{15}$$

Typically $p(\mathbf{y}_{0:n})$ cannot be expressed in closed form. However, one can observe that

$$I(f_n) = \frac{\mathbb{E}_{\pi(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})\, w(\mathbf{x}_{0:n})]}{\mathbb{E}_{\pi(\cdot|\mathbf{y}_{0:n})}[w(\mathbf{x}_{0:n})]} \tag{16}$$

where

$$w(\mathbf{x}_{0:n}) = p(\mathbf{y}_{0:n}|\mathbf{x}_{0:n})\, p(\mathbf{x}_{0:n}) / \pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) \tag{17}$$

Thus an estimate of $I(f_n)$ is given by the ratio of the estimates of the numerator and denominator obtained using the "classical" importance sampling method:

$$\widehat{I_N}(f_n) = \frac{\frac{1}{N}\sum_{i=1}^{N} f_n\left(\mathbf{x}_{0:n}^{(i)}\right) w_n^{(i)}}{\frac{1}{N}\sum_{j=1}^{N} w_n^{(j)}} = \sum_{i=1}^{N} f_n\left(\mathbf{x}_{0:n}^{(i)}\right) \widetilde{w}_n^{(i)} \tag{18}$$

where the *unnormalised importance weights* $\left\{w_n^{(i)}, i = 1, ..., N\right\}$ are equal to

$$w_n^{(i)} = w\left(\mathbf{x}_{0:n}^{(i)}\right) = p\left(\mathbf{y}_{0:n}|\mathbf{x}_{0:n}^{(i)}\right) p\left(\mathbf{x}_{0:n}^{(i)}\right) / \pi\left(\mathbf{x}_{0:n}^{(i)}\middle|\mathbf{y}_{0:n}\right) \tag{19}$$

$$\propto w_n^{*(i)} \tag{20}$$

($\propto$ means proportional to) and the normalised importance weights are equal to

$$\widetilde{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^{N} w_n^{(j)}} \tag{21}$$

The "true" importance weights $w_n^{*(i)}$ have been replaced by the following estimate:

$$\widehat{w}_n^{*(i)} = N\widetilde{w}_n^{(i)} \tag{22}$$

This method is well-known in the statistical literature as Bayesian IS, see for example [22, 51]. We recall here some classical results on this MC method.

Assumption 1

- $\left\{\mathbf{x}_{0:n}^{(i)}; i = 1, ..., N\right\}$ is a set of i.i.d. vectors distributed according to $\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$.
- The support $\overline{\pi} = \left\{\mathbf{x}_{0:n} \in \mathbb{R}^{n_{\mathbf{x}} \times (n+1)} / \pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) > 0\right\}$ of $\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ includes the support $\overline{p} = \left\{\mathbf{x}_{0:n} \in \mathbb{R}^{n_{\mathbf{x}} \times (n+1)} / p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) > 0\right\}$ of $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$.
- $I(f_n)$ exists and is finite.

Assumption 2

- $\mathbb{E}_{p(\cdot|\mathbf{y}_{0:n})}[w(\mathbf{x}_{0:n})] < +\infty$ and $\mathbb{E}_{p(\cdot|\mathbf{y}_{0:n})}[f_n^2(\mathbf{x}_{0:n})\, w(\mathbf{x}_{0:n})] < +\infty$.

A sufficient condition to verify assumption 2 is [22]:

$$var_{p(\cdot|\mathbf{y}_{0:n})}[f_n(\mathbf{x}_{0:n})] < +\infty \text{ and } w(\mathbf{x}_{0:n}) < C_n < +\infty \text{ for any } \mathbf{x}_{0:n} \in \overline{\pi} \tag{23}$$

Proposition 1. *For N finite, $\widehat{I_N}(f_n)$ is biased but asymptotically, under assumption 1, the SLLN yields:*

$$\widehat{I_N}(f_n) \underset{N\to+\infty}{\overset{a.s.}{\to}} I(f_n) \tag{24}$$

Under assumption 1, the previous proposition implies a convergence of the empirical distribution $\sum_{i=1}^{N} \widetilde{w}_n^{(i)} \delta_{\mathbf{x}_{0:n}^{(i)}}(d\mathbf{x}_{0:n})$ towards $P(d\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$ in the sense of a.s. convergence of $\widehat{I_N}(f_n)$ for any function f_n such that $I(f_n)$ exists and is finite. This result is important as it means that we can interpret the IS method as a simulation method to sample from $P(d\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n})$ rather than as an integration method, see [22] for a similar interpretation. Using the delta method, we also obtain the following proposition.

Proposition 2. *(Geweke 1989 [22]) Under assumptions 1 and 2,*

$$\sqrt{N}[\widehat{I_N}(f_n) - I(f_n)] \underset{N\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \sigma_{f_n}^2\right) \tag{25}$$

where

$$\sigma_{f_n}^2 = \mathbb{E}_{p(\,.\,|\mathbf{y}_{0:n})}\left(\left(f_n(\mathbf{x}_{0:n}) - \mathbb{E}_{p(\,.\,|\mathbf{y}_{0:n})}(f_n(\mathbf{x}_{0:n}))\right)^2 w(\mathbf{x}_{0:n})\right) \tag{26}$$

We show in the following subsection how it is possible to obtain easily a recursive MC filter using Bayesian IS.

2.3. Monte Carlo filter using sequential importance sampling. One can always rewrite the importance function as follows:

$$\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) = \pi(\mathbf{x}_0|\,\mathbf{y}_{0:n}) \prod_{k=1}^{n} \pi(\mathbf{x}_k|\,\mathbf{x}_{0:k-1}, \mathbf{y}_{0:n}) \tag{27}$$

where $\pi(\mathbf{x}_k|\,\mathbf{x}_{0:k-1}, \mathbf{y}_{0:n})$ is the probability density function of $\mathbf{x}_k$ conditional upon $\mathbf{x}_{0:k-1}$ and $\mathbf{y}_{0:n}$.

Our aim is to obtain at time k an estimate of the distribution $p(\mathbf{x}_{0:k}|\,\mathbf{y}_{0:k})$ and to be able to propagate this estimate in time without modifying subsequently the past simulated trajectories $\left\{\mathbf{x}_{0:k}^{(i)}; i = 1, ..., N\right\}$. This means that the importance function at time $k+1$ admits as a marginal distribution at time k the importance function $\pi(\mathbf{x}_{0:k}|\,\mathbf{y}_{0:k})$. This is possible if we restrict ourselves to importance functions of the following form:

$$\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) = \pi(\mathbf{x}_{0:n-1}|\,\mathbf{y}_{0:n-1})\,\pi(\mathbf{x}_n|\,\mathbf{x}_{0:n-1}, \mathbf{y}_{0:n}) \tag{28}$$

Iterating, it yields:

$$\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}) = \pi(\mathbf{x}_0|\,\mathbf{y}_0) \prod_{k=1}^{n} \pi(\mathbf{x}_k|\,\mathbf{x}_{0:k-1}, \mathbf{y}_{0:k}) \tag{29}$$

This importance function allows to evaluate recursively in time the importance weights (19) and (21).

Remark 1. *This assumption could be weakened. For example, one can consider the case where one is interested in an estimate of the fixed-lag distribution $p(\mathbf{x}_{0:k}|\,\mathbf{y}_{0:k+p})$, $p \in \mathbb{N}^*$ being fixed. In this case, one can choose:*

$$\pi(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n+p}) = \pi(\mathbf{x}_0|\,\mathbf{y}_{0:p}) \prod_{k=1}^{n} \pi(\mathbf{x}_k|\,\mathbf{x}_{0:k-1}, \mathbf{y}_{0:k+p})$$

Under the assumption (29), we obtain straightforwardly the following MC filter.
Algorithm : Sequential Importance Sampling (SIS)

1. At time $k = 0$,
   - For $i = 1, ..., N$, sample $\mathbf{x}_0^{(i)} \sim \pi(\mathbf{x}_0 | \mathbf{y}_0)$.
   - For $i = 1, ..., N$, evaluate the importance weights up to a normalizing constant:
   $$w_0^{(i)} = \frac{p\left(\mathbf{y}_0 | \mathbf{x}_0^{(i)}\right) p\left(\mathbf{x}_0^{(i)}\right)}{\pi\left(\mathbf{x}_0^{(i)} \middle| \mathbf{y}_0\right)} \tag{30}$$
   - For $i = 1, ..., N$, normalise the importance weights:
   $$\widetilde{w}_0^{(i)} = \frac{w_0^{(i)}}{\sum_{j=1}^N w_0^{(j)}} \tag{31}$$
2. For times $k \geq 1$,
   - For $i = 1, ..., N$, sample $\mathbf{x}_k^{(i)} \sim \pi\left(\mathbf{x}_k | \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{0:k}\right)$ and $\mathbf{x}_{0:k}^{(i)} \triangleq \left(\mathbf{x}_{0:k-1}^{(i)}, \mathbf{x}_k^{(i)}\right)$.
   - For $i = 1, ..., N$, evaluate the importance weights up to a normalizing constant:
   $$w_k^{(i)} = w_{k-1}^{(i)} \frac{p\left(\mathbf{y}_k | \mathbf{x}_k^{(i)}\right) p\left(\mathbf{x}_k^{(i)} \middle| \mathbf{x}_{k-1}^{(i)}\right)}{\pi(\mathbf{x}_k^{(i)} \middle| \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{0:k})} \tag{32}$$
   - For $i = 1, ..., N$, normalise the importance weights:
   $$\widetilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{j=1}^N w_k^{(j)}} \tag{33}$$

Numerous algorithms proposed in the literature are special cases of this general (and simple) algorithm. A particular case of this algorithm was introduced in 1969 by Handschin and Mayne [25, 26] !

The numerical complexity of this algorithm is $O(N)$. This is important as we take $N \gg 1$ in practice but it has the great advantage of being parallelizable. In the general case, the memory requirements are $O((k+1)N)$ as it is necessary to keep all the N simulated trajectories from time 0 to time k. However, if $\pi\left(\mathbf{x}_k | \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{0:k}\right) = \pi\left(\mathbf{x}_k | \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right)$ and if one is only interested in the filtering distribution $p(\mathbf{x}_k | \mathbf{y}_{0:k})$, the memory requirements are $O(N)$.

In the general case, one obtains at time k the following estimate of the joint posterior distribution:

$$\widehat{P}(d\mathbf{x}_{0:k} | \mathbf{y}_{0:k}) = \sum_{i=1}^N \widetilde{w}_k^{(i)} \delta_{\mathbf{x}_{0:k}^{(i)}}(d\mathbf{x}_{0:k}) \tag{34}$$

and an estimate of $I(f_k)$:

$$\begin{aligned} \widehat{I_N}(f_k) &= \int f_k(\mathbf{x}_{0:k}) \widehat{P}(d\mathbf{x}_{0:k} | \mathbf{y}_{0:k}) \\ &= \sum_{i=1}^N \widetilde{w}_k^{(i)} f_k\left(\mathbf{x}_{0:k}^{(i)}\right) \end{aligned} \tag{35}$$

Assumption 1 which ensures the asymptotic convergence of these estimates is quite weak. In practice, one obtains however poor performance of these estimates when the importance function is not well-chosen. The choice of the importance function is the topic of the following sections.

