Convergence of Sequential Monte Carlo Methods

by

Dan Crisan[‡] - Arnaud Doucet[†*]

‡Statistical Laboratory, DPMMS,

University of Cambridge, 16 Mill Lane,

Cambridge, CB2 1SB, UK.

Email: d.crisan@statslab.cam.ac.uk

[†]Signal Processing Group, Engineering Department,

University of Cambridge, Trumpington Street,

Cambridge CB2 1PZ, UK.

Email: ad2@eng.cam.ac.uk

[*]corresponding author.

Abstract

Bayesian estimation problems where the posterior distribution evolves over time through the accumulation of data arise in many applications in statistics and related fields. Recently, a large number of algorithms and applications based on sequential Monte Carlo methods (also known as particle filtering methods) have appeared in the literature to solve this class of problems; see (Doucet, de Freitas & Gordon, 2001) for a survey. However, few of these methods have been proved to converge rigorously. The purpose of this paper is to address this issue.

We present a general sequential Monte Carlo (SMC) method which includes most of the important features present in current SMC methods. This method generalizes and encompasses many recent algorithms. Under mild regularity conditions, we obtain rigorous convergence results for this general SMC method and therefore give theoretical backing for the validity of all the algorithms that can be obtained as particular cases of it.

Keywords: Bayesian estimation, Filtering, Markov chain Monte Carlo, Nonlinear non Gaussian state space model, Sequential Monte Carlo.

1 Introduction

Many real-world data analysis tasks involve estimating unknown quantities when only partial or inaccurate observations are available. In most of these applications, prior knowledge of the underlying system is available. This knowledge allows us to adopt a Bayesian approach to the problem; that is, to start with a prior distribution for the unknown quantities and a likelihood function relating these quantities to the observations. Within this setting, one performs inference on the unknown quantities based on the posterior distribution. Often, the observations arrive sequentially in time and one is interested in *estimating recursively in time* the evolving posterior distribution. This problem is known sometimes as the Bayesian or optimal filtering problem. It has many applications in statistics and related areas such as automatic control, econometrics, machine learning and signal processing.

Except in a few special cases, including linear Gaussian state space models, the posterior distributions cannot be obtained analytically. For over thirty years, many approximation schemes, such as the extended Kalman filter and Gaussian sum approximations, have been proposed to surmount this problem (West & Harrison, 1997). However, in many realistic problems, these approximating methods are unreliable. Standard iterative algorithms such as Markov chain Monte Carlo (MCMC) methods are also ineffective as the posterior distribution evolves over time (Robert & Casella, 1999).

Recently, there has been a surge of interest in SMC methods for solving sequential Bayesian estimation problems. These methods utilize a random sample (or particle) based representation of the posterior probability distributions. The particles are propagated over time using a combination of sequential importance sampling, selection/resampling and MCMC steps. Following the paper of Gordon, Salmon and Smith (1993) introducing the so-called bootstrap filter, see (Kitagawa, 1996; Isard & Blake, 1998; West, 1993a; West, 1993b) for related early work, many improved SMC methods have been proposed; see for example (Berzuini, Best, Gilks & Larizza, 1997; Doucet, Godsill & Andrieu, 2000; Doucet, de Freitas & Gordon, 2001; Gilks &

Berzuini, 1999; Hürzeler & Künsch, 1998; Liu & Chen, 1998; Pitt & Shephard, 1999; Tanizaki, 1996). These algorithms mainly rely on improved importance sampling/resampling strategies and MCMC steps.

Of all the algorithms available, the most extensively studied is the bootstrap filter (Gordon, Salmond & Smith, 1993), also known as the interacting particle systems/resolution algorithm; see (Del Moral, 1996; Del Moral, 1997) and subsequent papers. In (Crisan, Del Moral & Lyons, 1999), a rigorous treatment is given to a whole class of SMC methods. This class of methods does not however include most of the characteristics of standard methods commonly used by practitioners (Doucet, de Freitas & Gordon, 2001).

The aim of this paper is to present and study rigourously a general SMC method that includes most of these characteristics. The following are the main new features presented in this paper:

- We remove the standard assumptions that the signal process is Markov and that the observations are conditionally independent upon the signal in order to include the methods presented in (Liu & Chen, 1998).

- The importance sampling step is done using a general transition kernel which can depend on both the observations and the current Monte Carlo approximation of the posterior distribution.

- The conditions imposed on the resampling step are less restrictive - for example, the unbiasedness assumption is lifted.

- An MCMC step is added after the resampling procedure in order to address the problem of sample depletion. This method, originally proposed by Gilks & Berzuini (1999), is generalized by allowing the MCMC kernel to depend on the whole population of particles.

- The convergence results are given on the path space, that is, we prove the convergence not to the posterior distribution of the current state of the signal, but to the posterior distribution of the whole trajectory of the signal.

Under suitable regularity conditions on the dynamic model of interest, the importance sampling function, the selection scheme and the MCMC kernel used, we prove convergence of the empirical distribution of particles produced by the SMC method to the posterior distribution of the signal as the number of particles increases. The techniques used in this paper are close in spirit to those in (Crisan, Del Moral & Lyons, 1999), although the aims and the results are different. For a newcomer to this field, we recommend reading (Crisan, 2001) first.

As stated in the abstract, many algorithms presented in the literature (Carpenter, Clifford & Fearnhead, 1999; Chen & Liu, 2000; Doucet, Godsill & Andrieu, 2000; Gilks & Berzuini, 1999; Gordon, Salmond & Smith, 1993; Liu & Chen, 1998) are particular cases of the method presented here and, therefore, receive rigorous treatment in this paper. It is worth noting that a straightforward adaptation of those proofs would also allow rigourous convergence results to be obtained for other SMC algorithms such as the auxiliary particle filter (Pitt & Shephard, 1999) and also for some Monte Carlo algorithms developed in statistical physics such as Quantum Monte Carlo and Transfer Matrix Monte Carlo; see (Iba, 2000) for a brief description of these methods and their connections to the ones presented here.

The rest of the paper is organized as follows: In Section 2, the model and estimation objectives are given. In Section 3, the SMC algorithm is described and its different steps are detailed. The links with previous methods are briefly discussed. In Section 4, simple sufficient conditions are given to ensure asymptotic convergence of the empirical measures toward their true values. The rate of convergence of the average mean square error is also established. In Section 5, we apply these results to an algorithm developed to perform speech enhancement. All the proofs are given in the Appendix.

2 Problem Statement

Let (Ω, F, P) be a probability space on which we have defined two vector-valued stochastic processes $X = \{X_t,\ t \in \mathbb{N}\}$ and $Y = \{Y_t,\ t \in \mathbb{N}^*\}$. The process X is usually called the *signal* process and the process Y is called the *observation* process. Let n_x and n_y be the dimensions of the state space of X and of Y. We will denote by $X_{i:j}$ and $Y_{i:j}$ the path of the signal and of the observation process from time i to time j,

$$X_{i:j} \triangleq (X_i, X_{i+1}, ..., X_j), \quad Y_{i:j} \triangleq (Y_i, Y_{i+1}, ..., Y_j).$$

Let $x_{i:j}$ and $y_{i:j}$ be generic points in the space of paths of the signal and observation processes,

$$x_{i:j} \triangleq (x_i, x_{i+1}, ..., x_j) \in (\mathbb{R}^{n_x})^{j-i+1}, \; y_{i:j} \triangleq (y_i, y_{i+1}, ..., y_j) \in (\mathbb{R}^{n_y})^{j-i+1}.$$

The *signal* process X satisfies $X_0 \sim \pi_0(dx_0)$ and evolves according to the following equation,

$$\Pr(X_t \in A_t | Y_{0:t-1} = y_{0:t-1}, X_{0:t-1} = x_{0:t-1}) = \int_{A_t} k_t(y_{0:t}, x_{0:t-1}, dx_t), \; A_t \in \mathcal{B}(\mathbb{R}^{n_x}) \tag{1}$$

where $k_t(y_{0:t}, x_{0:t-1}, dx_t)$ is a probability transition kernel

$$\begin{aligned} k_t \quad &: \quad \mathcal{P}\left((\mathbb{R}^{n_x})^t\right) \longrightarrow \mathcal{P}\left((\mathbb{R}^{n_x})^{t+1}\right) \\ (k_t\mu)(dx_{0:t}) \quad &= \quad k_t(y_{0:t}, x_{0:t-1}, dx_t)\,\mu(dx_{0:t-1}). \end{aligned}$$

$\mathcal{P}\left((\mathbb{R}^{n_x})^{t+1}\right)$ and $\mathcal{P}\left((\mathbb{R}^{n_x})^{t}\right)$ respectively represent the set of probability measures defined on $\mathcal{B}\left((\mathbb{R}^{n_x})^{t+1}\right)$, the Borel $\sigma-$algebra on $(\mathbb{R}^{n_x})^{t+1}$ and on $\mathcal{B}\left((\mathbb{R}^{n_x})^{t}\right)$, the Borel $\sigma-$algebra on $(\mathbb{R}^{n_x})^{t}$. The *observation* process Y satisfies

$$\Pr(Y_t \in B_t | Y_{0:t-1} = y_{0:t-1}, X_{0:t} = x_{0:t}) = \int_{B_t} g_t(y_{0:t}, x_{0:t})\, dy_t, \; B_t \in \mathcal{B}(\mathbb{R}^{n_y}). \tag{2}$$

We assume that the sequence $(\pi_t, \rho_t)_{t\in\mathbb{N}}$ satisfies the following recurrence formula:

Bayes' recursion. For all $t \geq 0$ and $A_i \in \mathcal{B}(\mathbb{R}^{n_x})$, $i = 0, ..., t$, $A_{0:t} = A_0 \times \cdots \times A_t$, we have

$$\begin{aligned} &\text{Prediction} \quad \rho_t(A_{0:t}) = \textstyle\int_{A_{0:t-1}} k_t(y_{0:t}, x_{0:t-1}, A_t)\,\pi_{t-1}(dx_{0:t-1}) \\ &\text{Updating} \quad \pi_t(A_{0:t}) = C_t^{-1} \textstyle\int_{A_{0:t}} g_t(y_{0:t}, x_{0:t})\,\rho_t(dx_{0:t}), \end{aligned}$$

where C_t is a normalizing constant $C_t \triangleq \int_{(\mathbb{R}^{n_x})^{t+1}} g_t(y_{0:t}, x_{0:t})\,\rho_t(dx_{0:t})$.

