Sequential simulation-based estimation of jump Markov linear systems

Arnaud Doucet
Signal Processing Group, Department of Engineering,
University of Cambridge, Trumpington Street , CB2 1PZ Cambridge, UK.
Email: ad2@eng.cam.ac.uk

Neil J. Gordon[1]
Signal Processing and Imagery Department,
Defence Evaluation and Research Agency,
St Andrews Road, Malvern, Worcestershire, WR14 3PS, UK.
Email: N.Gordon@signal.dera.gov.uk

Vikram Krishnamurthy
Department of Electrical and Electronic Engineering,
The University of Melbourne, Parkville, Victoria 3052, Australia.
Email: v.krishnamurthy@ee.mu.oz.au

Abstract

Jump Markov linear systems (JMLS) are linear systems whose parameters evolve with time according to a finite state Markov chain. Our aim is to recursively compute optimal conditional mean state estimates for JMLS. We present efficient simulation-based algorithms called particle filters to solve the optimal filtering problem. Our algorithms combine sequential importance sampling, a selection scheme and Markov chain Monte Carlo methods. They use several variance reduction methods to make the most of the statistical structure of JMLS.

1 Introduction

Let r_t, $t = 1, 2, \ldots$ denote a discrete time Markov chain with known transition probabilities. A jump Markov linear system can be modelled as

$$x_{t+1} = A(r_{t+1})\, x_t + B(r_{t+1})\, v_{t+1} + F(r_{t+1})\, u_t \quad (1)$$

$$y_t = C(r_t)\, x_t + D(r_t)\, \varepsilon_t + G(r_t)\, u_t \quad (2)$$

where u_t denotes a known exogeneous input, v_t and ε_t denote independent white Gaussian noise sequences. A jump Markov linear system can be viewed as a linear system whose parameters ($A(r_t)$, $B(r_t)$, $C(r_t)$, $D(r_t)$, $F(r_t)$, $G(r_t)$) evolve with time according to a finite state Markov chain (r_t). Neither the continuous-state process x_t nor the finite state process r_t are observed – instead we observe the noisy measurement process y_t.

[1] The research of N.J. Gordon was sponsored by the UK MOD corporate research program TG9.

Jump Markov linear systems are widely used in several fields of signal processing including digital communications such as interference suppression in CDMA spread spectrum systems [14], target tracking [2] and de-interleaving of pulse trains [14]. They can be viewed as a generalization of the Hidden Markov Model (HMM) (which consists of a Markov chain r_t observed in white noise) to correlated noise.

It is well known that exact computation of the conditional mean filtered or smoothed state estimates of x_t and r_t involves a prohibitive computational cost exponential in the (growing) number of observations. This is unlike the standard HMM for which conditional mean state estimates can be computed with linear complexity in the number of observations via the HMM filter. Recently, efficient batch (off-line) deterministic and stochastic iterative algorithms have been proposed to compute fixed-interval smoothed conditional mean and maximum *a posteriori* (MAP) state estimates of x_t and r_t, see [5], [14]. However, in most real-world applications, one wants to compute state estimates of x_t and r_t recursively (on-line). The aim of this paper is to present **simulation-based recursive** filtering algorithms for computing conditional mean estimates of the states x_t and r_t given the observation history, namely, $\mathbb{E}\{x_t|\mathbf{y}_{1:t}\}$ and $\mathbb{E}\{r_t|\mathbf{y}_{1:t}\}$. Simulation based algorithms for computing fixed-lag smoothed state estimates $\mathbb{E}\{x_t|\mathbf{y}_{1:t+L}\}$ and $\mathbb{E}\{r_t|\mathbf{y}_{1:t+L}\}$ are also of interest – see [8] for details.

Due to the prohibitive computational cost required to compute filtered state estimates of x_t and r_t, it is necessary to consider in practice suboptimal estimation algorithms. A variety of algorithms has already been pro-

posed in the literature to solve these estimation problems [2]. Most of these algorithms are based on deterministic finite Gaussian mixture approximations like the popular Interacting Multiple Model (IMM) or the Generalised Pseudo Bayes (GPB) algorithms [2]. These methods are computationally cheap but they can fail in difficult situations.

The MC particle filters can be viewed as simulation-based approximations of the filtering distribution. Taking advantage of the increase of computational power and the availability of parallel computers, several authors have recently proposed such MC particle methods [11]. It has been shown that these methods outperform the standard suboptimal methods. In this paper, we propose improved simulation-based approximations of the optimal filter and smoother with novel variance reduction methods: the filtering distributions of interest are approximated by a Gaussian mixture of a large number, say N, of components which evolve stochastically over time and are driven by the observations.