2.4. Degeneracy of the algorithm. When interpreting IS as a MC sampling method rather than as a MC integration method, the best possible choice would consist of selecting the posterior distribution of interest $p(\mathbf{x}_{0:k}|\mathbf{y}_{0:k})$ as importance function $\pi(\mathbf{x}_{0:k}|\mathbf{y}_{0:k})$. Then we would obtain for the importance weights $\mathbb{E}_{\pi(\cdot|\mathbf{y}_{0:k})}(w^*(\mathbf{x}_{0:k})) = 1$ and $var_{\pi(\cdot|\mathbf{y}_{0:k})}(w^*(\mathbf{x}_{0:k})) = 0$. We would like to be close to this case. But for importance functions of the form (29), the variance of the importance weights can only increase (stochastically) over time.

Proposition 3. *The unconditional variance of the importance weights, i.e. with the observations $\mathbf{y}_{0:k}$ being interpreted as random variables, increases over time.*

The proof of this proposition is a straightforward extension of a Kong-Liu-Wong [37, p. 285] theorem to the case of an importance function of the form (29). Thus, it is impossible to avoid a degeneracy phenomenon. Practically, after a few iterations of the algorithm, all but one of the normalised importance weights are very close to zero, a large computational burden is devoted to updating trajectories whose contribution to the final estimate is almost zero.

2.5. Selection of the importance function. Practically, at time $k-1$, the importance weights $w_{k-1}^{*(i)}$, $i = 1, ..., N$ are fixed. To limit degeneracy of the algorithm, a natural strategy consists of selecting the importance function which minimizes the variance of the importance weights conditional upon the simulated trajectory $\mathbf{x}_{0:k-1}^{(i)}$ and the observations $\mathbf{y}_{0:k}$.

Proposition 4. *$p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right)$ is the importance function which minimizes the variance of the importance weight $w_k^{*(i)}$ conditional upon $\mathbf{x}_{0:k-1}^{(i)}$ and $\mathbf{y}_{0:k}$.*

The proof is straightforward [19]. First we present how to implement the optimal importance function $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right)$.

Optimal importance function. The optimal importance function $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right)$ has been introduced by Zaritskii *et al.* [61] then by Akashi *et al.* for a particular case [4]. More recently, this importance function has been used in [14, 15, 16, 30, 37, 38, 40]. For this distribution, we obtain using (32) the following expression for the importance weight:

$$\begin{aligned} w_k^{(i)} &= w_{k-1}^{(i)} p\left(\mathbf{y}_k|\,\mathbf{x}_k^{(i)}\right) p\left(\mathbf{x}_k^{(i)}\middle|\,\mathbf{x}_{k-1}^{(i)}\right) / p\left(\mathbf{x}_k^{(i)}\middle|\,\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right) \\ &= w_{k-1}^{(i)} p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right) \end{aligned} \tag{36}$$

Remark 2. *In this case, the importance weight $w_k^{(i)}$ does not depend on $\mathbf{x}_k^{(i)}$. This is interesting in practice as it allows parallelization of the simulation of $\left\{\mathbf{x}_k^{(i)}; i = 1, ..., N\right\}$ and the evaluation of $\left\{w_k^{(i)}; i = 1, ..., N\right\}$.*

Remark 3. *To verify Proposition 2, a sufficient condition which ensures that the importance weights are bounded consists of assuming that the likelihood is bounded. Unfortunately, this bound is time-dependent.*

The optimal importance function suffers from two major drawbacks. It requires the ability to sample from $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$ and to evaluate, up to a proportionality constant, $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ where

$$p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)=\int p\left(\mathbf{y}_k|\,\mathbf{x}_k\right)p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}\right)d\mathbf{x}_k \tag{37}$$

It requires the evaluation of an integral which does not admit an analytical expression in the general case. Nevertheless, this evaluation is possible for the important class of models presented below.

Example 5. *Partial Gaussian State Space Models. Let us consider the following model:*

$$\mathbf{x}_k=f\left(\mathbf{x}_{k-1}\right)+\mathbf{v}_k,\ \mathbf{v}_k\sim\mathcal{N}\left(\mathbf{0},\mathbf{\Sigma_v}\right) \tag{38}$$

$$\mathbf{y}_k=\mathbf{C}\mathbf{x}_k+\mathbf{w}_k,\ \mathbf{w}_k\sim\mathcal{N}\left(\mathbf{0},\mathbf{\Sigma_w}\right) \tag{39}$$

where $f:\mathbb{R}^{n_\mathbf{x}}\to\mathbb{R}^{n_\mathbf{x}}$, $\mathbf{C}$ *is a real* $n_\mathbf{y}\times n_\mathbf{x}$ *matrix,* $\mathbf{v}_k$ *and* $\mathbf{w}_k$ *are two mutually independent i.i.d. Gaussian sequences with* $\mathbf{\Sigma_v}>0$ *et* $\mathbf{\Sigma_w}>0$. *Denoting*

$$\mathbf{\Sigma}^{-1} = \mathbf{\Sigma_v}^{-1}+\mathbf{C}^t\mathbf{\Sigma_w}^{-1}\mathbf{C} \tag{40}$$

$$\mathbf{m}_k = \mathbf{\Sigma}\left(\mathbf{\Sigma_v}^{-1}f\left(\mathbf{x}_{k-1}\right)+\mathbf{C}^t\mathbf{\Sigma_w}^{-1}\mathbf{y}_k\right) \tag{41}$$

one obtains

$$\mathbf{x}_k|\,\mathbf{x}_{k-1},\mathbf{y}_k\sim\mathcal{N}\left(\mathbf{m}_k,\mathbf{\Sigma}\right) \tag{42}$$

and

$$p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}\right)\propto\exp\left(-\frac{1}{2}\left(\mathbf{y}_k-\mathbf{C}f\left(\mathbf{x}_{k-1}\right)\right)^t\left(\mathbf{\Sigma_v}+\mathbf{C}\mathbf{\Sigma_w}\mathbf{C}^t\right)^{-1}\left(\mathbf{y}_k-\mathbf{C}f\left(\mathbf{x}_{k-1}\right)\right)\right) \tag{43}$$

For many other models, such evaluations are impossible. We now present suboptimal methods which allow approximation of the optimal importance function. The first proposed method is based on a second MC step.

MC approximation of the optimal importance function. We assume here that $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ cannot be evaluated analytically and that it is not possible to sample from $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$.

If the likelihood $p\left(\mathbf{y}_k|\,\mathbf{x}_k\right)<M_k$ is bounded then the ratio $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)/p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ is bounded. It is possible to sample from $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$ using the accept/reject procedure.

`Accept/Reject procedure`

1. `Sample` $\widetilde{\mathbf{x}}_k\sim p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ `and` $u\sim\mathcal{U}_{[0,1]}$.
2. `Accept` $\mathbf{x}_k^{(i)}=\widetilde{\mathbf{x}}_k$ `if` $u\leq p\left(\mathbf{y}_k|\,\widetilde{\mathbf{x}}_k\right)/M_k$; `otherwise return to 1`.

Unfortunately, this procedure requires a random number of iterations before obtaining a random sample distributed according to $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$. In the framework of online applications, this strategy must be avoided. Another more severe problem is that $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ is not evaluated.

A naive approach consists of using a second MC step based on Bayesian IS to sample from $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$ and/or to evaluate $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$. For each $\mathbf{x}_{k-1}^{(i)}$ $(i=1,...,N)$, one can sample N' i.i.d. random variables $\left\{\mathbf{x}_k^{(i),(j)};j=1,...,N'\right\}$ distributed according to $\mathbf{x}_k^{(i),(j)}\sim p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$. We obtain the following approximation of $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$:

$$\widehat{P}\left(d\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)=\sum_{j=1}^{N'}p_{(i)}^{(j)}\delta_{\mathbf{x}_k^{(i),(j)}}\left(d\mathbf{x}_k\right) \tag{44}$$

where

$$p_{(i)}^{(j)}=\frac{\frac{1}{N'}p\left(\mathbf{y}_k|\,\mathbf{x}_k^{(i),(j)}\right)}{\widehat{p}\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)} \tag{45}$$

$\widehat{p}\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$ being an estimate of $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}\right)$:

$$\widehat{p}\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)=\frac{1}{N'}\sum_{j=1}^{N'}p\left(\mathbf{y}_k|\,\mathbf{x}_k^{(i),(j)}\right) \tag{46}$$

This approximation is theoretically valid only if $N'\to+\infty$. Moreover, this solution, although simple, is computationally very expensive. Other MC methods based on MCMC methods have been proposed to simulate approximately from $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}^{(i)},\mathbf{y}_k\right)$ and/or to evaluate $p\left(\mathbf{y}_k|\,\mathbf{x}_{k-1}^{(i)}\right)$, see [12, 18, 41]. These iterative algorithms appear to be of limited interest in an on-line framework and there is a lack of theoretical convergence results.

In fact, the general framework of SIS allows us to consider other importance functions built so as to approximate analytically the optimal importance function. The advantages of this alternative approach are that it is computationally less expensive that MC methods and that the previous given convergence results on Bayesian IS are still valid. There is no general method to build suboptimal importance functions and it is necessary to build these on a case by case basis, dependent on the model studied. To this end, it is possible to base these developments on previous work on standard suboptimal filtering methods [6, 60].

Importance distribution obtained by local linearisation. A simple choice consists of selecting as the importance function $\pi\left(\mathbf{x}_k|\,\mathbf{x}_{k-1},\mathbf{y}_k\right)$ a parametric distribution $\pi\left(\mathbf{x}_k|\,\boldsymbol{\theta}\left(\mathbf{x}_{k-1},\mathbf{y}_k\right)\right)$, of finite-dimensional parameter $\boldsymbol{\theta}$ $(\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^{n_{\boldsymbol{\theta}}})$ determined by $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$, $\boldsymbol{\theta}:\mathbb{R}^{n_{\mathbf{x}}}\times\mathbb{R}^{n_{\mathbf{y}}}\to\boldsymbol{\Theta}$ being a deterministic mapping. Many strategies are possible. To illustrate such methods, we present here two original methods that result in a Gaussian importance function whose parameters are evaluated using local linearisations, i.e. which are dependent on the simulated trajectory $i=1,...,N$.

Local linearisation of the Markov state space model. We propose to linearise the model locally as in the Extended Kalman Filter. However, in our case, this linearisation is performed with the aim of obtaining an importance function and the algorithm obtained still converges asymptotically towards the optimal solution under the assumptions given previously.