If μ is a measure, f is a function and K is a Markov kernel, we use the following standard notation

$$(\mu, f) \triangleq \int f d\mu,\ \mu K(A) \triangleq \int \mu(dx) K(A|x),\ Kf(x) \triangleq \int K(dz|x) f(z).$$

Using this notation, if $f_t : (\mathbb{R}^{n_x})^{t+1} \to \mathbb{R}$, then the recurrence formula implies that, for all $t \in \mathbb{N}$,

$$\begin{aligned} &\text{Prediction} \quad (\rho_t, f_t) = (\pi_{t-1}, k_t f_t) \\ &\text{Updating} \quad (\pi_t, f_t) = (\rho_t, f_t g_t)(\rho_t, g_t)^{-1}. \end{aligned} \tag{3}$$

Remark 1 *Let X be a Markov process with respect to the filtration $\mathcal{F}_t \triangleq \sigma(X_s, Y_s, s \in \{0, \ldots, t\})$ with transition kernel*

$$h_t(A, x_{t-1}) \triangleq P(X_t \in A | X_{t-1} = x_{t-1}), \quad A \in \mathcal{B}(\mathbb{R}^{n_x}),\ x_{t-1} \in \mathbb{R}^{n_x}.$$

Also, assume that $P\left(Y_t \in dy_t | \mathcal{F}_t^-\right) = P(Y_t \in dy_t | X_t)$, where $\mathcal{F}_t^- \triangleq \sigma(X_s, Y_s, s \in \{0, \ldots, t-1\}, X_t)$ and for all $x_t \in \mathbb{R}^{n_x}$, the conditional distribution of Y_t given the event $\{X_t = x_t\}$ is absolutely continuous with respect to the Lebesgue measure, that is there exists $i(y_t, x_t)$ such that for any $x_{t-1} \in \mathbb{R}^{n_x}$

$$P(Y_t \in dy_t | X_t = x_t) = i(y_t, x_t)\, dy_t.$$

*Then the sequence $(\pi_t, \rho_t)_{t \in \mathbb{N}}$ satisfies the **Bayes' recursion** condition with $k_t \triangleq h_t$ and $g_t \triangleq i_t$.*

Except for a very restricted number of dynamic models, it is impossible to evaluate the joint posterior distribution π_t in a closed form expression. In the next section, we present a SMC method to solve this problem.

3 A Sequential Monte Carlo Method

3.1 Algorithm

A SMC method is a recursive algorithm which produces, at each time t, a cloud of particles whose empirical measure closely "follows" the distribution π_t. The particles reside in the state space of π_t, hence in the space

$(\mathbb{R}^{n_x})^{t+1}$. Note that the dimension of the space where the particles live increases in time: at time t it is equal to $(t+1)\,n_x$. One can view the particles as (discrete) paths of length $t+1$ units in $\mathbb{R}^{n_x}$ (the signal's space).

We describe a general algorithm which generates at time t (for all $t>0$) N particles/paths $\left\{x_{0:t}^{(i)}\right\}_{i=1}^{N}$ with an associated empirical measure π_t^N

$$\pi_t^N\left(dx_{0:t}\right) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_{0:t}^{(i)}}\left(dx_{0:t}\right)$$

that is "close" to π_t. The algorithm is recursive in the sense that $\left\{x_{0:t}^{(i)}\right\}_{i=1}^{N}$ is produced using the observation obtained at time t and the previous set of particles $\left\{x_{0:t-1}^{(i)}\right\}_{i=1}^{N}$ produced at time $t-1$ (whose empirical measure π_{t-1}^N was "close" to π_{t-1}).

We also introduce a transition kernel $\Xi_t\left(y_{0:t}, x_{0:t-1}, \pi_{t-1}, dx_{0:t}\right)$ which is used in order to obtain an intermediate set of particles $\tilde{x}_{0:t}^{(i)}$. We denote by $\widetilde{\rho}_t$ the resulting *importance distribution*

$$\widetilde{\rho}_t \triangleq \pi_{t-1}\Xi_t.$$

We assume that ρ_t is absolutely continuous with respect to $\widetilde{\rho}_t$ and let h_t be the strictly positive Radon Nykodym derivative $\frac{d\rho_t}{d\tilde{\rho}_t}=h_t$, where $h_t(\cdot)=h_t\left(y_{0:t},\pi_{t-1},\cdot\right)$. Then π_t is also absolutely continuous with respect to $\tilde{\rho}_t$ and

$$\frac{d\pi_t}{d\tilde{\rho}_t} \propto g_t h_t. \tag{4}$$

We also assume that we can sample exactly from π_0 at $t=0$. The algorithm proceeds as follows.

At time $t=0$.

Step 0: <u>*Initialization*</u>

- For $i=1,...,N$, sample $x_0^{(i)} \sim \pi_0\left(dx_0\right)$ and set $t=1$.

At time $t \geq 1$,

Step 1: <u>*Importance Sampling step*</u>

- For $i = 1, ..., N$, sample $\widetilde{x}_{0:t}^{(i)} \sim \Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_{0:t}\right)$.

- For $i = 1, ..., N$, evaluate the normalized importance weights $\overline{w}_t^{(i)}$

$$\overline{w}_t^{(i)} \propto g_t\left(y_{0:t}, \widetilde{x}_{0:t}^{(i)}\right) h_t\left(y_{0:t}, \pi_{t-1}^N, \widetilde{x}_{0:t}^{(i)}\right) \tag{5}$$

and let $\widetilde{\rho}_t^N$ and $\overline{\pi}_t^N$ be the measures

$$\widetilde{\rho}_t^N(dx_{0:t}) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{x}_{0:t}^{(i)}}(dx_{0:t}), \tag{6}$$

$$\overline{\pi}_t^N(dx_{0:t}) \triangleq \sum_{i=1}^N \overline{w}_t^{(i)} \delta_{\widetilde{x}_{0:t}^{(i)}}(dx_{0:t}). \tag{7}$$

Step 2: _Selection step_

- Multiply/Discard particles $\left\{\widetilde{x}_{0:t}^{(i)}\right\}_{i=1}^N$ with respect to high/low importance weights $\overline{w}_t^{(i)}$ to obtain N particles $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N$. Let $\ddot{\pi}_t^N$ denote the associated empirical measure

$$\ddot{\pi}_t^N(dx_{0:t}) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\ddot{x}_{0:t}^{(i)}}(dx_{0:t}). \tag{8}$$

Step 3: _MCMC step_

- For $i = 1, ..., N$, sample $x_{0:t}^{(i)} \sim K_t\left(\left\{\ddot{x}_{0:t}^{(j)}\right\}_{j=1}^N, dx_{0:t}\right)$ where $K_t(\cdot,\cdot)$ is a Markov kernel of invariant distribution $\pi_t(dx_{0:t})$. Let π_t^N denote the associated empirical measure

$$\pi_t^N(dx_{0:t}) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x_{0:t}^{(i)}}(dx_{0:t}). \tag{9}$$

- Set $t \leftarrow t+1$ and go to **Step 1**.

3.2 Implementation Issues

We now detail the various steps of the method.

Importance Sampling Step

The new set of particles/paths is obtained by sampling from $\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, dx_{0:t}\right)$, which, obviously, depends on π_{t-1}^N, the observations $y_{0:t}$ and on the current paths $\left\{x_{0:t-1}^{(i)}\right\}_{i=1}^N$. The new paths $\left\{\widetilde{x}_{0:t}^{(i)}\right\}_{i=1}^N$ are then distributed approximately according to $\widetilde{\rho}_t\left(d\tilde{x}_{0:t}\right)$; $\widetilde{\rho}_t\left(d\tilde{x}_{0:t}\right)$ being chosen to be "as close as possible" to $\pi_t\left(d\tilde{x}_{0:t}\right)$. There are unlimited choices for $\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}, d\widetilde{x}_{0:t}\right)$, the only condition is that the weights $\overline{w}_t^{(i)}$ given by (5) are well-defined and can be computed analytically.