MC particle methods to solve optimal estimation problems were introduced in automatic control at the end of the 60's by Handschin and Mayne [10]. Interesting developments were then subsequently proposed in the 70's [1]. Most likely because of the primitive computers available at the time, these papers were overlooked and forgotten. In the beginning of the 90's, the great increase in computational power allowed for the rebirth of this field. In 1993, Gordon, Salmond and Smith [9] proposed an algorithm, the bootstrap filter, which introduced a selection step that statistically multiplies and/or discards particles at each time. This key step led to the first operational particle filter. Following this seminal paper, particle filters have stimulated great interest in the engineering and statistical literature. With these filters, complex non-linear non-Gaussian estimation problems can be solved efficiently in an on-line manner. Moreover, they are much easier to implement than classical numerical methods and, contrary to deterministic grid-based methods, their rate of convergence is theoretically not sensitive to the size of the state space [4].

2 Problem Formulation

2.1 Signal Model

Let r_t denote a discrete-time, time-homogeneous, s-state, first-order Markov chain with transition probabilities $p_{m,n} \triangleq \Pr\{r_{t+1} = n | r_t = m\}$ for any $m, n \in S$ where $S \triangleq \{1, 2, ..., s\}$. The transition probability matrix $[p_{m,n}]$, is thus an $s \times s$ matrix, with elements satisfying $p_{m,n} \geq 0$ and $\sum_{n=1}^{s} p_{m,n} = 1$, for each $m \in S$. Denote the initial probability distribution as $p_m \triangleq \Pr\{r_1 = m\}$, for $m \in S$, such that $p_m \geq 0, \forall m \in S$ and $\sum_{m=1}^{s} p_m = 1$. Consider the following JMLS given in equations (1)-(2) where $x_t \in \mathbb{R}^{n_x}$ is the system state, $y_t \in \mathbb{R}^{n_y}$ is the observation at time t, $u_t \in \mathbb{R}^{n_u}$ is a known deterministic input, $v_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{n_v}) \in \mathbb{R}^{n_v}$ and $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{n_w}) \in \mathbb{R}^{n_w}$ are i.i.d. Gaussian sequences, and $D(i) D^{\mathsf{T}}(i) > 0$ $(\forall i \in S)$. $A(\cdot), B(\cdot), C(\cdot), D(\cdot), F(\cdot)$ and $G(\cdot)$ are functions of the Markov chain state r_t, *i.e.* $(A(\cdot), B(\cdot), C(\cdot), D(\cdot), F(\cdot), G(\cdot)) \subset \{(A(m), B(m), C(m), D(m), F(m), G(m)); m \in S\}$ and they evolve according to the realisation of the finite state Markov chain r_t . We assume that $x_0 \sim \mathcal{N}(\hat{x}_0, P_0)$ where $P_0 > 0$ and let x_0, v_t and ε_t be mutually independent for all t. The model parameters $\lambda \triangleq \{p_m, p_{mn}, A(m), B(m), C(m), D(m), F(m), G(m), \hat{x}_0, P_0; m, n \in S\}$ are assumed known.

2.2 Estimation objectives

Given at time t the observations $\mathbf{y}_{1:t}$, assuming that the model parameters λ are exactly known, all Bayesian inference for JMLS relies on the joint posterior distribution $p(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t})$ where $p(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}) = p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t}, \mathbf{r}_{1:t}) p(\mathbf{r}_{1:t} | \mathbf{y}_{1:t})$. Given $\mathbf{r}_{1:t}$, $p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t}, \mathbf{r}_{1:t})$ is a Gaussian distribution whose parameters can be evaluated using a Kalman filter. $p(\mathbf{r}_{1:t} | \mathbf{y}_{1:t})$ could be computed exactly but this discrete distribution has s^t values and thus some approximations have to be made as time increases.

In this paper, we are interested in the following optimal estimation problem: Obtain the filtering distribution $p(r_t, x_t | \mathbf{y}_{1:t})$ as well as the MMSE estimate of $\varphi_{t|t}(r_t, x_t)$ given by $I\left(\varphi_{t|t}\right) \triangleq \mathbb{E}_{p(r_t, x_t | \mathbf{y}_{1:t})}\left(\varphi_{t|t}(r_t, x_t)\right)$ where $\varphi_{t|t} : S \times \mathbb{R}^{n_{\mathbf{x}}} \to \mathbb{R}^{n_{\varphi_{t|t}}}$.