Example 6. *Let us consider the following model*

$$\mathbf{x}_k = f\left(\mathbf{x}_{k-1}\right)+\mathbf{v}_k,\ \mathbf{v}_k\sim\mathcal{N}\left(\mathbf{0}_{n_{\mathbf{v}}\times1},\boldsymbol{\Sigma}_{\mathbf{v}}\right) \tag{47}$$

$$\mathbf{y}_k = g\left(\mathbf{x}_k\right)+\mathbf{w}_k,\ \mathbf{w}_k\sim\mathcal{N}\left(\mathbf{0}_{n_{\mathbf{w}}\times1},\boldsymbol{\Sigma}_{\mathbf{w}}\right) \tag{48}$$

where $f : \mathbb{R}^{n_{\mathbf{x}}} \to \mathbb{R}^{n_{\mathbf{x}}}$, $g : \mathbb{R}^{n_{\mathbf{x}}} \to \mathbb{R}^{n_{\mathbf{y}}}$ *is differentiable,* $\mathbf{v}_k$ *and* $\mathbf{w}_k$ *are two mutually independent i.i.d. sequences with* $\mathbf{\Sigma}_{\mathbf{v}} > 0$ *and* $\mathbf{\Sigma}_{\mathbf{w}} > 0$. *Performing an approximation up to the first order of the observation equation [6], we get*

$$\begin{aligned} \mathbf{y}_k &= g(\mathbf{x}_k) + \mathbf{w}_k \\ &\simeq g(f(\mathbf{x}_{k-1})) + \left.\frac{\partial g(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = f(\mathbf{x}_{k-1})} (\mathbf{x}_k - f(\mathbf{x}_{k-1})) + \mathbf{w}_k \end{aligned} \tag{49}$$

We have now defined a new model with a similar evolution equation to (47) but with a linear Gaussian observation equation (49), obtained by linearising $g(\mathbf{x}_k)$ *in* $f(\mathbf{x}_{k-1})$. *This model is not Markovian as (49) depends on* $\mathbf{x}_{k-1}$. *However, it is of the form (38)-(39) and one can perform similar calculations to obtain a Gaussian importance function* $\pi(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{y}_k) \sim \mathcal{N}(\mathbf{m}_k, \mathbf{\Sigma}_k)$ *with mean* $\mathbf{m}_k$ *and covariance* $\mathbf{\Sigma}_k$ *evaluated for each trajectory* $i = 1, ..., N$ *using the following formula:*

$$\mathbf{\Sigma}_k^{-1} = \mathbf{\Sigma}_{\mathbf{v}}^{-1} + \left[\left.\frac{\partial g(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = f(\mathbf{x}_{k-1})}\right]^t \mathbf{\Sigma}_{\mathbf{w}}^{-1} \left.\frac{\partial g(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = f(\mathbf{x}_{k-1})} \tag{50}$$

$$\mathbf{m}_k = \mathbf{\Sigma}_k \left(\mathbf{\Sigma}_{\mathbf{v}}^{-1} f(\mathbf{x}_{k-1}) + \left[\left.\frac{\partial g(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = f(\mathbf{x}_{k-1})}\right]^t \mathbf{\Sigma}_{\mathbf{w}}^{-1} \times \right. \tag{51}$$

$$\left. \times \left(\mathbf{y}_k - g(f(\mathbf{x}_{k-1})) + \left.\frac{\partial g(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = f(\mathbf{x}_{k-1})} f(\mathbf{x}_{k-1}) \right) \right) \tag{52}$$

The associated importance weight is evaluated using (32).

Local linearisation of the optimal importance function . We assume here that $l(\mathbf{x}_k) \triangleq \ln p(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{y}_k)$ is twice differentiable wrt $\mathbf{x}_k$ on $\mathbb{R}^{n_{\mathbf{x}}}$. We define:

$$l'(\mathbf{x}) \triangleq \left.\frac{\partial l(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = \mathbf{x}} \tag{53}$$

$$l''(\mathbf{x}) \triangleq \left.\frac{\partial^2 l(\mathbf{x}_k)}{\partial \mathbf{x}_k \partial \mathbf{x}_k^t}\right|_{\mathbf{x}_k = \mathbf{x}} \tag{54}$$

Using a second order Taylor expansion in $\mathbf{x}$, we get :

$$l(\mathbf{x}_k) \simeq l(\mathbf{x}) + [l'(\mathbf{x})]^t (\mathbf{x}_k - \mathbf{x}) + \frac{1}{2} (\mathbf{x}_k - \mathbf{x})^t l''(\mathbf{x}) (\mathbf{x}_k - \mathbf{x}) \tag{55}$$

The point $\mathbf{x}$ where we perform the expansion is arbitrary (but determined by a deterministic mapping of $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$). Under the additional assumption that $l''(\mathbf{x})$ is negative definite, which is true if $l(\mathbf{x}_k)$ is concave, then setting

$$\mathbf{\Sigma}(\mathbf{x}) = -l''(\mathbf{x})^{-1} \tag{56}$$

$$\mathbf{m}(\mathbf{x}) = \mathbf{\Sigma}(\mathbf{x}) l'(\mathbf{x}) \tag{57}$$

yields

$$\begin{aligned} &[l'(\mathbf{x})]^t (\mathbf{x}_k - \mathbf{x}) + \frac{1}{2} (\mathbf{x}_k - \mathbf{x})^t l''(\mathbf{x}) (\mathbf{x}_k - \mathbf{x}) \\ = \; & C - \frac{1}{2} (\mathbf{x}_k - \mathbf{x} - \mathbf{m}(\mathbf{x}))^t \mathbf{\Sigma}^{-1}(\mathbf{x}) (\mathbf{x}_k - \mathbf{x} - \mathbf{m}(\mathbf{x})) \end{aligned} \tag{58}$$

This suggests to adopt as importance function:

$$\pi\left(\mathbf{x}_k|\,\mathbf{x}_{k-1},\mathbf{y}_k\right)=\mathcal{N}\left(\mathbf{m}\left(\mathbf{x}\right)+\mathbf{x},\boldsymbol{\Sigma}\left(\mathbf{x}\right)\right) \tag{59}$$

If $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1},\mathbf{y}_k\right)$ is unimodal, it is judicious to adopt $\mathbf{x}$ as the mode of $p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1},\mathbf{y}_k\right)$, thus $\mathbf{m}\left(\mathbf{x}\right)=\mathbf{0}_{n_{\mathbf{x}}\times 1}$. The associated importance weight is evaluated using (32).

Example 7. *Linear Gaussian Dynamic/Observations according to a distribution from the exponential family. We assume that the evolution equation satisfies:*

$$\mathbf{x}_k=\mathbf{A}\mathbf{x}_{k-1}+\mathbf{v}_k \text{ where } \mathbf{v}_k\sim\mathcal{N}\left(\mathbf{0}_{n_{\mathbf{v}}\times 1},\boldsymbol{\Sigma}_{\mathbf{v}}\right) \tag{60}$$

where $\boldsymbol{\Sigma}_{\mathbf{v}}>0$ *and the observations are distributed according to a distribution from the exponential family, i.e.*

$$p\left(\mathbf{y}_k|\,\mathbf{x}_k\right)=\exp\left(\mathbf{y}_k^t\mathbf{C}\mathbf{x}_k-b\left(\mathbf{C}\mathbf{x}_k\right)+c\left(\mathbf{y}_k\right)\right) \tag{61}$$

where $\mathbf{C}$ *is a real* $n_{\mathbf{y}}\times n_{\mathbf{x}}$ *matrix,* $b:\mathbb{R}^{n_{\mathbf{y}}}\to\mathbb{R}$ *and* $c:\mathbb{R}^{n_{\mathbf{y}}}\to\mathbb{R}$*. These models have numerous applications and allow consideration of Poisson or binomial observations, see for example [60]. We have*

$$l\left(\mathbf{x}_k\right)=C+\mathbf{y}_k^t\mathbf{C}\mathbf{x}_k-b\left(\mathbf{C}\mathbf{x}_k\right)-\frac{1}{2}\left(\mathbf{x}_k-\mathbf{A}\mathbf{x}_{k-1}\right)^t\boldsymbol{\Sigma}_{\mathbf{v}}^{-1}\left(\mathbf{x}_k-\mathbf{A}\mathbf{x}_{k-1}\right) \tag{62}$$

This yields

$$\begin{aligned} l''\left(\mathbf{x}\right) &= -\left.\frac{\partial^2 b\left(\mathbf{C}\mathbf{x}_k\right)}{\partial\mathbf{x}_k\partial\mathbf{x}_k^t}\right|_{\mathbf{x}_k=\mathbf{x}}-\boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \\ &= -b''\left(\mathbf{x}\right)-\boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \end{aligned} \tag{63}$$

but $b''\left(\mathbf{x}\right)$ *is the covariance matrix of* $\mathbf{y}_k$ *for* $\mathbf{x}_k=\mathbf{x}$*, thus* $l''\left(\mathbf{x}\right)$ *is definite negative. One can determine the mode* $\mathbf{x}=\mathbf{x}^*$ *of this distribution by applying an iterative Newton-Raphson method initialised with* $\mathbf{x}_{(0)}=\mathbf{x}_{k-1}$*, which satisfies at iteration* j*:*

$$\mathbf{x}_{(j+1)}=\mathbf{x}_{(j)}-\left[l''\left(\mathbf{x}_{(j)}\right)\right]^{-1}l'\left(\mathbf{x}_{(j)}\right) \tag{64}$$

Remark 4. *This last method is close to the one developed independently by Shephard and Pitt [48] in a different framework. They propose a MCMC algorithm for off-line estimation of non-Gaussian measurements time series based on the Metropolis-Hastings algorithms. The proposal distribution of this algorithm is build in the case where* $l\left(\mathbf{x}_k\right)$ *is concave using a similar method* [3]*.*

We now present two simpler methods.

Prior importance function. A simple choice consists of selecting as importance function the prior distribution of the hidden Markov model. This is the choice made by Handschin et Mayne [25, 26] in their seminal work. This distribution has been recently adopted by Tanizaki *et al.* [56, 57]. In this case, we have

$$\pi\left(\mathbf{x}_k|\,\mathbf{x}_{0:k-1},\mathbf{y}_{0:k}\right)=p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}\right) \tag{65}$$

and

$$w_k^{(i)}=w_{k-1}^{(i)}p\left(\mathbf{y}_k|\,\mathbf{x}_k^{(i)}\right) \tag{66}$$

This method is often inefficient in simulations as the state space is explored without any knowledge of the observations. It is especially sensitive to outliers.

[3] In fact, all the methods developed in the literature to build "clever" proposal distributions for the Metropolis-Hastings (M-H) algorithms can be applied in a sequential framework and vice versa. But, while convergence of the M-H algorithm is ensured (under weak assumptions) when the number of iterations of the simulated Markov chain tends towards infinity, in the sequential framework, convergence of the algorithm is ensured (under weak assumptions) when the number N of simulated trajectories tends towards infinity.

Fixed importance function. A simpler choice consists of fixing an importance function independently of the simulated trajectories and from the observations. In this case, we have

$$\pi\left(\mathbf{x}_k \middle| \mathbf{x}_{0:k-1}, \mathbf{y}_{0:k}\right) = \pi\left(\mathbf{x}_k\right) \tag{67}$$

and (32) :

$$w_k^{(i)} = w_{k-1}^{(i)} p\left(\mathbf{y}_k \middle| \mathbf{x}_k^{(i)}\right) p\left(\mathbf{x}_k^{(i)} \middle| \mathbf{x}_{k-1}^{(i)}\right) / \pi\left(\mathbf{x}_k^{(i)}\right) \tag{68}$$

This is the choice adopted by Tanizaki *et al.* [54, 55] who presents this method as a stochastic alternative to the numerical integration method of Kitagawa [33]. The results obtained are rather poor as neither the dynamic of the model nor the observations are taken into account. It leads in most cases to unbounded importance weights.

3. Resampling

As it has been previously illustrated, the degeneracy of the algorithm based on SIS can not be avoided. In [15], a forgetting factor on the weights associated with the optimal importance function is introduced and, under stability and regularity assumptions on the Markov model, an interesting time-uniform convergence result is obtained as $N \to +\infty$. Practically, $N < +\infty$ and this regularization slows down but does not avoid degeneracy of the algorithm [16]. It is necessary to introduce another procedures.

