Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation

Olivier Cappé

Centre National de la Recherche Scientifique
et Ecole Nationale Supérieure des Télécommunications,* Paris

Corresponding Address:
O. Cappé
ENST, Dpt. TSI
46 rue Barrault
75634 Paris cedex 13, France
email: cappe@tsi.enst.fr
phone: +33 1 45 81 71 11 / fax: + 33 1 45 88 79 35

Abstract

We consider parameter estimation for a class of discrete-time partially observed Markovian processes, known as switching autoregressive models, which are used in a variety of applications which range from finance (stochastic volatility), to signal processing (deconvolution) or telecommunications (teletraffic modeling). For such models, maximum likelihood estimation (be it in the Expectation-Maximization approach or via direct computation of the log-likelihood and its derivatives) implies the computation of smoothed additive functionals of the hidden process. A little known property of the class of models under consideration is that there exists a generic filtering (or recursive in time) procedure for computing such smoothed additive functionals. However, when the hidden process is not finite valued, this procedure cannot, in general, be implemented exactly. We thus propose an approximate simulation-based filtering scheme based on the sequential Monte Carlo (or particle filtering) approach.

Keywords: Hidden Markov Models, State Space Model, Switching Autoregression, Sequential Monte Carlo, Particle Filtering, Maximum Likelihood, Expectation-Maximization

1 Introduction

1.1 Model, hypotheses and notations

We consider parametric models consisting of a discrete time homogeneous Markovian process $(X_t)_{t\geq 1}$ on a general Polish state space E with Borel σ-field $\mathcal{B}(E)$, indirectly observed through $\mathbb{R}^d$ valued observations $(Y_t)_{t\geq 1}$. A particular case of interest is when the Y_ts are conditionally independent given the X_ts, that is when

$$\mathbb{P}^\theta(Y_{t_1} \in B_1, \dots Y_{t_k} \in B_k | X_{t_1}, \dots X_{t_k}) = \prod_{i=1}^{k} \int_{B_i} g^\theta(y_{t_i}|X_{t_i}) dy_{t_i} \tag{1}$$

*This work has been supported by the European Union's TMR research network RB-FMRX-CT96-0095 on *Statistical and computational methods for the analysis of spatial data*

for any choice of the integer k, of the time indexes t_1 to t_k and of the Borel subsets of $\mathbb{R}^d$ $B_1, \dots B_k$. g^θ is a family of conditional probability density functions (or simply, pdf) with respect to (abbreviated to "wrt" henceforth) Lebesgue measure on $\mathbb{R}^d$, and the superscript θ denotes the dependence upon the parameter. Depending on the context, (1) is generally referred to as a Hidden Markov Model or HMM (MacDonald and Zucchini, 1997), or as a state space model (Brockwell and Davis, 1991, §12). It turns out that the methods investigated in this paper apply as well to a slightly more general model known as the switching autoregressive model which we shall consider for greater generality. In the switching autoregressive model (Hamilton, 1994, §22), it is assumed that

$$\mathbb{P}^\theta(Y_1 \in B_1, \dots Y_t \in B_t | X_1, \dots X_t) = \int_{B_1} \cdots \int_{B_t} \left[h^\theta(y_1|X_1) \prod_{s=2}^{t} g^\theta(y_s|X_s, y_{s-1}) \right] dy_1 \dots dy_t \tag{2}$$

where h^θ is a probability density wrt Lebesgue measure on $\mathbb{R}^d$. Thus (1) is a particular example of (2) where $g^\theta(y_s|x_s, y_{s-1})$ does not depend on the previous observation y_{s-1}.

Let R denote a generic transition kernel, λ and ν two measures and x a point, the following standard notations will be used: $d\lambda/d\nu$, the Radon-Nikodym derivative of the λ wrt ν; λR, the image of the measure λ obtained when applying one step of the transition kernel; δ_x, the Dirac mass in x, and R_x, the measure $\delta_x R$. $C(E, \mathbb{R}^q)$ and $C(E^2, \mathbb{R}^q)$ respectively denote the space of continuous $\mathbb{R}^q$ valued functions on E and $E \times E$.