Most of the algorithms presented in the literature are such that

$$\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_{0:t}\right) = \delta_{x_{0:t-1}^{(i)}}\left(d\widetilde{x}_{0:t-1}\right) \Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_t\right),$$

that is, the new path $\widetilde{x}_{0:t}^{(i)}$ is obtained by keeping the current path $x_{0:t-1}^{(i)}$ and extending it by adding a new point $\widetilde{x}_t^{(i)}$, where $\widetilde{x}_t^{(i)}$ is sampled from $\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_t\right)$. We now discuss several possible choices for $\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_t\right)$. A sensible selection criterion is to choose a proposal that minimises the conditional variance of the importance weights at time t, given $x_{0:t-1}^{(i)}$ and $y_{0:t}$. According to this strategy, it has been shown that the optimal distribution is $P\left(d\widetilde{x}_t \middle| y_{0:t}, x_{0:t-1}^{(i)}\right)$ (Doucet, Godsill & Andrieu, 2000).

- *Optimal sampling distribution.* If we sample $\widetilde{x}_t^{(i)}$ according to

$$P\left(d\widetilde{x}_t \middle| y_{0:t}, x_{0:t-1}^{(i)}\right) \propto g_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right) k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, d\widetilde{x}_t\right)$$

then the importance weight is equal to

$$\overline{w}_t^{(i)} \propto \int g_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right) k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, d\widetilde{x}_t\right).$$

If this integral does not admit an analytical expression, one has to use an alternative method.

- *Prior distribution.* If we use the prior distribution $k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, d\widetilde{x}_t\right)$ as importance distribution, the importance weight is proportional to $g_t\left(x_{0:t}^{(i)}, \widetilde{x}_t^{(i)}, y_{0:t}\right)$. However, as underlined in (Pitt & Shephard, 1999), this strategy is sensitive to outliers.

- *Likelihood distribution.* Assume the likelihood $g_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right)$ is integrable in argument $\widetilde{x}_t$, that is

$$\int g_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right) d\widetilde{x}_t < \infty,$$

then one can sample $\widetilde{x}_t^{(i)}$ according to

$$\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_t\right) \propto g_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right)$$

and

$$\overline{w}_t^{(i)} \propto k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \widetilde{x}_t\right).$$

This is useful when the observation noise is very low as the likelihood is then usually very peaked compared to the prior distribution $k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, d\widetilde{x}_t\right)$; see (Fox, Thrun, Burgard & Dellaert, 2001) for an alternative algorithm to handle a low observation noise level.

- *Alternative sampling distribution.* It is possible to design a variety of alternative sampling distributions (Doucet, Godsill & Andrieu, 2000; Pitt & Shephard, 1999). For example, one can use the results of a suboptimal deterministic algorithm to construct an importance sampling distribution $\Xi_t\left(y_{0:t}, x_{0:t-1}^{(i)}, \pi_{t-1}^N, d\widetilde{x}_t\right)$. This is very useful in applications where $P\left(d\widetilde{x}_t \middle| y_{0:t}, x_{0:t-1}^{(i)}\right)$ is too expensive or impossible to use and $k_t\left(y_{0:t}, x_{0:t-1}^{(i)}, d\widetilde{x}_t\right)$ is inefficient.

Resampling Step

The aim of the resampling/selection step is to obtain an "unweighted" empirical distribution approximation $\ddot{\pi}_t^N\left(dx_{0:t}\right)$ of the weighted measure $\overline{\pi}_t^N\left(dx_{0:t}\right)$. A selection procedure associates a number of offspring $N_t^{(i)} \in \mathbb{N}$ with each particle $\left\{\widetilde{x}_{0:t}^{(i)}\right\}_{i=1}^N$, where $\sum_{i=1}^N N_t^{(i)} = N$, to obtain N new particles $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N$ with associated empirical distribution $\ddot{\pi}_t^N$, where

$$\ddot{\pi}_t^N\left(dx_{0:t}\right) = N^{-1}\sum_{i=1}^N N_t^{(i)} \delta_{\widetilde{x}_{0:t}^{(i)}}\left(dx_{0:t}\right) = N^{-1}\sum_{i=1}^N \delta_{\ddot{x}_{0:t}^{(i)}}\left(dx_{0:t}\right).$$

We briefly describe three classes of selection schemes. All can be implemented in $O\left(N\right)$ iterations.

- *Sampling Importance Resampling (SIR)/Multinomial Sampling procedure.* This procedure, introduced originally by (Gordon, Salmond & Smith, 1993), is the most popular. One samples N times from $\widetilde{\pi}_t^N\left(dx_{0:t}\right)$ to obtain $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N$. This is equivalent to jointly drawing $\left\{N_t^{(i)}\right\}_{i=1}^N$ according to a multinomial distribution of parameters N and $\overline{w}_t^{(i)}$. It is possible to implement the SIR procedure exactly in $O(N)$ operations using a classical algorithm (Ripley, 1987). In this case, we have $E\left(N_t^{(i)}\right) = N\overline{w}_t^{(i)}$ and $var\left(N_t^{(i)}\right) = N\overline{w}_t^{(i)}\left(1-\overline{w}_t^{(i)}\right)$.

• *Residual Resampling* (Higuchi, 1997; Liu & Chen, 1998). Set $\widetilde{N}_t^{(i)} = \left\lfloor N\overline{w}_t^{(i)} \right\rfloor$, then perform a SIR procedure to select the remaining $\overline{N}_t = N - \sum_{i=1}^{N} \widetilde{N}_t^{(i)}$ samples with the new weights $w_t'^{(i)} = \overline{N}_t^{-1}\left(\overline{w}_t^{(i)} N - \widetilde{N}_t^{(i)}\right)$. Finally, add the results to the current $\widetilde{N}_t^{(i)}$. In this case, we again obtain $E\left(N_t^{(i)}\right) = N\overline{w}_t^{(i)}$ but $var\left(N_t^{(i)}\right) = \overline{N}_t w_t'^{(i)}\left(1-w_t'^{(i)}\right)$.

• *Minimal Variance Sampling.* This class includes the stratified/systematic sampling procedures introduced in (Carpenter, Clifford & Fearnhead, 1999; Kitagawa, 1996) where a set U of N points is generated in the interval $[0,1]$, each of the points a distance N^{-1} apart. The number $N_t^{(i)}$ is taken to be the number of points in U that lie between $\sum_{j=1}^{i-1} \overline{w}_t^{(j)}$ and $\sum_{j=1}^{i} \overline{w}_t^{(j)}$. It also includes the Tree Based Branching algorithm presented in (Crisan, 2001). If we denote $\left\{N\overline{w}_t^{(i)}\right\} \triangleq N\overline{w}_t^{(i)} - \left\lfloor N\overline{w}_t^{(i)} \right\rfloor$, then the variance of all the algorithms in this class is $var\left(N_t^{(i)}\right) = \left\{N\overline{w}_t^{(i)}\right\}\left(1-\left\{N\overline{w}_t^{(i)}\right\}\right)$.

Actually, as we will discuss later on, it is not necessary to design unbiased selection schemes, that is we can have $E\left(N_t^{(i)}\right) \neq N\overline{w}_t^{(i)}$. An example of a biased selection scheme is the deterministic selection scheme proposed by Kitagawa (1996).

MCMC Step

If the distribution of the importance weights is highly skewed then the particles $\widetilde{x}_{0:t}^{(i)}$ which have high importance weights $\overline{w}_t^{(i)}$ are selected many times and, thus, numerous particles $\ddot{x}_{0:t}^{(i)}$ will have identical positions. This amounts to a depletion of samples. One way to perform sample regeneration consists of considering a mixture approximation (Gordon, Salmond & Smith, 1993; Liu & West, 2001). An alternative method based on an MCMC procedure has recently been proposed in (Gilks & Berzuini, 1999). The rationale behind the use of the MCMC method is based on the following remark. Assume that the particles $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^{N}$ are distributed marginally according to $\pi_t(dx_{0:t})$. Then, if we apply to each particle $\ddot{x}_{0:t}^{(i)}$ a Markov transition kernel $K_t\left(\ddot{x}_{0:t}^{(i)}, dx_{0:t}\right)$ of invariant distribution $\pi_t(dx_{0:t})$, the new particles $\left\{x_{0:t}^{(i)}\right\}_{i=1}^{N}$ are still distributed according to the posterior distribution of interest. One can use any of the standard MCMC methods (Metropolis-Hastings, Gibbs sampler, etc.). Moreover, unlike for standard MCMC applications (Robert & Casella, 1999), we do not require that the kernel is ergodic[1]. For example, one can use a Gibbs sampling step which updates at time t the values of the process x_t from time $t-L+1$ to t, i.e., we sample $x_k^{(i)}$ for $k=t-L+1,\ldots,t$ according to $P(dx_k|\, Y_{0:t}, x_{0:k-1}^{(i)}, x_{k+1:t}^{(i)})$.