The basic idea of resampling methods consists of eliminating the trajectories which have weak normalised importance weights and to multiply trajectories with strong importance weights. We adopt as a measure of degeneracy of the algorithm the effective sample size. This criterion, introduced by Liu [37, 39], is defined using the variances of the estimates of $I\left(f_k\right)$ respectively obtained using (imaginary) i.i.d. samples according to $\pi\left(\mathbf{x}_{0:k} \middle| \mathbf{y}_{0:k}\right)$ and an importance sampling method based on i.i.d. samples distributed according to $p\left(\mathbf{x}_{0:k} \middle| \mathbf{y}_{0:k}\right)$. For functions $f\left(\mathbf{x}_{0:k}\right)$ which vary slowly with $\mathbf{x}_{0:k}$, Liu shows that:

$$\frac{var_{\pi(\cdot|\mathbf{y}_{0:k})}\left[\widehat{I_N}\left(f_k\right)\right]}{var_{p(\cdot|\mathbf{y}_{0:k})}\left[\widehat{I_N}\left(f_k\right)\right]} \approx \left[1 + var_{\pi(\cdot|\mathbf{y}_{0:k})}\left(w^*\left(\mathbf{x}_{0:k}\right)\right)\right] \tag{69}$$

The effective sample size N_{eff} is thus defined as:

$$\begin{aligned} N_{eff} &= \frac{N}{1 + var_{\pi(\cdot|\mathbf{y}_{0:k})}\left(w^*\left(\mathbf{x}_{0:k}\right)\right)} \\ &= \frac{N}{\mathbb{E}_{\pi(\cdot|\mathbf{y}_{0:k})}\left[\left(w^*\left(\mathbf{x}_{0:k}\right)\right)^2\right]} \leq N \end{aligned} \tag{70}$$

One can not evaluate exactly N_{eff} but, owing to (22), an estimate $\widehat{N_{eff}}$ of N_{eff} is given by:

$$\widehat{N_{eff}} = \frac{N}{\frac{1}{N}\sum_{i=1}^N \left(\widehat{w}_k^{*(i)}\right)^2} = \frac{1}{\sum_{i=1}^N \left(\widetilde{w}_k^{(i)}\right)^2} \tag{71}$$

When $\widehat{N_{eff}}$ is below a fixed threshold N_{thres}, we use a resampling procedure. The most popular resampling scheme is the SIR algorithm (Sampling Importance Resampling) introduced by Rubin [46, 50]. This scheme is based on two steps: a first step is an IS step, the second step is a sampling step based on the obtained discrete distribution.

3.1. SIS/Resampling Monte Carlo filter. At time $k-1$, we have the following approximation (34) :

$$\widehat{P}\left(d\mathbf{x}_{0:k-1} \middle| \mathbf{y}_{0:k-1}\right) = \sum_{i=1}^N \widetilde{w}_{k-1}^{(i)} \delta_{\mathbf{x}_{0:k-1}^{(i)}}\left(d\mathbf{x}_{0:k-1}\right) \tag{72}$$

At time k, the modified Monte Carlo filter proceeds as follows.

Algorithm : SIS/Resampling Monte Carlo filter

1. Importance sampling

   - For $i = 1, ..., N$, sample $\widetilde{\mathbf{x}}_k^{(i)} \sim \pi(\mathbf{x}_k | \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{0:k})$ and $\widetilde{\mathbf{x}}_{0:k}^{(i)} \triangleq \left(\mathbf{x}_{0:k-1}^{(i)}, \widetilde{\mathbf{x}}_k^{(i)}\right)$.
   - For $i = 1, ..., N$, evaluate the importance weights up to a normalizing constant:

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p\left(\mathbf{y}_k | \widetilde{\mathbf{x}}_k^{(i)}\right) p\left(\widetilde{\mathbf{x}}_k^{(i)} \middle| \widetilde{\mathbf{x}}_{k-1}^{(i)}\right)}{\pi(\widetilde{\mathbf{x}}_k^{(i)} \middle| \widetilde{\mathbf{x}}_{0:k-1}^{(i)}, \mathbf{y}_{0:k})} \tag{73}$$

   - For $i = 1, ..., N$, normalise the importance weights:

$$\widetilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{j=1}^N w_k^{(j)}} \tag{74}$$

   - Evaluate $\widehat{N_{eff}}$ using (71).

2. Resampling

   If $\widehat{N_{eff}} \geq N_{thres}$

   - $\mathbf{x}_{0:k}^{(i)} = \widetilde{\mathbf{x}}_{0:k}^{(i)}$ for $i = 1, ..., N$.

   otherwise

   - For $i = 1, ..., N$, sample an index $j(i)$ distributed according to the discrete distribution with N elements satisfying $\Pr\{j(i) = l\} = \widetilde{w}_k^{(l)}$ for $l = 1, ..., N$.
   - For $i = 1, ..., N$, $\mathbf{x}_{0:k}^{(i)} = \widetilde{\mathbf{x}}_{0:k}^{j(i)}$ and $w_k^{(i)} = \frac{1}{N}$.

If $\widehat{N_{eff}} \geq N_{thres}$, the algorithm presented in section 2 is thus not modified. If $\widehat{N_{eff}} < N_{thres}$ the SIR algorithm is applied and we obtain the following approximation of the joint distribution:

$$\widehat{P}(d\mathbf{x}_{0:k} | \mathbf{y}_{0:k}) = \frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{x}_{0:k}^{(i)}}(d\mathbf{x}_{0:k}) \tag{75}$$

Remark 5. *In [41], other more interesting resampling schemes are presented which reduce the MC variation of the SIR.*

3.2. Implementation of the resampling procedure. If $\widehat{N_{eff}} < N_{thres}$, it is necessary to implement the algorithm to sample N random variates according to a discrete distribution with N elements. A straightforward application of the SIR procedure has a complexity in $O(N \ln N)$ [23]. This complexity is very important and, so as to reduce it, Beadle *et al.* [10] have recently proposed several *ad hoc* methods. In fact, it is possible to implement exactly the SIR procedure in $O(N)$ operations by noticing that it is possible to sample in $O(N)$ operations N i.i.d. variables distributed according to $\mathcal{U}_{[0,1]}$ and **ordered**, *i.e.* $u_1 \leq u_2 \leq \cdots \leq u_N$, using a classical algorithm [45, pp. 96].

Algorithm ([45, pp. 96])

- For $i = 1, ..., N$, sample $\widetilde{u}_i \sim \mathcal{U}_{[0,1]}$.

- $u_N = [\widetilde{u}_N]^{1/N}$.
- For $i = N-1, ..., 1$, $u_i = [\widetilde{u}_i]^{1/i} u_{i+1}$.

We deduce straightforwardly the algorithm to sample N i.i.d. samples according to the discrete distribution in $O(N)$ operations.

Remark 6. *This algorithm is also presented in [49] which attributed the idea of using this algorithm to Carpenter, Clifford and Fearnhead.*

3.3. Limitations of the resampling scheme. The resampling procedure decreases algorithmically the degeneracy problem but introduces practical and theoretical problems. From a practical point of view, the resampling scheme seriously limits the parallelisability of the algorithm. From a theoretical point of view, after one resampling step, the simulated trajectories are no longer statistically independent and so we lose the simple convergence results given previously. Moreover the trajectories $\left\{\widetilde{\mathbf{x}}_{0:k}^{(i)}, i = 1, ..., N\right\}$ which have high importance weights $\widetilde{w}_k^{(i)}$ are statistically selected many times. In (75), numerous trajectories $\mathbf{x}_{0:k}^{(i_1)}$ and $\mathbf{x}_{0:k}^{(i_2)}$ are in fact equal for $i_1 \neq i_2 \in [1, \ldots, N]$. There is a loss of "diversity". Recently, Berzuini *et al.* [12] have however established a central limit theorem for the estimate of $I(f_k)$ which is obtained when the SIR procedure is applied at each iteration.

Despite its drawbacks, the SIR algorithm is the basis of numerous works. The popular bootstrap filter of Gordon, Salmond et Smith [9, 10, 12, 23, 31], simultaneously developed by Kitagawa [29, 34, 35, 36], applies at each iteration a resampling step using $\pi\left(\mathbf{x}_k | \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k\right) = p\left(\mathbf{x}_k | \mathbf{x}_{k-1}^{(i)}\right)$, see also [52] for a similar method developed in the closely related field of Bayesian networks. To limit the loss of "diversity", many *ad hoc* procedures have been proposed. In [23], the trajectories are artificially perturbed after the resampling step. Another simple solution consists of building a semi-parametric approximation of $\widehat{P}(d\mathbf{x}) = \sum_{i=1}^N p_i K\left(\mathbf{x} - \mathbf{x}^{(i)}\right)$ before resampling [18, 24] but the choice of a "good" kernel $K(\cdot)$ is difficult. Higuchi [27, 28] proposes various heuristic procedures taken from the genetic algorithms literature to introduce such a diversity among samples. One can notice that, in fact, the SIR procedure has a similar mathematical structure to the selection step of genetic algorithms. Interesting extensions of the SIR algorithm have been recently developed by Shephard and Pitt [49].

4. Rao-Blackwellisation for Sequential Importance Sampling

We propose here to improve SIS using variance reduction methods designed to make the most of the model studied. Numerous methods have been developed so as to reduce the variance of MC estimates including antithetic sampling [25, 26] and control variates [2, 26]. We apply here the Rao-Blackwellisation method [4]. We show how it is possible to apply this method successfully to an important class of HMM and obtain hybrid filters where a part of the calculations is realized analytically and the other part using MC methods.

Let us assume that we can partition the state $\mathbf{x}_k$ as $\left(\mathbf{x}_k^1, \mathbf{x}_k^2\right)$ and denote $\mathbf{x}_{0:n}^j \triangleq \left(\mathbf{x}_0^j, \ldots, \mathbf{x}_n^j\right)$. We have:

$$I(f_n) = \frac{\int f_n(\mathbf{x}_{0:n}) p(\mathbf{y}_{0:n} | \mathbf{x}_{0:n}) p(\mathbf{x}_{0:n}) d\mathbf{x}_{0:n}}{\int p(\mathbf{y}_{0:n} | \mathbf{x}_{0:n}) p(\mathbf{x}_{0:n}) d\mathbf{x}_{0:n}} \tag{76}$$

$$= \frac{\int \left[\int f_n\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) p\left(\mathbf{y}_{0:n} | \mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) p\left(\mathbf{x}_{0:n}^2 | \mathbf{x}_{0:n}^1\right) d\mathbf{x}_{0:n}^2\right] p\left(\mathbf{x}_{0:n}^1\right) d\mathbf{x}_{0:n}^1}{\int \left[\int p(\mathbf{y}_{0:n} | \mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2) p(\mathbf{x}_{0:n}^2 | \mathbf{x}_{0:n}^1) d\mathbf{x}_{0:n}^2\right] p(\mathbf{x}_{0:n}^1) d\mathbf{x}_{0:n}^1}$$

$$= \frac{\int g(\mathbf{x}_{0:n}^1) p\left(\mathbf{x}_{0:n}^1\right) d\mathbf{x}_{0:n}^1}{\int p(\mathbf{y}_{0:n} | \mathbf{x}_{0:n}^1) p(\mathbf{x}_{0:n}^1) d\mathbf{x}_{0:n}^1}$$

where

$$g(\mathbf{x}_{0:n}^1) \triangleq \int f_n\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) p\left(\mathbf{y}_{0:n} \middle| \mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) p\left(\mathbf{x}_{0:n}^2 \middle| \mathbf{x}_{0:n}^1\right) d\mathbf{x}_{0:n}^2 \tag{77}$$