We further assume that the transition kernel K^θ of the hidden chain (X_t) is dominated by some Radon measure μ on E (for all values of θ) and denote by $k^\theta(x, \bullet)$ the pdf of K^θ_x wrt μ. Likewise, we denote by π^θ_1 the probability measure corresponding to the initial state X_1 and its pdf (wrt μ) is denoted by l^θ.

1.2 Motivations

The main contribution of the paper consists in a systematic scheme for computing *recursively in the time index t* quantities of the form

$$Q^\theta_t = \sum_{s=1}^{t} \mathbb{E}^\theta \left(m^\theta_s(X_s) \middle| Y_{1:t} \right) + \sum_{s=2}^{t} \mathbb{E}^\theta \left(r^\theta_s(X_{s-1}, X_s) \middle| Y_{1:t} \right) \tag{3}$$

where the subscripting "$r\!:\!s$" is generically used to denote the collection of variables with time indexes from r to s (included); $(m_s)^\theta_{s\geq 1} \in C(E, \mathbb{R}^q)$ and $(r_s)^\theta_{s\geq 2} \in C(E^2, \mathbb{R}^q)$ (for some q) are functions which may depend on the parameter θ. Q^θ_t as defined in (3) is a smoothed additive functional of the hidden chain conditioned on the observations Y_1 up to Y_t. Such functionals are of prime importance for estimation of the parameter θ.

To illustrate this point, consider first the Expectation-Maximization (EM) framework of (Dempster et al., 1977). In this approach, the log-likelihood is optimized iteratively by repeated maximizations of intermediate quantities defined as

$$\mathcal{Q}_{\mathrm{EM}}(\theta|\hat\theta) = \mathbb{E}^{\hat\theta} \left(\log p^\theta(X_{1:t}, Y_{1:t}) \middle| Y_{1:t} \right) \tag{4}$$

where $\hat\theta$ denotes the current estimate of the parameters. Computation of (4) is generally referred to as the E step whereas the maximization $\hat\theta \hookleftarrow \arg\max_\theta \mathcal{Q}_{\mathrm{EM}}(\theta|\hat\theta)$ is the so-called M step (with the left arrow denoting variable substitution). Because the joint process $(X_t, Y_t)_{t\geq 1}$ is Markovian, (4) may be decomposed as

$$\begin{aligned} \mathcal{Q}_{\mathrm{EM}}(\theta|\hat\theta) = {} & \mathbb{E}^{\hat\theta} \left(\log l^\theta(X_1) + \log h^\theta(Y_1|X_1) \middle| Y_{1:t} \right) \\ & + \sum_{s=2}^{t} \mathbb{E}^{\hat\theta} \left(\log g^\theta(Y_s|X_s, Y_{s-1}) \middle| Y_{1:t} \right) + \sum_{s=2}^{t} \mathbb{E}^{\hat\theta} \left(\log k^\theta(X_{s-1}, X_s) \; Y_{1:t} \right) \end{aligned} \tag{5}$$

which is of the form defined in (3). In general settings however, numerical computation (5) for a given value of θ is not sufficient to implement the EM approach since maximization wrt θ is required. A very frequent case examined in detail by (Dempster et al., 1977), is when the complete data distribution (joint distribution of $X_{1:t}$ and $Y_{1:t}$) is from the exponential family, that is when

$$p^{\theta}(X_{1:t}, Y_{1:t}) = \exp\left[A(\theta)B(X_{1:t}, Y_{1:t}) + C(X_{1:t}, Y_{1:t}) + D(\theta)\right]$$

where $B(X_{1:t}, Y_{1:t})$ is a (possibly vector-valued) complete data sufficient statistic. Because of the Markovian dependence of $(X_t, Y_t)_{t\geq 1}$, B is a sum of terms which only involve two successive time indexes. Thus, maximization of $\mathcal{Q}_{\mathrm{EM}}(\theta|\hat{\theta})$ wrt θ only requires the computation of

$$\mathbb{E}^{\hat{\theta}}\left(B(X_{1:t}, Y_{1:t})|Y_{1:t}\right)$$

which still has the general form given in (3).