Further, one can allow the Markov transition kernel to be dependent on the whole population of particles $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^{N}$, provided it satisfies

$$\int K_t\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^{N}, dx_{0:t}\right)\prod_{i=1}^{N}\pi_t\left(d\ddot{x}_{0:t}^{(i)}\right) = \pi_t(dx_{0:t}).$$

This way one can use the information provided by the whole set of particles so as to design, for example, an efficient proposal for a Metropolis-Hastings sampler.

4 Convergence Study

Let $B(\mathbb{R}^n)$ and $C_b(\mathbb{R}^n)$ respectively be the space of bounded, Borel measurable functions and the space of bounded continuous functions on $\mathbb{R}^n$. We denote $\|f\| \triangleq \sup_{x\in\mathbb{R}^n}|f(x)|$. In this section, we first establish the convergence (and the rate of convergence) to 0 of the average mean square error $E\left[\left(\left(\pi_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right]$ for any $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$ under quite general conditions (where the expectation is over all the realizations of the random particle methods). Then we prove the (almost sure) convergence of π_t^N towards π_t under more restrictive conditions.

From now on we will assume that the observation process is fixed to a given observation record $Y_t = y_t$, $t > 0$. All the convergence results will be proven under this condition.

[1]Note that it does not need to be Markov either but in practice it is usually limited to this case.

4.1 Bounds for mean square errors

Let us consider the following assumptions.

Assumptions.

1. -A. *Importance distribution and weights.*

- ρ_t is assumed absolutely continuous with respect to $\widetilde{\rho}_t = \pi_{t-1}\Xi_t$, and for all $\mu \in \mathcal{P}\left((\mathbb{R}^{n_x})^t\right)$ the function $g_t\left(\widetilde{x}_{0:t}, y_{0:t}\right) h_t\left(\widetilde{x}_{0:t}, y_{0:t}, \mu\right)$ is a <u>bounded function</u> in argument $x_{0:t} \in (\mathbb{R}^{n_x})^{t+1}$.

The identity (4) becomes for all $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$(\pi_t, f_t) = \frac{(\tilde{\rho}_t, f_t g_t h_t)}{(\tilde{\rho}_t, g_t h_t)},$$

where $g_t(\cdot) = g_t(y_{0:t}, \cdot)$ and $h_t(\cdot) = h_t(y_{0:t}, \pi_{t-1}, \cdot)$. If $\mu,\, \nu \in \mathcal{P}\left((\mathbb{R}^{n_x})^t\right)$, we define

$$\begin{aligned}
\Xi_t^\mu &\triangleq \Xi_t\left(y_{0:t}, x_{0:t-1}, \mu, d\tilde{x}_{0:t}\right), \\
\Xi_t^\nu &\triangleq \Xi_t\left(y_{0:t}, x_{0:t-1}, \nu, d\tilde{x}_{0:t}\right), \\
h_t^\mu(\cdot) &= h_t\left(y_{0:t}, \mu, \cdot\right), \\
h_t^\nu(\cdot) &= h_t\left(y_{0:t}, \nu, \cdot\right).
\end{aligned}$$

- There exists a constant d_t such that, for all $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$, there exists $f_{t-1} \in B\left((\mathbb{R}^{n_x})^t\right)$ with $||f_{t-1}|| \leq ||f_t||$ such that

$$||\Xi_t^\mu f_t - \Xi_t^\nu f_t|| \leq d_t \left|(\mu, f_{t-1}) - (\nu, f_{t-1})\right| \tag{10}$$

- There exist f^h (independent of μ, ν) such that

$$||g_t h_t^\mu - g_t h_t^\nu|| \leq \left|\left(\mu, f^h\right) - \left(\nu, f^h\right)\right| \tag{11}$$

  and a constant e_t such that for any $x_{0:t} \in (\mathbb{R}^{n_x})^{t+1}$

$$|h_t^\mu(x_{0:t}) - h_t^\nu(x_{0:t})| \leq e_t \min\left(h_t^\mu(x_{0:t}), h_t^\nu(x_{0:t})\right). \tag{12}$$

2. -A. *Resampling/Selection scheme.* $\left\{N_t^{(i)}\right\}_{i=1}^N$ are integer valued random variables such that

$$E\left[\left|\sum_{i=1}^N \left(N_t^{(i)} - N\overline{w}_t^{(i)}\right) q^{(i)}\right|^2\right] \leq C_t N \max_{i=1,..N} \left|q^{(i)}\right|^2 \tag{13}$$

for all N-dimensional vectors $q = \left(q^{(1)}, q^{(2)}, ..., q^{(N)}\right) \in \mathbb{R}^N$ and $\sum_{i=1}^N N_t^{(i)} = N$.

The first assumption ensures that the importance function is chosen so that the corresponding importance weights are bounded above and that the sampling kernel and the importance weights depend "continuously" on the measure variable. The second assumption ensures that the selection scheme does not introduce too strong a "discrepancy".

The following lemmas essentially state that, at each step of the particle filtering algorithm, the approximation produced admits a mean square error of order $1/N$.

Lemma 1 *Let us assume that for any* $f_{t-1} \in B\left((\mathbb{R}^{n_x})^t\right)$

$$E\left[\left(\left(\pi_{t-1}^N, f_{t-1}\right) - \left(\pi_{t-1}, f_{t-1}\right)\right)^2\right] \leq c_{t-1} \frac{\|f_{t-1}\|^2}{N},$$

then, after ***Step 1*** *of the algorithm, for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$,

$$E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - \left(\widetilde{\rho}_t, f_t\right)\right)^2\right] \leq \widetilde{c}_t \frac{\|f_t\|^2}{N}.$$

Lemma 2 *Let us assume that for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\pi_{t-1}^N, f_{t-1}\right) - \left(\pi_{t-1}, f_{t-1}\right)\right)^2\right] \leq c_{t-1} \frac{\|f_{t-1}\|^2}{N},$$

$$E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - \left(\widetilde{\rho}_t, f_t\right)\right)^2\right] \leq \widetilde{c}_t \frac{\|f_t\|^2}{N},$$

then for any $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\overline{\pi}_t^N, f_t\right) - \left(\pi_t, f_t\right)\right)^2\right] \leq \overline{c}_t \frac{\|f_t\|^2}{N}.$$

Lemma 3 *Let us assume that for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\overline{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right] \leq \overline{c}_t \frac{\|f_t\|^2}{N},$$

*then, after **Step 2** of the algorithm, there exists a constant* $\ddot{c}_t$ *such that, for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\ddot{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right] \leq \ddot{c}_t \frac{\|f_t\|^2}{N}.$$

Lemma 4 *Let us assume that for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\ddot{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right] \leq \ddot{c}_t \frac{\|f_t\|^2}{N},$$

*then, after **Step 3** of the algorithm, for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\pi_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right] \leq c_t \frac{\|f_t\|^2}{N}.$$

By putting together lemmas 1, 2, 3 and 4, we obtain

Theorem 1 *For all* $t \geq 0$*, there exists* c_t *independent of* N *such that for any* $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$E\left[\left(\left(\pi_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right] \leq c_t \frac{\|f_t\|^2}{N}.$$

4.2 Convergence of empirical measures

Let $\mathcal{P}\left(\mathbb{R}^d\right)$ be the set of probability measures over $\mathcal{B}\left(\mathbb{R}^d\right)$. We endow $\mathcal{P}\left(\mathbb{R}^d\right)$ with the weak topology. i.e., if $(\mu_N)_{N=1}^{\infty}$ is a sequence of probability measures, then μ_N converges to $\mu \in \mathcal{P}\left(\mathbb{R}^d\right)$ in the weak topology ($\lim_{N\to\infty} \mu_N = \mu$), if for any $f \in C_b\left(\mathbb{R}^d\right)$

$$\lim_{N\to\infty} (\mu_N, f) = (\mu, f).$$

In our case, the particle filtering algorithm generates a sequence of measure valued random variables and N represents the number of particles used in the approximating particle system. In this section, we show that $\lim_{N\to\infty} \pi^N = \pi_t$ almost surely under the following assumptions.

Assumptions.

1. -B. *Importance distribution and weights.*

- $\Xi_t\left(y_{0:t}, x_{0:t-1}, \pi_{t-1}, d\widetilde{x}_{0:t}\right)$ is a Feller kernel.

- ρ_t is absolutely continuous with respect to $\widetilde{\rho}_t = \pi_{t-1}\Xi_t$ and, for any $\mu \in \mathcal{P}\left(\left(\mathbb{R}^{n_x}\right)^t\right)$, $g_t\left(x_{0:t}, y_{0:t}\right) h_t\left(x_{0:t}, y_{0:t}, \mu\right)$ is a bounded continuous function.