We restrict ourselves to the common case where $\mathbb{E}_{p(x_t | \mathbf{y}_{1:t}, \mathbf{r}_{1:t})}\left(\varphi_{t|t}(r_t, x_t)\right)$ and $\mathbb{E}_{p(x_t | \mathbf{y}_{1:t+L}, \mathbf{r}_{1:t})}\left(\varphi_{t|t+L}(r_t, x_t)\right)$ can be computed analytically.

Remark: In most filtering applications, we are interested in estimating the MMSE (conditional mean) state estimates $\mathbb{E}(x_t | \mathbf{y}_{1:t})$ and $cov(x_t | \mathbf{y}_{1:t})$. In these cases $\mathbb{E}_{p(x_t | \mathbf{y}_{1:t}, \mathbf{r}_{1:t})}\left(\varphi_{t|t}(r_t, x_t)\right)$ can be computed analytically using the Kalman filter for the sequence $\mathbf{r}_{1:t}$.

3 A Simulation-Based Optimal Filter

We describe the standard Bayesian importance sampling method, then show how variance reduction can be achieved by integrating out the states $\mathbf{x}_{0:t}$ using the Kalman filter. Then, we present a sequential version of Bayesian importance sampling for optimal filtering,

generalising current approaches in the literature. We show why it is necessary to introduce a selection scheme and then we propose a generic Monte Carlo filter.

3.1 Monte Carlo simulation for optimal estimation

For any $\varphi_{t|t}$, we will assume subsequently that $\left|I\left(\varphi_{t|t}\right)\right| < +\infty$. If we were able to sample N i.i.d. random samples, called particles, $\left\{\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right); i = 1, ..., N\right\}$ according to $p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$, then an empirical estimate of this distribution would be given by

$$\overline{p_N}\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right)}\left(d\mathbf{r}_{1:t}, d\mathbf{x}_{0:t}\right)$$

and, as a corollary, an estimate of $p\left(r_t, x_t | \mathbf{y}_{1:t}\right)$ is $\overline{p_N}\left(r_t, x_t | \mathbf{y}_{1:t}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\left(r_t^{(i)}, x_t^{(i)}\right)}\left(dr_t, dx_t\right)$. From this distribution, one can easily obtain an estimate of $I\left(\varphi_{t|t}\right)$ for any $\varphi_{t|t}$

$$\overline{I_N}\left(\varphi_{t|t}\right) = \int \varphi_{t|t}\left(r_t, x_t\right) \overline{p_N}\left(r_t, x_t | \mathbf{y}_{1:t}\right) dr_t dx_t$$

$$= \frac{1}{N}\sum_{i=1}^{N} \varphi_{t|t}\left(r_t^{(i)}, x_t^{(i)}\right)$$

This estimate is unbiased and, from the strong law of large numbers (SLLN), $\overline{I_N}\left(\varphi_{t|t}\right)$ converges almost surely (a.s.) towards $I\left(\varphi_{t|t}\right)$ as $N \to +\infty$. If $\sigma^2_{\varphi_{t|t}} \triangleq var_{p\left(r_t, x_t | \mathbf{y}_{1:t}\right)}\left[\varphi_{t|t}\left(r_t, x_t\right)\right] < +\infty$, then a central limit theorem (CLT) holds

$$\sqrt{N}\left[\overline{I_N}\left(\varphi_{t|t}\right) - I\left(\varphi_{t|t}\right)\right] \underset{N \to +\infty}{\Rightarrow} \mathcal{N}\left(0, \sigma^2_{\varphi_{t|t}}\right)$$

where " $\Rightarrow$ " denotes convergence in distribution. The advantage of the MC method is clear. One can easily estimate $I\left(\varphi_{t|t}\right)$ for any $\varphi_{t|t}$ and the rate of convergence of this estimate does not depend on t. Unfortunately, it is impossible to sample efficiently from the "target" posterior distribution $p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$ at any time t. So we focus on alternative methods.

A solution to estimate $p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$ and $I\left(\varphi_{t|t}\right)$ consists of using the well-known importance sampling method [3]. This method is based on the following remark. Let us introduce an arbitrary importance distribution $\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$, from which it is easy to obtain samples, and such that $p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right) > 0$ implies $\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right) > 0$, then

$$I\left(\varphi_{t|t}\right) = \frac{\mathbb{E}_{\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)}\left(\varphi_{t|t}\left(r_t, x_t\right) w\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t}\right)\right)}{\mathbb{E}_{\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)}\left(w\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t}\right)\right)}$$

where the importance weight is equal to

$$w\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t}\right) = \frac{p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)}{\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)}$$