Under the assumption that, conditional upon a realization of $\mathbf{x}_{0:n}^1$, $g(\mathbf{x}_{0:n}^1)$ and $p\left(\mathbf{y}_{0:n} \middle| \mathbf{x}_{0:n}^1\right)$ can be evaluated analytically, two estimates of $I(f_n)$ based on IS are possible. The first "classical" one is obtained using as importance distribution $\pi\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}\right)$:

$$\widehat{I_N}(f_n) = \frac{\widehat{N_N}(f_n)}{\widehat{D_N}(f_n)} = \frac{\sum_{i=1}^N f_n\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)}\right) w^*\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)}\right)}{\sum_{i=1}^N w^*\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)}\right)} \tag{78}$$

where

$$w^*\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)}\right) = \frac{p\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)} \middle| \mathbf{y}_{0:n}\right)}{\pi\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^{2,(i)} \middle| \mathbf{y}_{0:n}\right)} \tag{79}$$

The second "Rao-Blackwellised" estimate $\widehat{\widehat{I_N}}(f_n)$ is obtained by integrating out analytically $\mathbf{x}_{0:n}^2$ and using as importance distribution

$$\pi\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right) = \int \pi\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}\right) d\mathbf{x}_{0:n}^2 \tag{80}$$

The estimate is given by:

$$\widehat{\widehat{I_N}}(f_n) = \frac{\widehat{\widehat{N_N}}(f_n)}{\widehat{\widehat{D_N}}(f_n)} = \frac{\sum_{i=1}^N \mathbb{E}_{p\left(\mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}, \mathbf{x}_{0:n}^{1,(i)}\right)}\left(f_n\left(\mathbf{x}_{0:n}^{1,(i)}, \mathbf{x}_{0:n}^2\right)\right) w^*\left(\mathbf{x}_{0:n}^{1,(i)}\right)}{\sum_{i=1}^N w^*\left(\mathbf{x}_{0:n}^{1,(i)}\right)} \tag{81}$$

where

$$w^*\left(\mathbf{x}_{0:n}^{1,(i)}\right) = \frac{p\left(\mathbf{x}_{0:n}^{1,(i)} \middle| \mathbf{y}_{0:n}\right)}{\pi\left(\mathbf{x}_{0:n}^{1,(i)} \middle| \mathbf{y}_{0:n}\right)} \tag{82}$$

The following proposition shows that if one can integrate analytically one of the components then the variance of the obtained estimate is weaker than the one of the crude estimate.

Proposition 8. *The variances of the importance weights, the numerator and the denominator, obtained by Rao-Blackwellisation, are smaller than those obtained using a crude Monte Carlo method:*

$$var_{\pi\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)}\left[w^*\left(\mathbf{x}_{0:n}^1\right)\right] \leq var_{\pi\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}\right)}\left[w^*\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right)\right] \tag{83}$$

and

$$var_{\pi\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)}\left(\widehat{\widehat{N_N}}(f_n)\right) \leq var_{\pi\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}\right)}\left(\widehat{N_N}(f_n)\right) \tag{84}$$

$$var_{\pi\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)}\left(\widehat{\widehat{D_N}}(f_n)\right) \leq var_{\pi\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}\right)}\left(\widehat{D_N}(f_n)\right) \tag{85}$$

The proof is straightforward [19]. We can use this simple result to estimate the marginal distribution $p\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)$ but also:

- If $f_n\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) = f_n\left(\mathbf{x}_{0:n}^1\right)$ then $g(\mathbf{x}_{0:n}^1) = f_n\left(\mathbf{x}_{0:n}^1\right) p\left(\mathbf{y}_{0:n} \middle| \mathbf{x}_{0:n}^1\right)$ and

$$\widehat{\widetilde{I}}_N(f_n) = \frac{\sum_{i=1}^N f_n\left(\mathbf{x}_{0:n}^{1,(i)}\right) w\left(\mathbf{x}_{0:n}^{1,(i)}\right)}{\sum_{i=1}^N w\left(\mathbf{x}_{0:n}^{1,(i)}\right)} \tag{86}$$

- If $f_n\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right) = f_n\left(\mathbf{x}_{0:n}^2\right)$ then $g(\mathbf{x}_{0:n}^1) = \mathbb{E}_{p\left(\mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}, \mathbf{x}_{0:n}^1\right)}\left(f_n\left(\mathbf{x}_{0:n}^2\right)\right) p\left(\mathbf{y}_{0:n} \middle| \mathbf{x}_{0:n}^1\right)$ and

$$\widehat{\widetilde{I}}_N(f_n) = \frac{\sum_{i=1}^N \mathbb{E}_{p\left(\mathbf{x}_{0:n}^2 \middle| \mathbf{y}_{0:n}, \mathbf{x}_{0:n}^{1,(i)}\right)}\left(f_n\left(\mathbf{x}_{0:n}^2\right)\right) w\left(\mathbf{x}_{0:n}^{1,(i)}\right)}{\sum_{i=1}^N w\left(\mathbf{x}_{0:n}^{1,(i)}\right)} \tag{87}$$

In all cases, it is possible to use the MC methods developed in the previous sections to $\mathbf{x}_{0:n}^1$. Nevertheless, even if the observations $\mathbf{y}_{0:n}$ are independent conditional upon $\left(\mathbf{x}_{0:n}^1, \mathbf{x}_{0:n}^2\right)$, they are generally no longer independent conditional upon the single process $\mathbf{x}_{0:n}^1$. The modifications are straightforward. We obtain for the optimal importance function $p\left(\mathbf{x}_k^1 \middle| \mathbf{y}_{0:k}, \mathbf{x}_{0:k-1}^1\right)$ and its associated importance weight $p\left(\mathbf{y}_k \middle| \mathbf{y}_{0:k-1}, \mathbf{x}_{0:k-1}^1\right)$. We now present two important applications of this general method.

Example 9. *Conditionally linear Gaussian state space model*

Let us consider the following model

$$p\left(\mathbf{x}_k^1 \middle| \mathbf{x}_{k-1}^1\right) \tag{88}$$

$$\mathbf{x}_k^2 = \mathbf{A}_k\left(\mathbf{x}_k^1\right)\mathbf{x}_{k-1}^2 + \mathbf{B}_k\left(\mathbf{x}_k^1\right)\mathbf{v}_k \tag{89}$$

$$\mathbf{y}_k = \mathbf{C}_k\left(\mathbf{x}_k^1\right)\mathbf{x}_k^2 + \mathbf{D}_k\left(\mathbf{x}_k^1\right)\mathbf{w}_k \tag{90}$$

where $\mathbf{x}_k^1$ is a Markov process, $\mathbf{v}_k \sim \mathcal{N}\left(\mathbf{0}_{n_{\mathbf{v}}\times 1}, \mathbf{I}_{n_{\mathbf{v}}}\right)$ and $\mathbf{w}_k \sim \mathcal{N}\left(\mathbf{0}_{n_{\mathbf{w}}\times 1}, \mathbf{I}_{n_{\mathbf{w}}}\right)$. One wants to estimate $p\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)$, $\mathbb{E}\left(f\left(\mathbf{x}_n^1\right) \middle| \mathbf{y}_{0:n}\right)$, $\mathbb{E}\left(\mathbf{x}_n^2 \middle| \mathbf{y}_{0:n}\right)$ and $\mathbb{E}\left(\mathbf{x}_n^2\left(\mathbf{x}_n^2\right)^t \middle| \mathbf{y}_{0:n}\right)$. It is possible to use a MC filter based on Rao-Blackwellisation. Indeed, conditional upon $\mathbf{x}_{0:n}^1$, $\mathbf{x}_{0:n}^2$ is a linear Gaussian state space model and the integrations required by the Rao-Blackwellisation method can be realized using the Kalman filter.

Akashi and Kumamoto [1, 4, 58] introduced this algorithm under the name of RSA (Random Sampling Algorithm) in the particular case where $\mathbf{x}_k^1$ is a homogeneous finite state-space Markov chain [4]. In this case, they adopted the optimal importance function $p\left(\mathbf{x}_k^1 \middle| \mathbf{y}_{0:k}, \mathbf{x}_{0:k-1}^1\right)$. Indeed, it is possible to sample from this discrete distribution and to evaluate the importance weight $p\left(\mathbf{y}_k \middle| \mathbf{y}_{0:k}, \mathbf{x}_{0:k-1}^1\right)$ using the Kalman filter [4]. Similar developments have been proposed by Svetnik *et al.* [53]. The algorithm for blind deconvolution recently proposed by Liu *et al.* [38] is also a particular case of this method where $\mathbf{x}_k^2 = \mathbf{h}$ is a time-invariant channel of Gaussian prior distribution [5]. Using the Rao-Blackwellisation method in this framework is particularly attractive as, while $\mathbf{x}_k$ has some continuous components, we restrict ourselves to the exploration of a discrete state space.

Example 10. *Finite State-Space HMM*

Let us consider the following model

$$p\left(\mathbf{x}_k^1 \middle| \mathbf{x}_{k-1}^1\right)$$
$$p\left(\mathbf{x}_k^2 \middle| \mathbf{x}_k^1, \mathbf{x}_{k-1}^2\right)$$
$$p\left(\mathbf{y}_k \middle| \mathbf{x}_k^1, \mathbf{x}_k^2\right)$$

[4] Akashi and Kumamoto made the connections with the work of Handschin and Mayne in [2].

[5] In this framework, the extension to a time-varying channel $\mathbf{h}_k$ modeled by a linear Gaussian state-space model is straightforward.

where $\mathbf{x}_k^1$ is a Markov process and $\mathbf{x}_k^2$ is a finite state-space Markov chain whose parameters at time k depend on $\mathbf{x}_k^1$. We want to estimate $p\left(\mathbf{x}_{0:n}^1 \middle| \mathbf{y}_{0:n}\right)$, $\mathbb{E}\left(f\left(\mathbf{x}_n^1\right) \middle| \mathbf{y}_{0:n}\right)$ and $\mathbb{E}\left(f\left(\mathbf{x}_n^2\right) \middle| \mathbf{y}_{0:n}\right)$. It is possible to use a "Rao-Blackwellised" MC filter. Indeed, conditional upon $\mathbf{x}_{0:n}^1$, $\mathbf{x}_{0:n}^2$ is a finite state-space Markov chain of known parameters and thus the integrations require by the Rao-Blackwellisation method can be done analytically [6].

5. Prediction, smoothing and likelihood

The estimate of the joint distribution $p(\mathbf{x}_{0:k}|\mathbf{y}_{0:k})$ based on SIS, in practice coupled with a resampling procedure to limit the degeneracy, is at any time k of the following form:

$$\widehat{P}(d\mathbf{x}_{0:k}|\mathbf{y}_{0:k}) = \sum_{i=1}^{N} \widetilde{w}_k^{(i)} \delta_{\mathbf{x}_{0:k}^{(i)}}(d\mathbf{x}_{0:k}) \tag{91}$$

We show here how it is possible to obtain based on this distribution some approximations of the prediction and smoothing distributions as well as the likelihood.