- If μ, $\nu \in \mathcal{P}\left(\left(\mathbb{R}^{n_x}\right)^t\right)$, there exists a constant d_t such that, for all $f_t \in B\left(\left(\mathbb{R}^{n_x}\right)^{t+1}\right)$, there exists $f_{t-1} \in B\left(\left(\mathbb{R}^{n_x}\right)^t\right)$ with$||f_{t-1}|| \leq ||f_t||$ such that

$$||\Xi_t^\mu f_t - \Xi_t^\nu f_t|| \leq d_t \left|\left(\mu, f_{t-1}\right) - \left(\nu, f_{t-1}\right)\right|.$$

- There exists f^h (independent of μ, ν) such that

$$||g_t h_t^\mu - g_t h_t^\nu|| \leq \left|\left(\mu, f^h\right) - \left(\nu, f^h\right)\right|.$$

2. -B. *Selection scheme.* $N_t^{(i)}$ are integer valued random variables such that there exists $p > 1$ and $\alpha < p-1$ such that

$$E\left[\left|\sum_{i=1}^N \left(N_t^{(i)} - N\overline{w}_t^{(i)}\right) q^{(i)}\right|^p\right] \leq CN^\alpha \max_{i=1,..N}\left|q^{(i)}\right|^p \tag{14}$$

for all N-dimensional vectors $q = \left(q^{(1)}, q^{(2)}, ..., q^{(N)}\right)$ and $\sum_{i=1}^N N_t^{(i)} = 1$.

3. -B. *MCMC kernel.* $K_t\left(\left(\ddot{x}_{0:t}^{(1)}, \ldots, \ddot{x}_{0:t}^{(N)}\right), d\left(x_{0:t}^{(1)}, \ldots, x_{0:t}^{(N)}\right)\right)$ is a Feller kernel.

The additional Feller property and continuity of the importance weight function are required to prove convergence in the weak topology.

The following lemmas, put together, give us the almost sure convergence of the particle filter for all $t \geq 0$.

Lemma 5 *Let π_{t-1}^N be a sequence of random approximations of π_{t-1} such that, almost surely,*

$$\lim_{N\to\infty} \pi_{t-1}^N = \pi_{t-1}.$$

*Then, after **Step 1** of the algorithm, almost surely*

$$\lim_{N\to\infty} \tilde{\rho}_t^N = \tilde{\rho}_t.$$

Lemma 6 *Let $\widetilde{\rho}_t^N$ be a sequence of random approximations of ρ_t such that, almost surely,*

$$\lim_{N\to\infty} \widetilde{\rho}_{t-1}^N = \widetilde{\rho}_{t-1}.$$

*Then, after **Step 2** of the algorithm, almost surely*

$$\lim_{N\to\infty} \ddot{\pi}_t^N = \pi_t.$$

Lemma 7 *Let $\ddot{\pi}_t^N$ be a sequence of random approximations of ρ_t such that, almost surely,*

$$\lim_{N\to\infty} \ddot{\pi}_t^N = \pi_t.$$

*then, after **Step 3** of the algorithm, almost surely,*

$$\lim_{N\to\infty} \pi_t^N = \pi_t.$$

By putting together lemmas 5, 6 and 7 and observing that, almost surely, $\lim_{N\to\infty} \pi_0^N = \pi_0$ we obtain the following

Theorem 2 *For all $t \geq 0$ we have, almost surely, $\lim_{N\to\infty} \pi_t^N = \pi_t$.*

5 Application

Here we briefly describe the application of these convergence results to a particle filtering method developed in the context of speech enhancement; see (Vermaak, Andrieu, Doucet & Godsill, 2000) for details.

5.1 Model and algorithm

The speech signal of interest is modeled using a time-varying autoregressive (TVAR) process driven by white Gaussian noise. This signal is observed in white Gaussian noise. One wants to perform speech enhancement, that is estimating the signal based on the noisy observations. More formally, one has

$$s_t = \sum_{l=1}^{P} a_{l,t} s_{t-l} + \sigma_s v_t,\ \ v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{15}$$

where $s_{-l:-1} \triangleq (0,\ldots,0)$ and the TVAR coefficients $a_t \triangleq (a_{1,t},\ldots,a_{P,t})$ are assumed to follow a standard random walk

$$a_t = a_{t-1} + \sigma_a \varepsilon_t,\ \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),\ \ a_0 \sim \mathcal{N}(m_0, P_0). \tag{16}$$

One observes

$$y_t = s_t + \sigma_w w_t,\ w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{17}$$

The hyperparameters $(m_0, P_0, \sigma_s, \sigma_a, \sigma_w)$ are assumed known.

If one sets $x_t \triangleq (a_t, s_t)$ then it is easy to see that (15)-(16) define (1) whereas (17) defines (2). In this case, X is a Markov process and the observations are conditionally independent upon X. We could apply a SMC method to estimate the posterior distributions $\pi_t(da_{0:t}, ds_{0:t})$. It is possible however to design a more efficient algorithm by focusing on the estimation of the lower-dimensional posterior distributions $\pi_t(da_{0:t})$ as, conditional upon (a_t), the model (15)-(16) is a standard linear Gaussian state-space model; see (Chen & Liu, 2000; Doucet, Godsill & Andrieu, 2000) for details. Thus, if we set $x_t \triangleq a_t$, then the transition kernel is no longer Markovian and the observations are no longer conditionally independent. However, the framework (1)-(2) still applies.

The algorithm proposed in (Vermaak, Andrieu, Doucet & Godsill, 2000) is a particular case of the general SMC method presented in Section 3: the importance distribution is the prior distribution and the resampling scheme is the stratified scheme of Kitagawa (1996). Finally we use an MCMC kernel which updates at time t the values of the process x_t from time $t-L+1$ to t, i.e., we sample $x_k^{(i)}$ for $k = t-L+1,\ldots,t$ according to $P(dx_k|\, Y_{0:t}, x_{0:k-1}^{(i)}, x_{k+1:t}^{(i)})$ using a Metropolis-Hastings (M-H) algorithm (Robert & Casella, 1999).

5.2 Convergence results

In this case, the importance weight is equal to $g_t(y_{0:t}, x_{0:t})$. It is straightforward to check that this function, corresponds to the so-called innovation of the Kalman filter, and is bounded above and continuous. Conditions (10)-(11)-(12) are verified as the importance distribution does not depend on the empirical distribution so that Assumptions 1-A and 1-B are satisfied. One can also check easily that the stratified resampling scheme satisfies Assumptions 2-A and 2-B.

Assumptions 1-A and 2-A ensure that Theorem 1 holds. In this case, Assumption 3-B is not satisfied as the M-H kernel includes a delta-Dirac mass and is thus not Feller. Hence, we cannot ensure Theorem 2. If the MCMC kernel is omitted, then this theorem is valid.

5.3 Experimental results

In Figure 1, we present two "clean" speech signals and their noisy versions corrupted artificially by some white Gaussian noise. The Signal to Noise Ratios (SNR) are respectively equal to -0.61 decibels (dB) and 6.10dB.

Figure 1 about here.

We have applied the SMC method to these speech signals. In Table 1, we display the SNR improvement in dB obtained using $N = 500$ particles when the point estimate chosen for the signal is

$$\begin{aligned}
E\left(s_t \middle| y_{1:t+L-1}\right) &= \int s_t \pi_{t+L-1}\left(ds_{0:t+L-1}, da_{0:t+L-1}\right) \\
&= \int E\left(s_t \middle| y_{1:t+L-1}, a_{0:t+L-1}\right) \pi_{t+L-1}\left(da_{0:t+L-1}\right) \\
&\simeq \int E\left(s_t \middle| y_{1:t+L-1}, a_{0:t+L-1}\right) \pi^N_{t+L-1}\left(da_{0:t+L-1}\right) \\
&\simeq \frac{1}{N}\sum_{i=1}^{N} E\left(s_t \middle| y_{1:t+L-1}, a^{(i)}_{0:t+L-1}\right),
\end{aligned}$$

where L is the length of lag delay and $E\left(s_t \middle| y_{1:t+L-1}, a^{(i)}_{0:t+L-1}\right)$ is computed through the Kalman filter.

Table 1 about here.

The SNR improvements results were obtained by averaging over 50 independent runs of the algorithms. The fixed-lag estimates, *i.e.* $L > 0$, perform significantly better than the filtering estimates. The MCMC step significantly improves the point estimate $E\left(s_t | y_{1:t+L-1}\right)$ by limiting the sample depletion problem. One can listen to the results at `http://www-svr.eng.cam.ac.uk/~jv211/audio/phd/index.html`.

6 Discussion

We have given sufficient conditions to ensure convergence of a general SMC scheme. Essentially, one only requires that the importance weights are bounded and that the selection scheme has a small discrepancy in order to ensure convergence of the average mean square error. Further, if one assumes that the importance kernel and the MCMC kernel are Feller continuous, that the importance weight function is continuous and that the error introduced in the selection scheme is controlled by an inequality such as (14), then the resulting approximating measure converges weakly to the posterior measure.

Of course, a natural question would be to ask which models and algorithms have uniform rates of convergence (in time); that is for which the constant c_t in Theorem 1 is independent of t. However, our intention here was to present convergence results valid under very weak assumptions so as to include models and algorithms used routinely by practitioners. Therefore, the set-up we chose to work with was too general to allow for these more refined uniform convergence results. However, if one imposes appropriate additional assumptions on the signal and on the observation process then such uniform rates are obtainable for the class of methods we described above; see Del Moral & Guionnet (2001) for results of this kind. We will address this issue in a subsequent paper.

7 Acknowledgments

We are thankful to Jaco Vermaak for providing us with most of the simulation results.