If we have N i.i.d. random samples $\left\{\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right); i = 1, ..., N\right\}$ distributed according to $\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$ then a Monte Carlo estimate of $I\left(\varphi_{t|t}\right)$ is given by

$$\widehat{I_N^1}\left(\varphi_{t|t}\right) = \frac{\widehat{A_N^1}\left(\varphi_{t|t}\right)}{\widehat{B_N^1}\left(\varphi_{t|t}\right)} = \frac{\sum_{i=1}^{N} \varphi_{t|t}\left(r_t^{(i)}, x_t^{(i)}\right) w\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right)}{\sum_{i=1}^{N} w\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right)}$$

$$= \sum_{i=1}^{N} \widetilde{w}_{1:t}^{(i)} \varphi_{t|t}\left(r_t^{(i)}, x_t^{(i)}\right)$$

where the normalised importance weights $\widetilde{w}_{1:t}^{(i)}$ are

$$\widetilde{w}_{1:t}^{(i)} = \frac{w\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right)}{\sum_{j=1}^{N} w\left(\mathbf{r}_{1:t}^{(j)}, \mathbf{x}_{0:t}^{(j)}\right)}$$

This method is equivalent to the following point mass approximation of $p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$

$$\widehat{p_N}\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right) = \sum_{i=1}^{N} \widetilde{w}_{1:t}^{(i)} \delta_{\left(\mathbf{r}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)}\right)}\left(d\mathbf{r}_{1:t}, d\mathbf{x}_{0:t}\right)$$

and thus $\widehat{p_N}\left(r_t, x_t | \mathbf{y}_{1:t}\right) = \sum_{i=1}^{N} \widetilde{w}_{1:t}^{(i)} \delta_{\left(r_t^{(i)}, x_t^{(i)}\right)}\left(dr_t, dx_t\right)$. The "perfect" simulation case, *i.e.* $\pi\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right) = p\left(\mathbf{r}_{1:t}, \mathbf{x}_{0:t} | \mathbf{y}_{1:t}\right)$, would correspond to $\widetilde{w}_{1:t}^{(i)} = N^{-1}$ for any i. In practice, we will try to select the importance distribution as close as possible to the target distribution in a given sense. For N finite, $\widehat{I_N^1}\left(\varphi_{t|t}\right)$ is biased (ratio of estimates), but asymptotically according to the SLLN, $\widehat{I_N}\left(\varphi_{t|t}\right)$ converges a.s. towards $I\left(\varphi_{t|t}\right)$. Under additional assumptions, a CLT also holds. However, we first show, in the next subsection, how the variance of the estimate $\widehat{I_N^1}\left(\varphi_{t|t}\right)$ can be decreased.

3.2 Variance Reduction

It is possible to reduce the problem of estimating $p\left(r_t, x_t | \mathbf{y}_{1:t}\right)$ and $I\left(\varphi_{t|t}\right)$ to one of sampling from $p\left(\mathbf{r}_{1:t} | \mathbf{y}_{1:t}\right)$. Indeed $p\left(\mathbf{r}_{1:t}, x_t | \mathbf{y}_{1:t}\right) = p\left(\mathbf{r}_{1:t} | \mathbf{y}_{1:t}\right) p\left(x_t | \mathbf{y}_{1:t}, \mathbf{r}_{1:t}\right)$ where $p\left(x_t | \mathbf{y}_{1:t}, \mathbf{r}_{1:t}\right)$ is a Gaussian distribution whose parameters are given by the Kalman filter. Thus, given an approximation of $p\left(\mathbf{r}_{1:t} | \mathbf{y}_{1:t}\right)$, ones gets straightforwardly an approximation of $p\left(r_t, x_t | \mathbf{y}_{1:t}\right)$. Moreover, if $\mathbb{E}_{p\left(x_t | \mathbf{y}_{1:t}, \mathbf{r}_{1:t}\right)}\left(\varphi_{t|t}\left(r_t, x_t\right)\right)$ can be evaluated in a closed-form expression, then the following alternative Bayesian importance sampling estimate of $I\left(\varphi_{t|t}\right)$ can

be proposed

$$\widehat{I_N^2}\left(\varphi_{t|t}\right) = \frac{\widehat{A_N^2}\left(\varphi_{t|t}\right)}{\widehat{B_N^2}\left(\varphi_{t|t}\right)}$$

$$\frac{\sum_{i=1}^N \mathbb{E}_{p\left(x_t|\mathbf{y}_{1:t},\mathbf{r}_{1:t}^{(i)}\right)}\left(\varphi_{t|t}\left(r_t^{(i)},x_t\right)\right) w\left(\mathbf{r}_{1:t}^{(i)}\right)}{\sum_{i=1}^N w\left(\mathbf{r}_{1:t}^{(i)}\right)}$$

where $w(\mathbf{r}_{1:t}) = \frac{p(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}$

$$\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t}) = \int \pi(\mathbf{r}_{1:t},\mathbf{x}_{0:t}|\mathbf{y}_{1:t})\, d\mathbf{x}_{0:t}$$