5.1. Prediction. Based on the approximation of the filtering distribution $\widehat{P}(d\mathbf{x}_k|\mathbf{y}_{0:k})$, we want to estimate the p step-ahead prediction distribution, $p \geq 2 \in \mathbb{N}^*$, given by:

$$p(\mathbf{x}_{k+p}|\mathbf{y}_{0:k}) = \int p(\mathbf{x}_k|\mathbf{y}_{0:k}) \left[\prod_{j=k+1}^{k+p} p(\mathbf{x}_j|\mathbf{x}_{j-1})\right] d\mathbf{x}_{k:k+p-1} \tag{92}$$

Replacing $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ in (92) by its approximation obtained from (91), we obtain:

$$\sum_{i=1}^{N} \widetilde{w}_k^{(i)} \int p\left(\mathbf{x}_{k+1}|\mathbf{x}_k^{(i)}\right) \prod_{j=k+2}^{k+p} p(\mathbf{x}_j|\mathbf{x}_{j-1})\, d\mathbf{x}_{k+1:k+p-1} \tag{93}$$

To evaluate these integrals, it is sufficient to extend the trajectories $\mathbf{x}_{0:k}^{(i)}$ using the evolution equation.

Algorithm. p **step-ahead prediction**

- `For` $j = 1$ `to` p
  - `For` $i = 1, ..., N$, `sample` $\mathbf{x}_{k+j}^{(i)} \sim p\left(\mathbf{x}_{k+j}|\mathbf{x}_{k+j-1}^{(i)}\right)$ `and` $\mathbf{x}_{0:k+j}^{(i)} \triangleq \left(\mathbf{x}_{0:k+j-1}^{(i)}, \mathbf{x}_{k+j}^{(i)}\right)$.

We obtain random samples $\left\{\mathbf{x}_{0:k+p}^{(i)}; i = 1, ..., N\right\}$. An estimate of $\widehat{P}(d\mathbf{x}_{0:k+p}|\mathbf{y}_{0:k})$ is given by

$$\widehat{P}(d\mathbf{x}_{0:k+p}|\mathbf{y}_{0:k}) = \sum_{i=1}^{N} \widetilde{w}_k^{(i)} \delta_{\mathbf{x}_{0:k+p}^{(i)}}(d\mathbf{x}_{0:k+p})$$

Thus

$$\widehat{P}(d\mathbf{x}_{k+p}|\mathbf{y}_{0:k}) = \sum_{i=1}^{N} \widetilde{w}_k^{(i)} \delta_{\mathbf{x}_{k+p}^{(i)}}(d\mathbf{x}_{k+p}) \tag{94}$$

5.2. Fixed-Lag smoothing. We want to estimate the fixed-lag smoothing distribution $p(\mathbf{x}_k|\mathbf{y}_{0:k+p})$, $p \in \mathbb{N}^*$ being the length of the lag. At time $k+p$, the MC filter yields the following approximation of $p(\mathbf{x}_{0:k+p}|\mathbf{y}_{0:k+p})$:

$$\widehat{P}(d\mathbf{x}_{0:k+p}|\mathbf{y}_{0:k+p}) = \sum_{i=1}^{N} \widetilde{w}_{k+p}^{(i)} \delta_{\mathbf{x}_{0:k+p}^{(i)}}(d\mathbf{x}_{0:k+p}) \tag{95}$$

By marginalising, we obtain an estimate of the fixed-lag smoothing distribution:

$$\widehat{P}\left(d\mathbf{x}_k|\,\mathbf{y}_{0:k+p}\right)=\sum_{i=1}^{N}\widetilde{w}_{k+p}^{(i)}\delta_{\mathbf{x}_k^{(i)}}\left(d\mathbf{x}_k\right) \tag{96}$$

When p is high, such an approximation will generally perform poorly.

Remark 7. *To estimate $p\left(\mathbf{x}_k|\,\mathbf{y}_{0:k+p}\right)$, it would be better to use an importance function of the form $\pi\left(\mathbf{x}_k|\,\mathbf{x}_{0:k-1},\mathbf{y}_{0:k+p}\right)$, see Remark 1. Under a straightforward modification of the criterion proposed previously, the optimal importance function and the associated importance weight are respectively equal to $p\left(\mathbf{x}_k|\,\mathbf{x}_{0:k-1},\mathbf{y}_{0:k+p}\right)$ and $p\left(\mathbf{y}_{k+p}|\,\mathbf{x}_{0:k-1},\mathbf{y}_{k:k+p-1}\right)$. Usually, it is difficult to sample from $p\left(\mathbf{x}_k|\,\mathbf{x}_{0:k-1},\mathbf{y}_{0:k+p}\right)$ and impossible to evaluate analytically $p\left(\mathbf{y}_{k+p}|\,\mathbf{x}_{0:k-1},\mathbf{y}_{k:k+p-1}\right)$. It is possible to build suboptimal importance functions based for example on extended Kalman smoother techniques but it remains to evaluate the term $p\left(\mathbf{y}_{k+p}|\,\mathbf{x}_{0:k},\mathbf{y}_{k:k+p-1}\right)$ which occurs in the expression of the importance weight. It is possible to evaluate this term using MC integration.*

5.3. Fixed-interval smoothing. Given $\mathbf{y}_{0:n}$, we want to estimate $p\left(\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)$ for any $k=0,...,n$. At time n, the filtering algorithm yields the following approximation of $p\left(\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}\right)$:

$$\widehat{P}\left(d\mathbf{x}_{0:n}|\,\mathbf{y}_{0:n}\right)=\sum_{i=1}^{N}\widetilde{w}_n^{(i)}\delta_{\mathbf{x}_{0:n}^{(i)}}\left(d\mathbf{x}_{0:n}\right) \tag{97}$$

Thus one can theoretically obtain $p\left(\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)$ for any k by marginalising this distribution. Practically, this method cannot be used as soon as $(n-k)$ is significant as the degeneracy problem requires use of a resampling algorithm. At time n, the simulated trajectories $\left\{\mathbf{x}_{0:n}^{(i)};i=1,...,N\right\}$ have been usually resampled many times: there are thus only a few distinct trajectories at times k for $k\ll n$ and the above approximation of $p\left(\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)$ is bad. This problem is even more severe for the bootstrap filter where one resamples at each time instant.

It is necessary to develop an alternative algorithm. We propose an original algorithm to solve this problem. This algorithm is based on the following formula [8, 33]:

$$p\left(\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)=p\left(\mathbf{x}_k|\,\mathbf{y}_{0:k}\right)\int\frac{p\left(\mathbf{x}_{k+1}|\,\mathbf{y}_{0:n}\right)p\left(\mathbf{x}_{k+1}|\,\mathbf{x}_k\right)}{p\left(\mathbf{x}_{k+1}|\,\mathbf{y}_{0:k}\right)}d\mathbf{x}_{k+1} \tag{98}$$

We seek here an approximation of the fixed-interval smoothing distribution with the following form:

$$\widehat{P}\left(d\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)\triangleq\sum_{i=1}^{N}\widetilde{w}_{k|n}^{(i)}\delta_{\mathbf{x}_k^{(i)}}\left(d\mathbf{x}_k\right) \tag{99}$$

i.e. $\widehat{P}\left(d\mathbf{x}_k|\,\mathbf{y}_{0:n}\right)$ has the same support $\left\{\mathbf{x}_k^{(i)};i=1,\ldots,N\right\}$ as the filtering distribution $\widehat{P}\left(d\mathbf{x}_k|\,\mathbf{y}_{0:k}\right)$ but the weights are different. An algorithm to obtain these weights $\left\{\widetilde{w}_{k|n}^{(i)};i=1,\ldots,N\right\}$ is the following.

Algorithm. Fixed-interval smoothing.

1. `Initialisation at time` $k=n$.
   - `For` $i=1,...,N$, $\widetilde{w}_{n|n}^{(i)}=\widetilde{w}_n^{(i)}$.
2. `For` $k=n-1,...,0$.

- For $i = 1, ..., N$, evaluate the importance weight

$$\widetilde{w}_{k|n}^{(i)} = \sum_{j=1}^{N} \widetilde{w}_{k+1|n}^{(j)} \frac{\widetilde{w}_k^{(i)} p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k^{(i)}\right)}{\left[\sum_{l=1}^{N} \widetilde{w}_k^{(l)} p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k^{(l)}\right)\right]} \tag{100}$$

This algorithm is obtained by the following argument. Replacing $p(\mathbf{x}_{k+1}|\mathbf{y}_{0:n})$ by its approximation (99) yields

$$\int \frac{p(\mathbf{x}_{k+1}|\mathbf{y}_{0:n})\, p(\mathbf{x}_{k+1}|\mathbf{x}_k)}{p(\mathbf{x}_{k+1}|\mathbf{y}_{0:k})} d\mathbf{x}_{k+1} \simeq \sum_{i=1}^{N} \widetilde{w}_{k+1|n}^{(i)} \frac{p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{x}_k\right)}{p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{y}_{0:k}\right)} \tag{101}$$

where, owing to (91), $p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{y}_{0:k}\right)$ can be approximated by

$$\begin{aligned} p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{y}_{0:k}\right) &= \int p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{x}_k\right) p(\mathbf{x}_k|\mathbf{y}_{0:k})\, d\mathbf{x}_k \\ &\simeq \sum_{j=1}^{N} \widetilde{w}_k^{(j)} p\left(\mathbf{x}_{k+1}^{(i)} \middle| \mathbf{x}_k^{(j)}\right) \end{aligned} \tag{102}$$

An approximation $\widehat{P}(d\mathbf{x}_k|\mathbf{y}_{0:n})$ of $p(\mathbf{x}_k|\mathbf{y}_{0:n})$ is thus

$$\begin{aligned} & \widehat{P}(d\mathbf{x}_k|\mathbf{y}_{0:n}) \\ = & \left[\sum_{i=1}^{N} \widetilde{w}_k^{(i)} \delta_{\mathbf{x}_k^{(i)}}(d\mathbf{x}_k)\right] \sum_{j=1}^{N} \widetilde{w}_{k+1|n}^{(j)} \frac{p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k\right)}{\left[\sum_{l=1}^{N} \widetilde{w}_k^{(l)} p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k^{(l)}\right)\right]} \\ = & \sum_{i=1}^{N} \widetilde{w}_k^{(i)} \left[\sum_{j=1}^{N} \widetilde{w}_{k+1|n}^{(j)} \frac{p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k^{(i)}\right)}{\left[\sum_{l=1}^{N} \widetilde{w}_k^{(l)} p\left(\mathbf{x}_{k+1}^{(j)} \middle| \mathbf{x}_k^{(l)}\right)\right]}\right] \delta_{\mathbf{x}_k^{(i)}}(d\mathbf{x}_k) \\ \triangleq & \sum_{i=1}^{N} \widetilde{w}_{k|n}^{(i)} \delta_{\mathbf{x}_k^{(i)}}(d\mathbf{x}_k) \end{aligned} \tag{103}$$

The algorithm follows.

This algorithm requires storage of the marginal distributions $\widehat{P}(d\mathbf{x}_k|\mathbf{y}_{0:k})$ (weights and supports) for any $k = 0, ..., n$. The memory requirement is $O(nN)$. Its complexity is $O(nN^2)$, which is quite important as $N \gg 1$. However this complexity is a little lower than the one of the previous developed algorithms of Kitagawa [35, 36] and Tanizaki [56, 57] as it does not require any new simulation step.