A Proofs of Lemmas and Theorems

Proof of Lemma 1. Let $\mathcal{G}_{t-1}$ be the σ-field generated by $\left\{x_{0:t-1}^{(i)}\right\}_{i=1}^{N}$, then

$$E\left[\left(\widetilde{\rho}_t^N, f_t\right)\Big|\, \mathcal{G}_{t-1}\right] = \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right)$$

and

$$\begin{aligned} E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - E\left[\left(\widetilde{\rho}_t^N, f_t\right)\Big|\,\mathcal{G}_{t-1}\right]\right)^2\Big|\,\mathcal{G}_{t-1}\right] &= E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right)\right)^2\Big|\,\mathcal{G}_{t-1}\right] \\ &= \frac{1}{N}\left(\left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t^2\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right)^2\right) \\ &\leq \frac{\|f_t\|^2}{N}. \end{aligned}$$

From (10) we have that

$$\begin{aligned} \left|\left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}} f_t\right)\right| &= \left|\left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t - \Xi_t^{\pi_{t-1}} f_t\right)\right| \\ &\leq \left\|\Xi_t^{\pi_{t-1}^N} f_t - \Xi_t^{\pi_{t-1}} f_t\right\| \\ &\leq d_t\left|\left(\pi_{t-1}^N, f_{t-1}\right) - \left(\pi_{t-1}, f_{t-1}\right)\right|, \end{aligned}$$

hence,

$$E\left[\left(\left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}} f_t\right)\right)^2\right] \leq d_t c_{t-1}\frac{\|f_t\|^2}{N}.$$

Then, as $\Xi_t \triangleq \Xi_t^{\pi_{t-1}}$,

$$\begin{aligned} \left|\left(\widetilde{\rho}_t^N, f_t\right) - \left(\widetilde{\rho}_t, f_t\right)\right| \leq\; & \left|\left(\widetilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N \Xi_t^{\pi_{t-1}^N}, f_t\right)\right| + \left|\left(\pi_{t-1}^N \Xi_t^{\pi_{t-1}^N}, f_t\right) - \left(\pi_{t-1}^N \Xi_t, f_t\right)\right| \\ & + \left|\left(\pi_{t-1}^N \Xi_t, f_t\right) - \left(\pi_{t-1} \Xi_t, f_t\right)\right| \end{aligned}$$

and

$$E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N, \Xi_t f_t\right)\right)^2\right]^{\frac{1}{2}} \leq E\left[\left(\left(\widetilde{\rho}_t^N, f_t\right) - \left(\Xi_t \pi_{t-1}^N, f_t\right)\right)^2\right]^{\frac{1}{2}}$$

$$
\begin{aligned}
&+E\left[\left(\left(\pi_{t-1}^N \Xi_t^{\pi_{t-1}^N}, f_t\right) - \left(\pi_{t-1}^N \Xi_t, f_t\right)\right)^2\right]^{\frac{1}{2}} \\
&+E\left[\left(\left(\pi_{t-1}^N \Xi_t, f_t\right) - \left(\pi_{t-1} \Xi_t, f_t\right)\right)^2\right]^{\frac{1}{2}} \\
\leq\ & \sqrt{\tilde{c}_t}\frac{\|f_t\|}{\sqrt{N}},
\end{aligned}
$$

where $\tilde{c}_t = \left(1 + \sqrt{d_t c_{t-1}} + \sqrt{c_{t-1}}\right)^2$. ■

Proof of Lemma 2. We have using (11) and (12) that

$$
\begin{aligned}
\left|\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)} - \frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t\right)}\right| &= \frac{\left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right| \left|\left(\widetilde{\rho}_t^N, g_t\left(h_t^{\pi_{t-1}^N} - h_t\right)\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)\left(\widetilde{\rho}_t^N, g_t h_t\right)} \\
&\leq ||f_t|| \frac{\left|\left(\widetilde{\rho}_t^N, g_t\left(h_t^{\pi_{t-1}^N} - h_t\right)\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} \\
&\leq ||f_t|| \left|\left(\widetilde{\rho}_t^N, g_t\left(h_t^{\pi_{t-1}^N} - h_t\right)\right)\right| \left|\frac{1}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} - \frac{1}{\left(\widetilde{\rho}_t, g_t h_t\right)}\right| \\
&\quad + ||f_t|| \frac{\left|\left(\widetilde{\rho}_t^N, g_t\left(h_t^{\pi_{t-1}^N} - h_t\right)\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \\
&\leq ||f_t||\, e_t \left(\widetilde{\rho}_t^N, g_t h_t\right) \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\widetilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t\right)\left(\widetilde{\rho}_t, g_t h_t\right)} \\
&\quad + ||f_t|| \frac{\left|\left(\pi_{t-1}^N, f^h\right) - \left(\pi_{t-1}, f^h\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \\
&\leq \frac{||f_t||}{\left(\widetilde{\rho}_t, g_t h_t\right)} \left(e_t \left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\widetilde{\rho}_t, g_t h_t\right)\right|\right. \\
&\quad \left. + \left|\left(\pi_{t-1}^N, f^h\right) - \left(\pi_{t-1}, f^h\right)\right|\right).
\end{aligned}
$$

Hence,

$$
E\left[\left(\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)} - \frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t\right)}\right)^2\right]^{\frac{1}{2}} \leq \frac{||f_t|| \left(e_t \sqrt{\tilde{c}_t}\, ||g_t h_t|| + \sqrt{c_{t-1}}\, \left|\left|f^h\right|\right|\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)\sqrt{N}}. \qquad (18)
$$

We have, using (12),

$$
\begin{aligned}
\left|\frac{1}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)} - \frac{1}{\left(\widetilde{\rho}_t^N, g_t h_t\right)}\right| &= \frac{\left|\left(\widetilde{\rho}_t^N, g_t\left(h_t^{\pi_{t-1}^N} - h_t\right)\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)\left(\widetilde{\rho}_t^N, g_t h_t\right)} \\
&\leq \frac{e_t\left(\widetilde{\rho}_t^N, g_t h_t\right)}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)\left(\widetilde{\rho}_t^N, g_t h_t\right)} \\
&= \frac{e_t}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)},
\end{aligned}
$$

so

$$
\begin{aligned}
\left|\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} - \frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)}\right| &= \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \frac{\left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} \\
&\leq \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right| \\
&\quad \times \left|\frac{1}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} - \frac{1}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)}\right| \\
&\quad + \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)\left(\widetilde{\rho}_t, g_t h_t\right)} \left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right| \\
&\leq \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right| \frac{e_t}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)} \\
&\quad + \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t^N, g_t h_t^{\pi_{t-1}^N}\right)\left(\widetilde{\rho}_t, g_t h_t\right)} \left|\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)\right| \\
&\leq (e_t + 1)\, ||f_t||\, \frac{\left|\left(\widetilde{\rho}_t^N, g_t h_t\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)}
\end{aligned}
$$

and

$$
\begin{aligned}
\left|\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)} - \frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)}\right| &\leq \frac{\left(\widetilde{\rho}_t^N, |f_t| \left|g_t h_t^{\pi_{t-1}^N} - g_t h_t\right|\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)} \leq ||f_t|| \frac{\left|\left(\pi_{t-1}^N, f^h\right) - \left(\pi_{t-1}, f^h\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)} \\
\left|\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)} - \frac{\left(\widetilde{\rho}_t, f_t g_t h_t\right)}{\left(\widetilde{\rho}_t, g_t h_t\right)}\right| &\leq \frac{\left|\left(\widetilde{\rho}_t^N, f_t g_t h_t\right) - \left(\widetilde{\rho}_t, f_t g_t h_t\right)\right|}{\left(\widetilde{\rho}_t, g_t h_t\right)}.
\end{aligned}
$$

Hence,

$$E\left[\left(\frac{\left(\widetilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)}{\left(\widetilde{\rho}_t^N, g_t h_t\right)} - \frac{(\widetilde{\rho}_t, f_t g_t h_t)}{(\widetilde{\rho}_t, g_t h_t)}\right)^2\right]^{\frac{1}{2}} \leq \frac{||f_t|| \left((e_t+2)\sqrt{\widetilde{c}_t}\,||g_t h_t|| + \sqrt{c_{t-1}}\,\left|\left|f^h\right|\right|\right)}{(\widetilde{\rho}_t, g_t h_t)\sqrt{N}}. \tag{19}$$

From (18) and (19), the result follows with

$$\overline{c}_t = \frac{\left(2\,(e_t+1)\sqrt{\widetilde{c}_t}\,||g_t h_t|| + 2\sqrt{c_{t-1}}\,\left|\left|f^h\right|\right|\right)^2}{(\widetilde{\rho}_t, g_t h_t)^2}.$$