Intuitively, to reach a given precision, $\widehat{I_N^2}\left(\varphi_{t|t}\right)$ will require a reduced number N of samples over $\widehat{I_N^1}\left(\varphi_{t|t}\right)$ as we only need to sample from a lower-dimensional distribution. This is proven in the following propositions where it is shown that, if one can integrate $\mathbf{x}_{0:t}$ analytically, then the variances of the resulting estimates are lower than the ones of the "crude" estimates.

Proposition 1 *The variances of the importance weights, the numerators and the denominators satisfy for any N*

$$var_{\pi(\mathbf{r}_{1:t},\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}\left[w(\mathbf{r}_{1:t},\mathbf{x}_{0:t})\right] - var_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left[w(\mathbf{r}_{1:t})\right]$$
$$= \mathbb{E}_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left[var_{\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t},\mathbf{r}_{1:t})}\left[w(\mathbf{r}_{1:t},\mathbf{x}_{0:t})\right]\right]$$
$$var_{\pi(\mathbf{r}_{1:t},\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}\left(\widehat{A_N^1}\left(\varphi_{t|t}\right)\right) - var_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left(\widehat{A_N^2}\left(\varphi_{t|t}\right)\right)$$
$$= \mathbb{E}_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left[var_{\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t},\mathbf{r}_{1:t})}\left[\widehat{A}_N^1\left(\varphi_{t|t}\right)\right]\right]$$
$$var_{\pi(\mathbf{r}_{1:t},\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}\left(\widehat{B_N^1}\left(\varphi_{t|t}\right)\right) - var_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left(\widehat{B_N^2}\left(\varphi_{t|t}\right)\right)$$
$$= \mathbb{E}_{\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})}\left[var_{\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t},\mathbf{r}_{1:t})}\left[\widehat{B}_N^1\left(\varphi_{t|t}\right)\right]\right]$$

A sufficient condition for $\widehat{I_N^1}\left(\varphi_{t|t}\right)$ to satisfy a CLT is $var_{p(r_t,x_t|\mathbf{y}_{1:t})}\left\{\varphi_{t|t}(r_t,x_t)\right\} < +\infty$ and $w(\mathbf{r}_{1:t},\mathbf{x}_{0:t}) < C_t < +\infty$ for any $(\mathbf{r}_{1:t},\mathbf{x}_{0:t}) \in S^t \times (\mathbb{R}^{n_\mathbf{x}})^{t+1}$ [3]. This trivially implies that $\widehat{I_N^2}\left(\varphi_{t|t}\right)$ also satisfies a CLT, see [8] for details.

Given these results, we now focus on importance sampling methods to get an approximation of $p(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})$ and $I\left(\varphi_{t|t}\right)$ using an importance distribution $\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})$. Up to now, the methods we have described are batch methods. We show in the next subsection how to obtain a sequential method.

3.3 Sequential Importance Sampling

One can always rewrite the importance function at time t as follows

$$\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t}) = \pi(r_1|\mathbf{y}_{1:t}) \prod_{k=2}^{t} \pi(r_k|\mathbf{y}_{1:t},\mathbf{r}_{1:k-1})$$

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

$$\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t}) = \pi(r_1|y_1) \prod_{k=2}^{t} \pi(r_k|\mathbf{y}_{1:k},\mathbf{r}_{1:k-1}) \quad (3)$$

Such an importance function allows for a recursive evaluation of $w(\mathbf{r}_{1:t}) = w(\mathbf{r}_{1:t-1})\, w_t$ and thus of $\widetilde{w}_{1:t}$, the *incremental weight* w_t is given by $w_t = \frac{p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t})p(r_t|r_{t-1})}{p(y_t|\mathbf{y}_{1:t-1})\pi(r_t|\mathbf{y}_{1:t},\mathbf{r}_{1:t-1})}$

$$\propto \frac{p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t})\, p(r_t|r_{t-1})}{\pi(r_t|\mathbf{y}_{1:t},\mathbf{r}_{1:t-1})}$$

Further on, $\widetilde{w}_t$ denotes the normalized version of w_t, i.e. $\widetilde{w}_t^{(i)} = \left[\sum_{j=1}^N w_t^{(j)}\right]^{-1} w_t^{(i)}$.