As a representative example of this situation, consider the stochastic volatility model (Kim et al., 1998) where

$$Y_t = \mathrm{e}^{X_t} N_t$$

where $(N_t)_{t\geq 1}$ is an iid sequence of standard Gaussian random variables, independent of the volatility process $(X_t)_{t\geq 1}$ which is described by a first order Gaussian autoregressive model

$$X_{t+1} = \theta_1 + \theta_2(X_t - \theta_1) + \theta_3 E_{t+1}$$

$(E_t)_{t\geq 2}$ being a standardized Gaussian iid sequence. The distribution of the initial state X_1 is here chosen such that the observed process $(Y_t)_{t\geq 1}$ is stationary. For this model,

$$\begin{aligned} k^{\theta}(x_{t-1}, x_t) &= n(x_t; \theta_1 + \theta_2(x_{t-1} - \theta_1), \theta_3^2) \\ g^{\theta}(y_t|x_t, y_{t-1}) &= n(y_t; 0, \mathrm{e}^{x_t}) \\ h^{\theta}(y_1|x_1) &= n(y_1; 0, \mathrm{e}^{x_1}) \\ l^{\theta}(x_1) &= n(x_1; 0, \theta_3^2/(1-\theta_2^2)) \end{aligned} \tag{6}$$

where $n(\bullet\,; \mu, \sigma^2)$ denotes the Gaussian pdf with mean μ and variance σ^2. For this particular model, the complete data sufficient statistic is four dimensional and thus each iteration of the EM algorithm can be carried out by computing the vector

$$\left(\sum_{s=1}^{t} \mathbb{E}^{\hat{\theta}}\left[Y_s \mathrm{e}^{-X_s}\middle| Y_{1:t}\right], \sum_{s=1}^{t} \mathbb{E}^{\hat{\theta}}\left[X_s^2\middle| Y_{1:t}\right], \sum_{s=2}^{t-1} \mathbb{E}^{\hat{\theta}}\left[X_s^2\middle| Y_{1:t}\right], \sum_{s=2}^{t} \mathbb{E}^{\hat{\theta}}\left[X_{s-1}X_s\middle| Y_{1:t}\right]\right)$$

whose components all are particular cases of (3).

Another important application is the computation of the gradient of the log-likelihood, through Fisher identity (Dempster et al., 1977, discussion by B. Efron): Under standard regularity assumptions, the gradient of the log-likelihood may be written as

$$\nabla_{\theta} \log p^{\theta}(Y_{1:t}) = \mathbb{E}^{\theta}\left(\nabla_{\theta} \log p^{\theta}(X_{1:t}, Y_{1:t})\middle| Y_{1:t}\right) \tag{7}$$

From (5) it is easily seen that for switching autoregressive models, (7) is an instance of (3) for the particular choice

$$\begin{aligned} m_1^{\theta}(x_1) &= \nabla_{\theta} \log l^{\theta}(x_1) + \nabla_{\theta} \log h^{\theta}(Y_1|x_1) \\ m_t^{\theta}(x_t) &= \nabla_{\theta} \log g^{\theta}(Y_t|x_t, Y_{t-1}) \quad (t \geq 2) \\ r_t^{\theta}(x_{t-1}, x_t) &= \nabla_{\theta} \log k^{\theta}(x_{t-1}, x_t) \end{aligned} \tag{8}$$

The EM approach is well known for being simple to implement and numerically well-behaved. On the other hand, optimization of the log-likelihood using its gradient is potentially much faster thanks to the availability of quadratically converging optimization strategies (quasi Newton, or conjugate directions) – see (Cappé et al., 1998) for a comparison of both approaches in a simple case. Depending on the constraints of the application under consideration, both strategies can thus be useful.

1.3 Known solutions and open problems

For general models, the main difficulty in computing (3) lies in the evaluation of the smoothing distributions. For finite state space HMMs (when E is finite), the smoothed distributions can be evaluated efficiently by a procedure known as the forward-backward due to Baum and his coworkers (MacDonald and Zucchini, 1997). A similar procedure is available for linear Gaussian state space models (when $E = \mathbb{R}^q$ for some q and the joint distribution of $(X_{t:t+1}, Y_{t:t+1})$ for any index t is multivariate normal) (De Jong, 1989). Both procedures however share the same shortcoming that a double recursion, for increasing time indexes and then for decreasing time indexes, is required. In practice, this means that a storage space that grows linearly with the number of observations is needed, which can be problematic for applications involving large datasets such as finance of bioinformatics.