■

Proof of Lemma 3. Since

$$\left|\left(\ddot{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right| \leq \left|\left(\ddot{\pi}_t^N, f_t\right) - \left(\overline{\pi}_t^N, f_t\right)\right| + \left|\left(\overline{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right|,$$

we get from lemma 2 and (13) that

$$\begin{aligned} E\left[\left(\left(\ddot{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right]^{\frac{1}{2}} &\leq E\left[\left(\left(\ddot{\pi}_t^N, f_t\right) - \left(\overline{\pi}_t^N, f_t\right)\right)^2\right]^{\frac{1}{2}} + E\left[\left(\left(\overline{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right]^{\frac{1}{2}} \\ &\leq \sqrt{\ddot{c}_t}\frac{\|f_t\|^2}{\sqrt{N}}, \end{aligned}$$

where $\sqrt{\ddot{c}_t} = \sqrt{C_t} + \sqrt{\overline{c}_t}$. ■

Proof of Lemma 4. Let $\mathcal{H}_t$ be the σ-field generated by $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N$, then

$$E\left[\left(\pi_t^N, f_t\right)\middle|\,\mathcal{H}_t\right] = \left(\ddot{\pi}_t^N, K_t f_t\right)$$

and one has

$$\begin{aligned} E\left[\left(\left(\pi_t^N, f_t\right) - E\left[\left(\pi_t^N, f_t\right)\middle|\,\mathcal{H}_t\right]\right)^2\middle|\,\mathcal{H}_t\right] &= \frac{1}{N}\left(\left(\ddot{\pi}_t^N, K_t f_t^2\right) - \left(\ddot{\pi}_t^N, K_t f_t\right)^2\right) \\ &\leq \frac{1}{N}\left(\ddot{\pi}_t^N, K_t f_t^2\right) \leq \frac{\|f_t\|^2}{N}, \end{aligned}$$

thus,

$$\begin{aligned} E\left[\left(\left(\ddot{\pi}_t^N, f_t\right) - (\pi_t, f_t)\right)^2\right]^{\frac{1}{2}} &\leq E\left[\left(\left(\pi_t^N, f_t\right) - \left(\ddot{\pi}_t^N, K_t f_t\right)\right)^2\right]^{\frac{1}{2}} + E\left[\left(\left(\ddot{\pi}_t^N, K_t f_t\right) - (\pi_t, K_t f_t)\right)^2\right]^{\frac{1}{2}} \\ &\leq \sqrt{c_t}\frac{\|f_t\|}{\sqrt{N}}, \end{aligned}$$

where $c_t = \left(\sqrt{\bar{c}_t}+1\right)^2$. ■

Remark 2 *It is actually possible to consider a global Markov transition kernel on the N-fold space admitting the N-fold posterior distribution as invariant distribution, that is,*

$$\int K_t\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, d\left\{x_{0:t}^{(i)}\right\}_{i=1}^N\right)\prod_{i=1}^N \pi_t\left(d\ddot{x}_{0:t}^{(i)}\right) = \prod_{i=1}^N \pi_t\left(dx_{0:t}^{(i)}\right).$$

Let us introduce the following notation: for any set of indices $i_1,\ldots,i_n$ ($i_1 \neq i_2 \neq \ldots \neq i_n$)

$$K_t^{i_1,\ldots,i_n}\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, d\left(x_{0:t}^{(i_1)},\ldots,dx_{0:t}^{(i_n)}\right)\right) \triangleq \int K_t\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, d\left\{x_{0:t}^{(i)}\right\}_{i=1}^N\right),$$

where the integration is over $\left\{x_{0:t}^{(i)}; i \in \{1,\ldots,N\}\setminus\{i_1,\ldots,i_n\}\right\}$. If we restrict ourselves to the case where the kernels $K_t^i\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, dx_{0:t}\right)$ are equal for all $i = 1,\ldots,N$, convergence is still ensured under the following "mixing" assumption on the MCMC kernel: for any $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$

$$\sum_{i\neq j}\left|\int K_t^{i,j}\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, d\left(x_{0:t}^{(i)},x_{0:t}^{(j)}\right)\right) f_t\left(x_{0:t}^{(i)}\right) f_t\left(x_{0:t}^{(j)}\right) - \left(\int K_t\left(\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N, dx_{0:t}\right) f_t\left(x_{0:t}\right)\right)^2\right| < M_t\frac{\|f_t\|^2}{N}. \tag{20}$$

■

Proof of Theorem 1. At time $t=0$, we have assumed to sample N i.i.d. particles from π_0, so

$$E\left[\left(\left(\pi_0^N, f_0\right) - \left(\pi_0, f_0\right)\right)^2\right] \leq \frac{\|f_0\|^2}{N}$$

and the result follows straightforwardly from Lemmas 1, 2, 3 and 4. ■

Proof of Lemma 5. Let $\mathcal{G}_{t-1}$ be the σ-field generated by $\left\{x_{0:t-1}^{(i)}\right\}_{i=1}^N$, then

$$E\left[\left(\tilde{\rho}_t^N, f_t\right)|\mathcal{G}_{t-1}\right] = \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right). \tag{21}$$

Since, conditionally on $\mathcal{G}_{t-1}$, $\left\{\widetilde{x}_{0:t}^{(i)}\right\}_{i=1}^N$ are i.i.d. random variables, there exists a constant C independent of N such that

$$E\left[\left(\left(\tilde{\rho}_t^N, f_t\right) - E\left[\left(\tilde{\rho}_t^N, f_t\right)|\mathcal{G}_{t-1}\right]\right)^4 |\mathcal{G}_{t-1}\right] \leq \frac{C\,||f_t||^4}{N^2}. \tag{22}$$

From (21) and (22), we get that

$$E\left[\left(\left(\tilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right)\right)^4\right] \leq \frac{C\,||f_t||^4}{N^2},$$

hence, via a Borel-Cantelli argument, we have, almost surely,

$$\lim_{N\to\infty} \left(\tilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N, \Xi_t^{\pi_{t-1}^N} f_t\right) = 0. \tag{23}$$

Now

$$\left\|\left(\pi_{t-1}^N, \left(\Xi_t^{\pi_{t-1}^N} - \Xi_t\right) f_t\right)\right\| \leq d_t \left|\left(\pi_{t-1}^N, f_t\right) - \left(\pi_{t-1}, f_t\right)\right|$$

so almost surely

$$\lim_{N\to\infty} \left(\tilde{\rho}_t^N, f_t\right) - \left(\pi_{t-1}^N, \Xi_t f_t\right) = 0.$$

But since $\Xi_t f_t$ is a continuous function (using the Feller property of Ξ_t) and, almost surely, $\lim_{N\to\infty} \pi_{t-1}^N = \pi_{t-1}$, then, from (23), one has

$$\lim_{N\to\infty} \left(\rho_t^N, f_t\right) = \left(\pi_{t-1}, \Xi_t f_t\right) = \left(\tilde{\rho}_t, f_t\right).$$

■

Proof of Lemma 6. From the definition of $\bar{\pi}_t^N$ we have that for any $f_t \in B\left(\left(\mathbb{R}^{n_x}\right)^{t+1}\right)$

$$\left(\bar{\pi}_t^N, f\right) = \frac{\left(\tilde{\rho}_t^N, f_t g_t h_t\right)}{\left(\tilde{\rho}_t^N, g_t h_t\right)}.$$

Since, almost surely, $\lim_{N\to\infty} \tilde{\rho}_t^N = \rho_t$, we have that $\left(\tilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right)$ as by assumption $g_t h_t^{\pi_{t-1}^N}$ is a bounded continuous function

$$\lim_{N\to\infty} \left(\tilde{\rho}_t^N, f_t g_t h_t^{\pi_{t-1}^N}\right) = \left(\tilde{\rho}_t, f_t g_t h_t^{\pi_{t-1}^N}\right) \quad \lim_{N\to\infty} \left(\rho_t^N, g_t h_t^{\pi_{t-1}^N}\right) = \left(\rho_t, g_t h_t^{\pi_{t-1}^N}\right)$$

and

$$\begin{aligned} \left\|\left(\tilde{\rho}_t, f_t g_t h_t^{\pi_{t-1}^N}\right) - \left(\tilde{\rho}_t, f_t g_t h_t\right)\right\| &\leq \|f_t\| \left|\left(\pi_{t-1}^N, f^h\right) - \left(\pi_{t-1}, f^h\right)\right| \\ \left\|\left(\tilde{\rho}_t, g_t h_t^{\pi_{t-1}^N}\right) - \left(\tilde{\rho}_t, g_t h_t\right)\right\| &\leq \left|\left(\pi_{t-1}^N, f^h\right) - \left(\pi_{t-1}, f^h\right)\right|, \end{aligned}$$

so since, almost surely, $\lim_{N\to\infty} \pi_{t-1}^N = \pi_{t-1}$ then

$$\lim_{N\to\infty} \left(\tilde{\rho}_t^N, f_t g_t h_t\right) = \left(\widetilde{\rho}_t, f_t g_t h_t\right), \quad \lim_{N\to\infty} \left(\rho_t^N, g_t h_t\right) = \left(\rho_t, g_t h_t\right)$$

as by assumption $g_t h_t$ is a bounded continuous function. Hence, almost surely, $\lim_{N\to\infty} \left(\bar{\pi}_t^N, f_t\right) = \frac{(\widetilde{\rho}_t, f_t g_t h_t)}{(\rho_t, g_t h_t)} = (\pi_t, f_t)$ for any $f_t \in B\left(\left(\mathbb{R}^{n_x}\right)^{t+1}\right)$ and, therefore, $\lim_{N\to\infty} \bar{\pi}_t^N = \pi_t$. Now, from (14), we have

$$E\left[\left|\left(\ddot{\pi}_t^N, f_t\right) - \left(\bar{\pi}_t^N, f_t\right)\right|^p\right] \leq \frac{C_t \left|\left|f_t\right|\right|^p}{N^{1+\varepsilon}}, \tag{24}$$

where $\varepsilon = p - \alpha - 1 > 0$. From (24) again via a Borel-Cantelli argument, we have, almost surely, that

$$\lim_{N\to\infty} \left(\ddot{\pi}_t^N, f_t\right) - \left(\bar{\pi}_t^N, f_t\right) = 0,$$

hence the claim. ■

Proof of Lemma 7. The proof is identical to that of lemma 5. Let $\mathcal{H}_t$ be the σ-field generated by $\left\{\ddot{x}_{0:t}^{(i)}\right\}_{i=1}^N$, then

$$E\left[\left(\pi_t^N, f\right) | \mathcal{H}_t\right] = \left(\ddot{\pi}_t^N, K_t f\right). \tag{25}$$