5.4. Likelihood. In some applications, in particular for model choice [33, 36], we may wish to estimate the likelihood of the data $p(\mathbf{y}_{0:n})$. A simple estimate of the likelihood is given, using to (15) and (18), by

$$\widehat{p}(\mathbf{y}_{0:n}) = \frac{1}{N} \sum_{j=1}^{N} w_n^{(j)} \tag{104}$$

In practice, the introduction of resampling steps makes this approach impossible. We will use an alternative decomposition of the likelihood:

$$p(\mathbf{y}_{0:n}) = p(\mathbf{y}_0) \prod_{k=1}^{n} p(\mathbf{y}_k|\mathbf{y}_{0:k-1}) \tag{105}$$

where:

$$p(\mathbf{y}_k|\mathbf{y}_{0:k-1}) = \int p(\mathbf{y}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{y}_{0:k-1})\, d\mathbf{x}_k \quad (106)$$

$$= \int p(\mathbf{y}_k|\mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})\, d\mathbf{x}_{k-1} \quad (107)$$

Using (106), an estimate of this quantity is given by

$$\widehat{p}(\mathbf{y}_k|\mathbf{y}_{0:k-1}) = \sum_{i=1}^{N} p\left(\mathbf{y}_k|\widetilde{\mathbf{x}}_k^{(i)}\right) \widetilde{w}_{k-1}^{(i)} \quad (108)$$

where the samples $\left\{\widetilde{\mathbf{x}}_k^{(i)}; i=1,\ldots,N\right\}$ are obtained using a one-step ahead prediction based on the approximation $\widehat{P}(d\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})$ of $p(\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})$. Using expression (107), it is possible to avoid a MC integration if we know analytically $p\left(\mathbf{y}_k|\mathbf{x}_{k-1}^{(i)}\right)$:

$$\widehat{p}(\mathbf{y}_k|\mathbf{y}_{0:k-1}) = \sum_{i=1}^{N} p\left(\mathbf{y}_k|\mathbf{x}_{k-1}^{(i)}\right) \widetilde{w}_{k-1}^{(i)} \quad (109)$$

6. Simulations

In this section, we apply the methods developed previously to a linear Gaussian state space model and to a classical nonlinear model. We make for these two models $M = 100$ simulations of length $n = 500$ and we evaluate the empirical standard deviation for the filtering estimates $\mathbf{x}_{k|k} = \mathbb{E}[\mathbf{x}_k|\mathbf{y}_{0:k}]$ obtained by the MC methods:

$$\sqrt{VAR}\left(\mathbf{x}_{k|l}\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{M}\sum_{j=1}^{M}\left(\mathbf{x}_{k|l}^{j} - \mathbf{x}_k^j\right)^2\right)^{1/2}$$

where:

- $\mathbf{x}_k^j$ is the simulated state for the j^{th} simulation, $j = 1, ..., M$.
- $\mathbf{x}_{k|l}^j \triangleq \sum_{i=1}^{N} \widetilde{w}_{k|l}^{(i)} \mathbf{x}_k^{j,(i)}$ is the MC estimate of $\mathbb{E}[\mathbf{x}_k|\mathbf{y}_{0:l}]$ for the j^{th} test signal and $\mathbf{x}_k^{j,(i)}$ is the i^{th} simulated trajectory, $i = 1, ..., N$, associated with the signal j. (We denote $\widetilde{w}_{k|k}^{(i)} \triangleq \widetilde{w}_k^{(i)}$.)

These calculations have been realized for $N = 100$, 250, 500, 1000, 2500 and 5000. The implemented filtering algorithms are the bootstrap filter, the SIS with the prior importance function and the SIS with the optimal or a suboptimal importance function. The fixed-interval smoothers associated with these SIS filters are then computed.

For the SIS-based algorithms, the SIR procedure has been used when $\widehat{N_{eff}} < N_{thres} = N/3$. We state the percentage of iterations where the SIR step is used for each importance function.

6.1. Linear Gaussian model. Let us consider the following model

$$x_k = x_{k-1} + v_k \quad (110)$$

$$y_k = x_k + w_k \quad (111)$$

where $x_0 \sim \mathcal{N}(0,1)$, v_k and w_k are white Gaussian noises mutually independent, $v_k \sim \mathcal{N}\left(0, \sigma_v^2\right)$ and $w_k \sim \mathcal{N}\left(0, \sigma_w^2\right)$ with $\sigma_v^2 = \sigma_w^2 = 1$. For this model, the optimal filter is the Kalman filter [6].

Optimal importance function. The optimal importance function is

$$x_k \,|\, x_{k-1}, y_k \sim \mathcal{N}\left(m_k, \sigma_k^2\right) \tag{112}$$

where

$$\begin{aligned} [\sigma_k]^{-2} &= \sigma_w^{-2} + \sigma_v^{-2} && (113)\\ m_k &= \sigma_k^2 \left(\frac{x_{k-1}}{\sigma_v^2} + \frac{y_k}{\sigma_w^2}\right) && (114) \end{aligned}$$

and the associated importance weight is equal to:

$$p\left(y_k \,|\, x_{k-1}\right) \propto \exp\left(-\frac{1}{2}\frac{\left(y_k - x_{k-1}\right)^2}{\left(\sigma_v^2 + \sigma_w^2\right)}\right) \tag{115}$$

Results. For the Kalman filter, we obtain $\sqrt{VAR}\left(\mathbf{x}_{k|k}\right) = 0.79$. For the different MC filters, the results are presented in Tab. 1 and Tab. 2.

$\sqrt{VAR}\left(\mathbf{x}_{k\|k}\right)$	bootstrap	prior dist.	optimal dist.
$N = 100$	0.80	0.86	0.83
$N = 250$	0.81	0.81	0.80
$N = 500$	0.79	0.80	0.79
$N = 1000$	0.79	0.79	0.79
$N = 2500$	0.79	0.79	0.79
$N = 5000$	0.79	0.79	0.79

Table 1: MC filters: linear Gaussian model

Percentage SIR	prior dist.	optimal dist.
$N = 100$	40	16
$N = 250$	23	10
$N = 500$	20	8
$N = 1000$	15	6
$N = 2500$	13	5
$N = 5000$	11	4

Table 2: Percentage of SIR steps: linear Gaussian model

With $N = 500$ trajectories, the estimates obtained using MC methods are similar to those obtained by Kalman. The SIS algorithms have similar performances to the bootstrap filter for a smaller computational cost. The most interesting algorithm is based on the optimal importance function which limits seriously the number of resampling steps.

6.2. Nonlinear series. We consider here the following nonlinear reference model [7, 23, 35, 56]:

$$\begin{aligned} x_k &= f\left(x_{k-1}\right) + v_k && (116)\\ &= \frac{1}{2}x_{k-1} + 25\frac{x_{k-1}}{1 + \left(x_{k-1}\right)^2} + 8\cos\left(1.2k\right) + v_k \\ y_k &= g\left(x_k\right) + w_k && (117)\\ &= \frac{\left(x_k\right)^2}{20} + w_k \end{aligned}$$

where $x_0 \sim \mathcal{N}(0,5)$, v_k and w_k are mutually independent white Gaussian noises, $v_k \sim \mathcal{N}\left(0,\sigma_v^2\right)$ and $w_k \sim \mathcal{N}\left(0,\sigma_w^2\right)$ with $\sigma_v^2 = 10$ and $\sigma_w^2 = 1$. In this case, it is not possible to evaluate analytically $p(y_k|x_{k-1})$ or to sample simply from $p(x_k|x_{k-1},y_k)$. We propose to apply the method described in 2.5 which consists of linearising locally the observation equation.

Importance function obtained by local linearisation. We get

$$\begin{aligned} y_k &\simeq g(f(x_{k-1})) + \left.\frac{\partial g(x_k)}{\partial x_k}\right|_{x_k=f(x_{k-1})} (x_k - f(x_{k-1})) + w_k \\ &= \frac{f^2(x_{k-1})}{20} + \frac{f(x_{k-1})}{10}(x_k - f(x_{k-1})) + w_k \\ &= -\frac{f^2(x_{k-1})}{20} + \frac{f(x_{k-1})}{10} x_k + w_k \end{aligned} \tag{118}$$

Then we obtain the linearised importance function $\pi(x_k|x_{k-1},y_k) = \mathcal{N}\left(x_k; m_k, (\sigma_k)^2\right)$ where

$$(\sigma_k)^{-2} = \sigma_v^{-2} + \sigma_w^{-2}\frac{f^2(x_{k-1})}{100} \tag{119}$$

and

$$m_k = (\sigma_k)^2\left[\sigma_v^{-2} f(x_{k-1}) + \sigma_w^{-2}\frac{f(x_{k-1})}{10}\left(y_k + \frac{f^2(x_{k-1})}{20}\right)\right] \tag{120}$$

Results. In this case, it is not possible to estimate the optimal filter. For the MC filters, the results are displayed in Tab. 3. The average percentages of SIR steps are presented in Tab. 4.

$\sqrt{VAR}(\mathbf{x}_{k\|k})$	bootstrap	prior dist.	linearised dist.
$N = 100$	5.67	6.01	5.54
$N = 250$	5.32	5.65	5.46
$N = 500$	5.27	5.59	5.23
$N = 1000$	5.11	5.36	5.05
$N = 2500$	5.09	5.14	5.02
$N = 5000$	5.04	5.07	5.01

Table 3: MC filters: nonlinear time series

Percentage SIR	prior dist.	linearised dist.
$N = 100$	22.4	8.9
$N = 250$	19.6	7.5
$N = 500$	17.7	6.5
$N = 1000$	15.6	5.9
$N = 2500$	13.9	5.2
$N = 5000$	12.3	5.3

Table 4: Percentage of SIR steps: nonlinear time series

This model requires simulation of more samples than the preceding one. In fact, the variance of the dynamic noise is more important and more trajectories are necessary to explore the space. The most interesting algorithm is the SIS with a suboptimal importance function which greatly limits the number of resampling steps over the prior importance function while avoiding a MC integration step needed to evaluate the optimal importance

function. This can be roughly explained by the fact that the observation noise is rather small so that y_k is highly informative and allows a limitation of the regions explored.

7. Conclusion

In this report, we have presented an overview of sequential simulation-based methods for Bayesian filtering of general hidden Markov models. This overview includes in the general framework of SIS numerous approaches that have been previously proposed independently in the literature for nearly 30 years. Several original extensions have also been presented. In this re-emerging area, there are numerous ways of improvement including among many others new variance reduction methods [20, 21] or efficient hybrid IS/MCMC methods.

8. Acknowledgments

I acknowledge Dr. T. Higuchi, Pr. G. Kitagawa and Dr. H. Tanizaki who sent me their preprints on this subject. I also acknowledge C. Andrieu, T. Clapp and Dr. S. Godsill for various comments that helped me to improve this report.

References

[1] H. Akashi and H. Kumamoto, "State Estimation for Systems under Measurements Noise with Markov Dependent Statistical Property - an Algorithm based on Random Sampling", in *Proc. 6^{th} Conf. IFAC*, 1975.

[2] H. Akashi and H. Kumamoto, "Construction of Discrete-time Nonlinear Filter by Monte Carlo Methods with Variance-reducing Techniques", *Sys. Cont.*, vol. 19, no. 4, 1975, pp. 211-221 (in Japanese).