In addition, the non causal nature of these smoothing procedures is a real obstacle when trying to devise efficient on-line (recursive in the time index t) strategies to estimate the parameter θ. This last problem is considered by (LeGland and Mevel, 1997) and (Collings and Ryden, 1998) who used the fact that the gradient of the log-likelihood can be updated recursively using formulas obtained by formal differentiation of the filtering recursion. More specifically, the log-likelihood may be decomposed as

$$\log p^\theta(Y_{1:t+1}) = \log p^\theta(Y_{1:t}) + \log\left(\int_E g^\theta(Y_{t+1}|x_{t+1}, Y_t)\pi^\theta_{t+1}(dx_{t+1})\right) \quad (\text{for } t \geq 2) \tag{9}$$

Thus, differentiation, wrt θ, of (9) together with the filtering relations (11)-(12) described in section 2.1 yields recursive update formulas for computing $\nabla_\theta \log p^\theta(Y_{1:t+1})$.

Another approach (upon which we will draw in the next section) is based on the EM intermediate quantity for which exact recursive filters have been proposed by (Zeitouni and Dembo, 1988) and further developed by Elliot and coworkers – see (Elliott and Krishnamurthy, 1999) for instance. The fact that the same principle can be applied generically for all additive functionals of the form given in (3), and hence for the computation of the log-likelihood and its gradient, has apparently not been recognized by these authors.

Of course, except in some specific cases (including finite state space HMMs and linear Gaussian state space models), even the forward-backward approach can not be applied anymore because the smoothing distributions no longer have closed form expressions. This is already the case for the simple stochastic volatility model defined in (6). The techniques used in this situation are usually based on Markov Chain Monte Carlo (MCMC) simulations (Kim et al., 1998), (Cappé et al., 1999) which are very similar in principle to the forward-backward approach in that they imply conditioning both on past and future indexes of the hidden process (X_t).

Attempts to circumvent this limitation with sequential Monte Carlo methods (also known as particle filtering) include (Pitt and Shephard, 1999) and (Hürzeler and Kunsch, 2000). These authors however only consider the evaluation of the log-likelihood which can be straightforwardly approximated from (9). Maximization of the log-likelihood is then carried out by a grid search which would clearly be impractical for large (multidimensional) parameter spaces.

2 Recursive update formulas

2.1 Preliminary: Standard prediction and filtering

Let π_s^θ and φ_s^θ denote respectively the prediction and filtering probability measures defined as

$$\pi_s^\theta(\bullet) = \mathbb{P}^\theta\left(X_s \in \bullet | Y_{1:s-1}\right) \quad \text{and} \quad \varphi_s^\theta(\bullet) = \mathbb{P}^\theta\left(X_s \in \bullet | Y_{1:s}\right)$$

Both of these quantities may be computed according to the following recursions:

$$\frac{d\varphi_1^\theta}{d\pi_1^\theta}(x_1) = w_1^\theta(x_1) = \frac{h^\theta(Y_1|x_1)}{\int_E h^\theta(Y_1|x_1)\pi_1^\theta(dx_1)} \qquad (initialization) \tag{10}$$

For $s = 1, \dots t-1$,

$$\pi_{s+1}^\theta = \varphi_s^\theta K^\theta \qquad (prediction) \tag{11}$$

$$\frac{d\varphi_{s+1}^\theta}{d\pi_{s+1}^\theta} = w_{s+1}^\theta \qquad (filtering) \tag{12}$$

where

$$w_{s+1}^\theta(x_{s+1}) = \frac{g^\theta(Y_{s+1}|x_{s+1}, Y_s)}{\int_E g^\theta(Y_{s+1}|x_{s+1}, Y_s)\pi_{s+1}^\theta(dx_{s+1})} \tag{13}$$

The filter-to-predictor update thus consists of one step of the transition kernel of the hidden chain, while the predictor-to-filter update corresponds to an application of Bayes rule.