Choice of the importance distribution: There are infinitely many possible choices for $\pi(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})$, the only condition being that its support includes the one of $p(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})$, that is the support of $p(\mathbf{r}_{1:t})$. A sensible selection criterion is to choose a proposal that minimizes the variance of the importance weights at time t, given $\mathbf{r}_{1:t-1}$ and $\mathbf{y}_{1:t}$. According to this strategy, the following proposition establishes what the optimal importance distribution is.

Proposition 2 *$p(r_t|\mathbf{y}_{1:t},\mathbf{r}_{1:t-1})$ is the distribution which minimises the variance of the importance weights conditional upon $\mathbf{r}_{1:t-1}$ and $\mathbf{y}_{1:t}$.*

The proof is straightforward as one can easily check that the conditional variance is equal to zero in this case. We show how to implement this "optimal" distribution and then describe several suboptimal methods.

- *Optimal sampling distribution.* The optimal distribution satisfies $p(r_t = m|\mathbf{r}_{1:t-1},\mathbf{y}_{1:t}) =$

$$\frac{p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t-1},r_t=m)\, p(r_t=m|r_{t-1})}{p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t-1})}$$

and the associated importance weight w_t is proportional to $p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t-1}) =$

$$\sum_{m=1}^{s} p(y_t|\mathbf{y}_{1:t-1},\mathbf{r}_{1:t-1},r_t=m)\, p(r_t=m|r_{t-1})$$
$$= \sum_{\substack{m=1 \\ p_{r_{t-1},m}}}^{s} \Phi\left(\widetilde{y}_{t|t-1}(\mathbf{r}_{1:t-1},r_t=m), S_t(\mathbf{r}_{1:t-1},r_t=m)\right)$$

where $\widetilde{y}_{t|t-1}(\mathbf{r}_{1:t})$ and $S_t(\mathbf{r}_{1:t-1}, r_t = m)$ are respectively the innovation and the one-step ahead prediction covariance of the observation conditional upon $(\mathbf{r}_{1:t-1}, r_t = m)$. Computing $p(y_t|\mathbf{y}_{1:t-1}, \mathbf{r}_{1:t-1})$ requires the evaluation of s one-step ahead Kalman filter steps. It is thus computationally intensive if s is large.

- *Prior distribution.* If we use the prior distribution $p(r_t|r_{t-1})$ as importance distribution, the importance weight is proportional to $p(y_t|\mathbf{y}_{1:t-1}, \mathbf{r}_{1:t}) = \Phi\left(\widetilde{y}_{t|t-1}(\mathbf{r}_{1:t}), S_t(\mathbf{r}_{1:t})\right)$. It only requires one step of a Kalman filter to be evaluated.

- *Alternative sampling distribution.* It is possible to design a variety of alternative sampling distributions. For example, one can use the results of a suboptimal deterministic algorithm to construct an importance sampling distribution.

Degeneracy of the algorithm: The following proposition shows that, for importance functions of the form (3), the variance of $w(\mathbf{r}_{1:t})$ can only increase (stochastically) over time. The proof of this proposition is an extension of a Kong-Liu-Wong [12, p. 285] theorem to the case of an importance function of the form (3) and it is omitted here.

Proposition 3 *The unconditional variance (*i.e. *with the observations $\mathbf{y}_{1:t}$ being interpreted as random variables) of the importance weights $w(\mathbf{r}_{1:t})$ increases over time.*

It is thus 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 and a large computational burden is devoted to updating trajectories whose contribution to the final estimate is almost zero. That is why it is of crucial importance to introduce a selection step in the algorithm. The aim of this selection step is to discard the particles $\mathbf{r}_{1:t}^{(i)}$ with low normalised importance weights $\widetilde{w}\left(\mathbf{r}_{1:t}^{(i)}\right)$ and to multiply the ones with high $\widetilde{w}\left(\mathbf{r}_{1:t}^{(i)}\right)$ so as to avoid the degeneracy of the algorithm and to jump into the interesting zones of the space. Each time a selection step is used the weights are reset to N^{-1}.