Also it is relatively easy to prove that there exists a constant C independent of N such that

$$E\left[\left(\left(\pi_t^N, f\right) - E\left[\left(\pi_t^N, f\right) | \mathcal{H}_t\right]\right)^4 | \mathcal{H}_t\right] \leq \frac{C \left|\left|f\right|\right|^4}{N^2}. \tag{26}$$

From (25) and (26), we get that

$$E\left[\left(\left(\pi_t^N, f\right) - \left(\ddot{\pi}_t^N, K_t f\right)\right)^4\right] \leq \frac{C_t \left|\left|f\right|\right|^4}{N^2}$$

hence, via a Borel-Cantelli argument, we have, almost surely,

$$\lim_{N\to\infty} \left(\pi_t^N, f\right) - \left(\ddot{\pi}_t^N, K_t f\right) = 0. \tag{27}$$

But since $K_t f$ is a continuous function (using the Feller property of K_t) and, almost surely, $\lim_{N\to\infty} \pi_{t-1}^N = \pi_{t-1}$, we have, from (27), that

$$\lim_{N\to\infty} \left(\pi_t^N, f\right) = \left(\pi_t, K_t f\right) = \left(\pi_t K_t, f\right) = \left(\pi_t, f\right).$$

which gives our claim. ■

Proof of Theorem 2. By putting together lemmas 5, 6 and 7 and observing that, almost surely, $\lim_{N\to\infty} \pi_0^N = \pi_0$ we straightforwardly obtain the result. ■

B Figure and Table

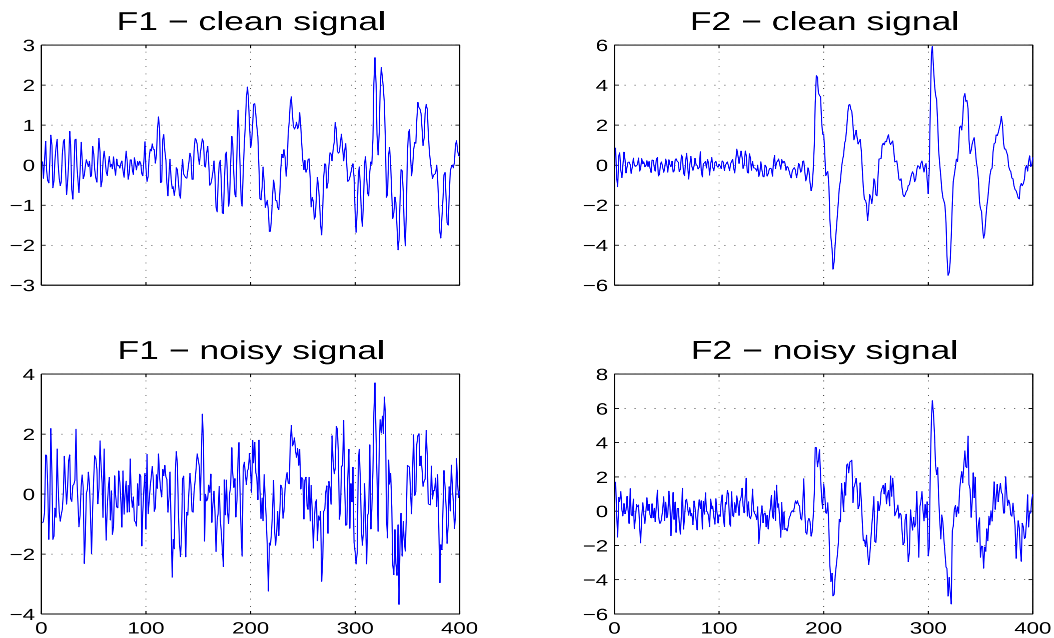

Figure 1: Clean (top) and noisy (bottom) speech frames depicting the transitions between /sh/ and /uw/ in “should” (left) and /s/ and /er/ in “service” (right).

L	0	10	20	30	40
F_1	2.83	3.10	3.40	3.30	3.43
F_2	1.72	1.92	2.07	2.09	2.06

Table 1: SNR improvement results in dB *vs.* the lag L

References

Berzuini C., Best N., Gilks W.R. & Larizza C. (1997) Dynamic conditional independence models and Markov chain Monte Carlo methods. *J. Am. Statist. Ass.*, **92**, 1403-1412.

Carpenter J., Clifford P. & Fearnhead P. (1999) Building robust simulation-based filters for evolving data sets. Technical report, University of Oxford, Dept. of Statistics.

Chen R. & Liu J.S. (2000) Mixture Kalman filters. *J. R. Statist. Soc.* B, **62**, 493-508.

Crisan D., Del Moral P. & Lyons T. (1999) Discrete filtering using branching and interacting particle systems. *Markov Proc. Rel. Fields*, **5**, 293-318.

Crisan D. (2001) Particle filters - A theoretical perspective. In *Sequential Monte Carlo Methods in Practice* (eds A. Doucet, J.F.G. de Freitas & N.J. Gordon). New-York: Springer-Verlag.

Del Moral P. (1996) Non linear filtering: interacting particle solution. *Markov Proc. Rel. Fields*, **2**, 555-580.

Del Moral P. (1997) Measure valued processes and interacting particle systems. Application to non linear filtering problems. *Ann. Appl. Proba.,* **8**, 438-495.

Del Moral P. & Guionnet A. (2001) On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Instit. Henri Poincaré*. to appear.

Doucet A., de Freitas J.F.G. & Gordon N.J. (eds.) (2001) *Sequential Monte Carlo Methods in Practice.* New York: Springer-Verlag.

Doucet A., Godsill S.J. & Andrieu C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, **10**, 197-208.

Fox D., Thrun S., Burgard W. & Dellaert F. (2001). Particle filters for mobile robot localization. In *Sequential Monte Carlo Methods in Practice* (eds A. Doucet, J.F.G. de Freitas and N.J. Gordon). New York: Springer-Verlag.

Gilks W.R. & Berzuini C. (1999) Following a moving target - Monte Carlo inference for dynamic Bayesian models. submitted to *J. R. Statist. Soc.* B.

Gordon N.J., Salmond D.J. & Smith A.F.M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F*, **140**, 107-113.

Higuchi T. (1997) Monte Carlo filter using the genetic algorithm operators. *J. Statist. Comput. Simul.*, **59**, 1-23.

Kitagawa G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, **5**, 1-25.

Hürzeler M. & Künsch, H.R. (1998) Monte Carlo approximations for general state space models. *J. Comput. Graph. Statist.*, **7**, 175-193.

Iba Y. (2000) Population-based Monte Carlo algorithms. Research memorandum no. 757, Institute of Statistical Mathematics, Tokyo.

Isard M. & Blake A. (1998) CONDENSATION – conditional density propagation for visual tracking, *Int. J. Computer Vision*, **29**, 5–28.

Liu J.S. & Chen R. (1998) Sequential Monte Carlo methods for dynamic systems. *J. Am. Statist. Ass.*, **93**, 1032-1044.

Liu J. and West M. (2001) Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (eds A. Doucet, J.F.G. de Freitas and N.J. Gordon). New York: Springer-Verlag.

Pitt M.K. & Shephard N. (1999) Filtering via simulation: auxiliary particle filter. *J. Am. Statist. Ass.*, **94**, 590-599.

Ripley B.D. (1987) *Stochastic Simulation*, New York: Wiley.

Robert C.P. & Casella G. (1999) *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Tanizaki H. (1996) *Nonlinear Filters: Estimation and Applications*. 2nd edition, Berlin: Springer-Verlag.

Vermaak J., Andrieu C., Doucet A. & Godsill S.J. (2000) On-line Bayesian modelling and enhancement of speech signals. Technical report CUED/F-INFENG/TR.361, Cambridge University.

West M. (1993a) Approximating posterior distributions by mixtures. *J. R. Statist. Soc.* B, **55**, 409-422.

West M. (1993b) Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Comput. Science Statist.*, **24**, 325-333.

West M. & Harrison P.J. (1997) *Bayesian Forecasting and Dynamic Models*. 2nd edition, New York: Springer-Verlag.