2.2 Extension: computing smoothed functionals

For a time index t, let $A \in \sigma(X_{1:t})$ denote a past event. The important remark used by (Zeitouni and Dembo, 1988) is that, whereas $\mathbb{P}^\theta(A|Y_{1:t+1})$ cannot be directly computed from $\mathbb{P}^\theta(A|Y_{1:t})$, it is possible to update the (unnormalized) measure $\mathbb{P}^\theta(A, X_t \in \bullet|Y_{1:t})$ so as to obtain $\mathbb{P}^\theta(A, X_{t+1} \in \bullet|Y_{1:t+1})$. To build on this remark, first note that the Markovian structure implies that

$$\mathbb{P}^\theta(A, X_{t+1} \in \bullet|Y_{1:t}) = \int_E \mathbb{P}^\theta(A, dx_t|Y_{1:t})K^\theta(x_t, \bullet) \tag{14}$$

Next, apply Bayes' rule to obtain

$$\frac{d\mathbb{P}^\theta(A, X_{t+1} \in \bullet|Y_{1:t+1})}{d\mathbb{P}^\theta(A, X_{t+1} \in \bullet|Y_{1:t})} = w_{t+1}^\theta \tag{15}$$

where w_{t+1}^θ is defined in (13). Perhaps surprisingly, the above equations show that the unnormalized measure $\mathbb{P}^\theta(A, X_t \in \bullet|Y_{1:t})$ can be updated recursively using the same formulas as for the standard filtering probability measure.

In order to generalize this observation to the computation of general Q_t^θ functionals given in (3), define the signed measures

$$\Pi_t^\theta(\bullet) = \sum_{s=1}^t \int_E m_s^\theta(x_s)\mathbb{P}^\theta(dx_s, X_t \in \bullet|Y_{1:t-1}) + \sum_{s=2}^t \int_{E^2} r_s^\theta(x_{s-1}, x_s)\mathbb{P}^\theta(dx_{s-1}, dx_s, X_t \in \bullet|Y_{1:t-1}) \tag{16}$$

and

$$\Phi_t^\theta(\bullet) = \sum_{s=1}^{t} \int_E m_s^\theta(x_s)\mathbb{P}^\theta(dx_s, X_t \in \bullet | Y_{1:t}) \\ + \sum_{s=2}^{t} \int_{E^2} r_s^\theta(x_{s-1}, x_s)\mathbb{P}^\theta(dx_{s-1}, dx_s, X_t \in \bullet | Y_{1:t}) \quad (17)$$

for $t \geq 2$, with $d\Pi_1^\theta / d\pi_1^\theta = m_1^\theta$ and $d\Phi_1^\theta / d\varphi_1^\theta = m_1^\theta$. Proceeding as for (14) and (15), one obtains the following updating equations

$$\Pi_{t+1}^\theta(B) = \int_{x_t \in E} \int_{x_{t+1} \in B} \varphi_t^\theta(dx_t) K^\theta(x_t, dx_{t+1}) \left(m_{t+1}^\theta(x_{t+1}) + r_{t+1}^\theta(x_t, x_{t+1}) \right) \\ + \int_E \Phi_t^\theta(dx_t) K^\theta(x_t, B) \qquad (\text{for } B \in \mathcal{B}(E)) \quad (18)$$

and

$$\frac{d\Phi_{t+1}^\theta}{d\Pi_{t+1}^\theta} = w_{t+1} \quad (19)$$

An important remark to be used in what follows is that $\varphi_t^\theta(dx_t)K^\theta(x_t, dx_{t+1})$ featured in (16) is the joint distribution of X_t and X_{t+1} given $Y_{1:t}$.

For any time index t, the quantity of interest can be evaluated by integration wrt Π_t^θ or Φ_t^θ with

$$Q_t^\theta = \int_E w_t^\theta(x_t)\Pi_t^\theta(dx_t) \quad \text{or} \quad Q_t^\theta = \Phi_t^\theta(E) \quad (20)$$

Thus (18)-(19) and (20) together with (11)-(12) define our recursive algorithm for computing Q_t^θ for all times indexes.

3 Particle approximation

This part of the paper deals with sequential Monte Carlo approximation to the recursive mechanism presented in the previous section. A hat sign is placed over approximate quantities computed from Monte Carlo averages to distinguish them from their exacts counterparts introduced so far.

3.1 Particle filtering

The motivation for the basic approach to particle filtering, usually referred to as "the bootstrap filter" (Doucet et al., 2000), is the following: Assume that at time index t, the predictive distribution is approximated the empirical probability measure associated with a sample $\{X_t^\theta[i]\}_{1 \leq i \leq p}$,

$$\hat{\pi}_t^\theta = 1/p \sum_{i=1}^{p} \delta_{X_t^\theta[i]} \quad (21)$$

where $X_t^\theta[i] \in E$ are generally referred to as the "particles".