3.4 Selection step

A selection procedure associates to each particle, say $\widetilde{\mathbf{r}}_{1:t}^{(i)}$ $(i = 1, \ldots, N)$, a number of "children" $N_i \in \mathbb{N}$, such that $\sum_{i=1}^{N} N_i = N$, to obtain N new particles $\mathbf{r}_{1:t}^{(i)}$. If $N_i = 0$, then $\widetilde{\mathbf{r}}_{1:t}^{(i)}$ is discarded, otherwise it has N_i "children" at time t. If we use a selection scheme *at each time step* then, before the selection scheme, we have a weighted distribution $\widetilde{p_N}(\mathbf{r}_{1:t}|\mathbf{y}_{1:t}) = \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta_{\widetilde{\mathbf{r}}_{1:t}^{(i)}}(d\mathbf{r}_{1:t})$ and, after the selection step, we have $\widehat{p_N}(\mathbf{r}_{1:t}|\mathbf{y}_{1:t}) = N^{-1} \sum_{i=1}^{N} \delta_{\mathbf{r}_{1:t}^{(i)}}(d\mathbf{r}_{1:t})$.

Some selection schemes: We describe here some selection schemes and show how to implement them in $O(N)$ iterations.

- *Sampling Importance Resampling (SIR)/Multinomial Sampling procedure.* This procedure, introduced originally by Gordon *et al.* [9], is the most popular one. One samples N times from $\widetilde{p_N}(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})$ to obtain $\left(\mathbf{r}_{1:t}^{(i)}; i = 1, \ldots, N\right)$. This is equivalent to drawing jointly $(N_i; i = 1, \ldots, N)$ according to a multinomial distribution of parameters N and $\widetilde{w}_t^{(i)}$. This algorithm has been originally implemented in $O(N \log N)$ operations [9]. 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 uniformly distributed in $[0,1]$ and **ordered**, *i.e.* $u_1 \leq u_2 \leq \cdots \leq u_N$, using a classical algorithm [15, p. 96]. In this case, we have $\mathbb{E}[N_i] = N\widetilde{w}_t^{(i)}$ and $var[N_i] = N\widetilde{w}_t^{(i)}\left(1 - \widetilde{w}_t^{(i)}\right)$. However, as pointed out in [13], it is possible and better to use selection schemes with a reduced variance.

- *Residual Resampling* [13]. This procedure performs as follows. Set $\widetilde{N}_i = \left\lfloor N\widetilde{w}_t^{(i)} \right\rfloor$ then perform a SIR procedure to select the remaining $\overline{N}_t = N - \sum_{i=1}^{N} \widetilde{N}_i$ samples with the new weights $w_t'^{(i)} = \left(\widetilde{w}_t^{(i)} N - \widetilde{N}_i\right)/\overline{N}_t$, finally add the results to the current $\widetilde{N}_i$. In this case, we obtain $\mathbb{E}[N_i] = N\widetilde{w}_t^{(i)}$ but $var[N_i] = \overline{N}_t w_t'^{(i)}\left(1 - w_t'^{(i)}\right)$.

Recent theoretical results obtained in [4] suggest that it is not necessary to design unbiased selection schemes, *i.e.* we can have $\mathbb{E}[N_i] \neq N\widetilde{w}_t^{(i)}$.

On the use of a selection scheme: Two estimates of $I\left(\varphi_{t|t}\right)$ can be proposed before, $\widetilde{I_N}\left(\varphi_{t|t}\right)$, and after, $\widehat{I_N}\left(\varphi_{t|t}\right)$, the selection scheme at time t where

$$\widetilde{I_N}\left(\varphi_{t|t}\right) = \int \mathbb{E}_{p(x_t|\mathbf{y}_{1:t}, \mathbf{r}_{1:t})}\left(\varphi_{t|t}(r_t, x_t)\right) \widetilde{p_N}(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})\, d\mathbf{r}_{1:t}$$

$$\widehat{I_N}\left(\varphi_{t|t}\right) = \int \mathbb{E}_{p(x_t|\mathbf{y}_{1:t}, \mathbf{r}_{1:t})}\left(\varphi_{t|t}(r_t, x_t)\right) \widehat{p_N}(\mathbf{r}_{1:t}|\mathbf{y}_{1:t})\, d\mathbf{r}_{1:t}$$

Using the variance decomposition, it is straightforward to show that if the selection scheme used is unbiased then

$$var\left(\widehat{I_N}\left(\varphi_{t|t}\right)\right) \geq var\left(\widetilde{I_N}\left(\varphi_{t|t}\right)\right)$$

So it is better to estimate $I\left(\varphi_{t|t}\right)$ using $\widetilde{I_N}\left(\varphi_{t|t}\right)$ as the selection scheme can only increase the variance of the estimate.

3.5 Implementation Issues

Given at time $t-1$, $N \in \mathbb{N}^*$ random samples $\left(\mathbf{r}_{1:t-1}^{(i)}; i = 1, \ldots, N\right)$ distributed approximately ac-

cording to $p(\mathbf{r}_{1:t-1}|\mathbf{y}_{1:t-1})$, the MC filter proceeds as follows at time t.