[3] H. Akashi, H. Kumamoto and K. Nose, "Application of Monte Carlo Method to Optimal Control for Linear Systems under Measurement Noise with Markov Dependent Statistical Property", *Int. J. Cont.*, vol. 22, no. 6, 1975, pp. 821-836.

[4] H. Akashi and H. Kumamoto, "Random Sampling Approach to State Estimation in Switching Environments", *Automatica*, vol. 13, 1977, pp. 429-434.

[5] D.L. Aspach and H.W. Sorenson, "Nonlinear Bayesian Estimation using Gaussian Sum Approximation", *IEEE Trans. Auto. Cont.*, vol. 17, no. 4, pp. 439-448, 1972.

[6] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Englewood Cliffs, 1979.

[7] M.L. Andrade, L. Gimeno and M.J. Mendes, "On the Optimal and Suboptimal Nonlinear Filtering Problem for Discrete Time Systems", *IEEE Trans. Auto. Cont.*, vol. 17, 1978, pp. 439-448.

[8] M. Askar and H. Derin, "A Recursive Algorithm for the Bayes Solution of the Smoothing Problem", *IEEE Trans. Auto. Cont.*, vol. 26, no. 2, 1981, pp. 558-561.

[9] D. Avitzour, "A Stochastic Simulation Bayesian Approach to Multitarget Tracking", *IEE Proc. on Radar, Sonar and Navigation*, vol. 142, no. 2, 1995, pp. 41-44.

[10] E.R. Beadle and P.M. Djuric, "A Fast Weighted Bayesian Bootstrap Filter for Nonlinear Model State Estimation", *IEEE Trans. Aeros. Elec. Sys.*, vol. 33, no. 1, 1997, pp. 338-343.

[11] P.R. Benyon, "Monte Carlo and Other Methods for Nonlinear Non-Gaussian Estimation", *Math. Comp. Simul.*, no. 32, 1990, pp. 215-220.

[12] C. Berzuini, N. Best, W. Gilks and C. Larizza, "Dynamic Conditional Independence Models and Markov Chain Monte Carlo Methods", forthcoming *J. Am. Stat. Assoc.*, 1997.

[13] R.S Bucy and K.D. Senne, "Digital Synthesis of Nonlinear Filters", *Automatica*, vol. 7, 1971, pp. 287-298.

[14] R. Chen and J.S. Liu, "Predictive Updating Methods with Application to Bayesian Classification", *J. Roy. Stat. Soc. B*, vol. 58, no. 2, 1996, pp. 397-415.

[15] P. Del Moral and G. Salut, "Filtrage non-linéaire : résolution à la Monte Carlo", *C.R.A.S.,* vol. 320, 1995, pp. 1147-1152 (in French).

[16] P. Del Moral, J.C. Noyer, G. Rigal and G. Salut, "Résolution particulaire en traitement non-linéaire du signal: application Radar/Sonar", *Trait. Signal*, vol. 12, no. 4, 1995, pp. 287-301 (in French).

[17] A. Doucet, E. Barat and P. Duvaut, "A Monte Carlo approach to Recursive Bayesian State Estimation", in *Proc. IEEE Work. HOS*, June 1995, Spain.

[18] A. Doucet, E. Barat and P. Duvaut, "Implantation du paradigme bayésien pour l'estimation récursive d'état", in *Proc. 15*ème *colloque GRETSI*, Juan-les-Pins, 1995, pp. 73-76 (in French).

[19] A. Doucet, *Monte Carlo Methods for Bayesian Estimation of Hidden Markov Models. Application to Radiation Signals*, Ph.D. Thesis, Univ. Paris-Sud, Orsay, 1997 (in French with chapters 4 and 5 in English).

[20] A. Doucet and P. Duvaut, "A New Computational Method for Optimal Estimation of Nonlinear Non-Gaussian Dynamic Models", in *Proc. ICSPC'98*, Feb. 1998.

[21] A. Doucet and C. Andrieu, "A Killing and Splitting Scheme for Sequential Importance Sampling Applied to Bayesian Filtering", in preparation.

[22] J. Geweke, "Bayesian Inference in Econometrics Models using Monte Carlo Integration", *Econometrica*, vol. 57, 1989, pp. 1317-1339.

[23] N.J. Gordon, D.J. Salmond and A.F.M. Smith, "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation", *IEE-Proceedings-F*, vol. 140, no. 2, 1993, pp. 107-113.

[24] N. Gordon, "A Hybrid Bootstrap Filter for Target Tracking in Clutter", *IEEE Trans. Aero. Elec. Sys.*, vol. 33, no. 1, 1997, pp. 353-358.

[25] J.E. Handschin and D.Q. Mayne, "Monte Carlo Techniques to Estimate the Conditional Expectation in Multi-stage Non-linear Filtering", *Int. J. Cont.*, vol. 9, no. 5, 1969, pp. 547-559.

[26] J.E. Handschin, "Monte Carlo Techniques for Prediction and Filtering of Non-Linear Stochastic Processes", *Automatica*, vol. 6, 1970, pp. 555-563.

[27] T. Higuchi, "Kitagawa Monte Carlo Filter from the Perspective of Genetic Algorithm", Research Memorandum, The Institute of Statistical Mathematics, Tokyo, Japan, 1995.

[28] T. Higuchi, "Kitagawa Monte Carlo Filter using the Genetic Algorithm Operators", Research Memorandum, The Institute of Statistical Mathematics, Tokyo, Japan, 1995.

[29] T. Higuchi, "Bayesian Model for Seasonal Small Count Time Series and Monte Carlo Filter Approach", technical report, January 1997.

[30] M. Irwin, N. Cox and A. Kong, "Sequential Imputation for Multilocus Linkage Analysis", *Proc. Nat. Acad. Sci. USA*, vol. 91, 1994, pp. 11684-11688.

[31] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of the Conditional Density", in *Proc. Europ. Conf. Comp. Vision*, Cambridge, 1996, pp. 343-356.

[32] A.H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.

[33] G. Kitagawa, "Non-Gaussian State-Space Modeling of Nonstationary Time Series", *J. Am. Stat. Assoc.*, vol. 82, no. 400, 1987, pp. 1032-1063.

[34] G. Kitagawa, "A Monte Carlo Filtering and Smoothing Method for Non-Gaussian Nonlinear State Space Models", in *Proc. 2^{nd} US-Japan Joint Seminar on Statistical Time Series Analysis*, Honolulu, Hawaii, pp. 110-131, 1993.

[35] G. Kitagawa, "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models", *J. Comp. Graph. Stat.*, vol. 5, no. 1, pp. 1-25, 1996.

[36] G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, vol. 116, Springer, 1996.

[37] A. Kong, J.S. Liu and W.H. Wong, "Sequential Imputations and Bayesian Missing Data Problems", *J. Am. Stat. Assoc.*, vol. 89, no. 425, 1994, pp. 278-288.

[38] J.S. Liu and R. Chen, "Blind Deconvolution via Sequential Imputation", *J. Am. Stat. Assoc.*, vol. 90, no. 430, 1995, pp. 567-576.

[39] J.S. Liu, "Metropolized Independent Sampling with Comparison to Rejection Sampling and Importance Sampling", *Stat. Comp.*, vol. 6, 1996, pp. 113-119.

[40] J.S. Liu, "Nonparametric Hierarchical Bayes via Sequential Imputations", *Ann. Stat.*, 1996.

[41] J.S. Liu and R. Chen, "Monte Carlo Methods for Dynamic Systems", technical report, Department of Statistics, Stanford University, 1997.

[42] R.S. Mariano and H. Tanizaki, "Simulation-Based Inference in Nonlinear State-Space Models: Application to Testing the Permanent Income Hypothesis", in *Simulation-Based Inference in Econometrics: Methods and Applications* (R.S. Mariano, M. Weeks and T. Schuermann, Eds.), Cambridge University Press, 1996.

[43] P. Müller, "Monte Carlo Integration in General Dynamic Models", *Contemporary Math.*, vol. 115, 1991, pp. 145-163.

[44] P. Müller, "Posterior Integration in Dynamic Models", *Comp. Science Stat.*, vol. 24, 1992, pp. 318-324.

[45] B.D. Ripley, *Stochastic Simulation*, Wiley, New York, 1987.

[46] D.B. Rubin, "Using the SIR Algorithm to Simulate Posterior Distributions", in Bayesian Statistics 3 (Eds J.M. Bernardo, M.H. DeGroot, D.V. Lindley et A.F.M. Smith), Oxford University Press, pp. 395-402, 1988.

[47] E.I. Shapiro, "The Random Distribution Method and its Applications to the Solution of the Problem of Nonlinear Filtering in Discrete Time", *Radio Eng. Elec. Phys.*, vol. 26, no. 6, 1981, pp. 48-54.

[48] N. Shephard and M.K. Pitt, "Likelihood Analysis of Non-Gaussian Measurement Time Series", *Biometrika*, forthcoming, 1997.

[49] N. Shephard and M.K. Pitt, "Filtering via Simulation: Auxiliary Particle Filters", technical report, Department of Statistics, Imperial College, London, October 1997.

[50] A.F.M. Smith and A.E. Gelfand, "Bayesian Statistics without Tears: a Sampling-Resampling Perspective", *Am. Stat.*, vol. 46, no. 2, 1992, pp. 84-88.

[51] L. Stewart, "Bayesian Analysis using Monte Carlo Integration - a Powerful Methodology for Handling some Difficult Problems", *The Stat.*, vol. 32, 1983, pp. 195-200.

[52] L. Stewart and P. McCarty, "The Use of Bayesian Belief Networks to Fuse Continuous and Discrete Information for Target Recognition, Tracking and Situation Assessment", in *Proc. SPIE*, vol. 1699, 1992, pp. 177-185.

[53] V.B. Svetnik, "Applying the Monte Carlo Method for Optimum Estimation in Systems with Random Disturbances", *Auto. Remo. Cont.*, vol. 47, no. 6, 1986, pp. 818-825.

[54] H. Tanizaki, *Nonlinear Filters: Estimation and Applications*, Lecture Notes in Economics and Mathematical Systems, no. 400, Springer, Berlin, 1993.

[55] H. Tanizaki and R.S. Mariano, "Prediction, Filtering and Smoothing in Non-linear and Non-normal Cases using Monte Carlo Integration", *J. App. Econometrics*, vol. 9, 1994, pp. 163-179.

[56] H. Tanizaki and R.S. Mariano, "Nonlinear and Nonnormal State-Space Modeling with Monte-Carlo Stochastic Simulations", forthcoming *J. Econometrics,* 1997.

[57] H. Tanizaki, "Nonlinear and Nonnormal Filters using Monte-Carlo Methods", forthcoming *Comp. Stat. Data Ana.*, 1997.

[58] J.K. Tugnait, "Detection and Estimation for Abruptly Changing Systems", *Automatica*, vol. 18, no. 5, 1982, pp. 607-615.

[59] M. West, "Mixtures Models, Monte Carlo, Bayesian Updating and Dynamic Models", *Comp. Science Stat.*, vol. 24, 1993, pp. 325-333.

[60] M. West and J.F. Harrison, *Bayesian Forecasting and Dynamic Models*, Springer Verlag Series in Statistics, 2nd edition, 1997.

[61] V.S. Zaritskii, V.B. Svetnik and L.I. Shimelevich, "Monte Carlo Technique in Problems of Optimal Data Processing", *Auto. Remo. Cont.*, vol. 12, 1975, pp. 95-103.