Particle Filter for JMLS

Sequential Importance Sampling step

- For $i = 1, ..., N$, sample $\widetilde{r}_t^{(i)} \sim \pi(r_t|\mathbf{y}_{1:t}, \mathbf{r}_{1:t-1}^{(i)})$ and set $\widetilde{\mathbf{r}}_{1:t}^{(i)} \triangleq \left(\mathbf{r}_{1:t-1}^{(i)}, \widetilde{r}_t^{(i)}\right)$.
- For $i = 1, ..., N$, evaluate the importance weights up to a normalising constant:

$$w_t^{(i)} \propto \frac{p\left(y_t|\mathbf{y}_{1:t-1}, \widetilde{\mathbf{r}}_{1:t}^{(i)}\right) p\left(\widetilde{r}_t^{(i)} \middle| \widetilde{r}_{t-1}^{(i)}\right)}{\pi\left(\widetilde{r}_t^{(i)} \middle| \mathbf{y}_{1:t}, \mathbf{r}_{1:t-1}^{(i)}\right)} \qquad (4)$$

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

$$\widetilde{w}_t^{(i)} = \left[\sum_{j=1}^{N} w_t^{(j)}\right]^{-1} w_t^{(i)} \qquad (5)$$

Selection step

- Multiply/Discard particles $\left(\widetilde{\mathbf{r}}_{1:t}^{(i)}; i = 1, \ldots, N\right)$ with respect to high/low normalised importance weights $\widetilde{w}_t^{(i)}$ to obtain N particles $\left(\mathbf{r}_{1:t}^{(i)}; i = 1, \ldots, N\right)$.

Clearly, the computational complexity of this algorithm at each iteration is $O(N)$. At first sight, it could appear that one needs to keep in memory the paths of all trajectories, that is $\left(\widetilde{\mathbf{r}}_{1:t}^{(i)}; i = 1, \ldots, N\right)$. In this case, the storage requirements would increase linearly over time. Actually, under the standard assumption that $\pi(r_t|\mathbf{y}_{1:t}, \mathbf{r}_{1:t-1})$ only depends on $\mathbf{r}_{1:t-1}$ via the set of low-dimensional sufficient statistics $m_{t|t-1}\left(\mathbf{r}_{1:t}^{(i)}\right)$ and $P_{t|t-1}\left(\mathbf{r}_{1:t}^{(i)}\right)$, this is the case for $p(r_t|r_{t-1})$ and $p(r_t|\mathbf{y}_{1:t}, \mathbf{r}_{1:t-1})$, then one only needs to keep in memory these statistics. So the storage requirements are still $O(N)$ and do not increase over time.

References

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

[2] Y. Bar-Shalom and X.R. Li, Multitarget-Multisensor Tracking: Principles and Techniques, 1995.

[3] J.M. Bernardo and A.F.M. Smith, Bayesian Theory, Wiley, 1994.

[4] D. Crisan, P. Del Moral and T. Lyons, "Discrete filtering using branching and interacting particle systems", Markov Processes and Related Fields, vol. 5, no. 3, pp. 293-318, 1999.

[5] A. Doucet, A. Logothetis and V. Krishnamurthy, "Stochastic sampling algorithms for state estimation of jump Markov linear systems", IEEE Trans. Automatic Control, vol. 45, no. 2, pp. 188-201, 2000.

[6] A. Doucet, J.F.G. de Freitas and N.J. Gordon (eds.), Sequential Monte Carlo Methods in Practice, Springer-Verlag, 2000.

[7] A. Doucet, S.J. Godsill and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering", Statistics and Computing, vol. 10, no. 3, pp. 197-208, 2000.

[8] A. Doucet, N.J. Gordon and V. Krishnamurthy, "Particle filtering algorithms for state estimation of jump Markov linear systems", technical report Cambridge University, CUED-F-INFENG TR. 359, 1999 – to appear in IEEE Trans Signal Processing, 2001.

[9] 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, pp. 107-113, 1993.

[10] 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.

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

[12] A. Kong, J.S. Liu and W.H. Wong, "Sequential imputations and Bayesian missing data problems", J. Am. Stat. Assoc., vol. 89, pp. 278-288, 1994.

[13] J.S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems", J. Am. Stat. Assoc., vol. 93, pp. 1032-1044, 1998.

[14] A. Logothetis and V. Krishnamurthy, "Expectation-maximization algorithms for MAP estimation of jump Markov linear systems", IEEE Trans. Signal Processing, vol. 47, no. 8, pp.2139-2156, 1999.

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