Applying one complete step of the mapping defined by (12)-(11) yields (for $t \geq 2$)

$$\tilde{\pi}_{t+1}^\theta = \sum_{i=1}^{p} w_t^\theta[i] \, K_{X_t^\theta[i]}^\theta \quad (22)$$

where $w_{t+1}^\theta[i] = g^\theta(Y_t | X_t^\theta[i], Y_{t-1}) / \sum_{i=1}^{p} g(Y_t | X_t^\theta[i], Y_{t-1})$. The resulting predictive distribution $\tilde{\pi}_{t+1}^\theta$ defined by (22) is a mixture distribution, from which it is possible to obtain p (conditionally) independent samples by

1. Drawing $(I_t^\theta[1], \dots I_t^\theta[p])$ from a discrete distribution with probabilities $(w_t^\theta[1], \dots w_t^\theta[p])$ (p iid draws with replacement).

2. Drawing independent samples $X_{t+1}^\theta[i]$ from each of the the distributions $K^\theta_{X_t^\theta[I_t^\theta[i]]}$ for $i = 1, \dots p$.

The predictive measure at time step $t+1$ is then approximated by the empirical probability measure associated with the new particles $\{X_{t+1}^\theta[i]\}_{1\le i\le p}$. For obvious reasons, the step 1 above is generally referred to as "multinomial resampling". Although the bootstrap filter is certainly not the only approach for constructing a sequential Monte Carlo approximation, (Del Moral and Miclo, 2000) show that the fact that each step of the algorithm can be decomposed into, first, an application of the *exact* prediction mapping to the current approximation of the predictive measure, followed by, the approximation of the resulting distribution by an empirical measure, is instrumental in proving the convergence of the approximation (as the number p of particle increases) under reasonable conditions. Thus, the distinctive feature of the bootstrap filter compared to other approaches to sequential Monte Carlo — see (Doucet et al., 2000) for a recent review of these — is that given $\mathcal{F}_t = \sigma\left(X_t^\theta[1\!:\!p], (Y_t)_{t\ge 1}\right)$, the "particles" at time index $t+1$ are conditionally iid with a distribution that satisfies

$$\mathbb{P}^\theta\left(X_{t+1}^\theta[i] \in B \,\middle|\, \mathcal{F}_t\right) = \sum_{i=1}^p \frac{g^\theta(Y_t|X_t^\theta[i], Y_{t-1})}{\sum_{i=1}^p g^\theta(Y_t|X_t^\theta[i], Y_{t-1})} K^\theta_{X_t^\theta[i]}(B) \quad (\text{for } B \in \mathcal{B}(E)) \tag{23}$$

which coincides with the result of (12)-(11) applied to $\hat{\pi}_t^\theta$.

Note that both

$$\sum_{i=1}^p w_t^\theta[i]\, \delta_{X_t^\theta[i]} \quad \text{and} \quad 1/p \sum_{i=1}^p \delta_{X_t^\theta[I_t^\theta[i]]}$$

provide approximations to the filtering probability measure ϕ_t, the latter having an increased (conditional) variance due to the resampling. For the same reason,

$$\sum_{i=1}^p \delta_{(X_t^\theta[I_t^\theta[i]], X_{t+1}^\theta[i])}$$

is an approximation of the joint distribution of (X_t, X_{t+1}) given $Y_{1:t}$.

3.2 Approximation of smoothed functionals

We now consider approximating general functionals of the form (3) with a recursive particle type algorithm which follows the bootstrap filter philosophy outlined in the previous section. The closing remark of the previous section suggests a very simple way of approximating Π_t^θ since the updating equation obtained in (18) essentially involves integrating wrt the joint distribution of X_t and X_{t+1} given $Y_{1:t}$. First note that from its definition in (16), Π_t^θ is absolutely continuous wrt to the (standard) prediction distribution π_t^θ. In the context of the particle approximation, it is thus reasonable to approximate Π_t^θ with

$$\hat{\Pi}_t^\theta = 1/p \sum_{i=1}^p \gamma_t^\theta[i] \delta_{X_t^\theta[i]} \tag{24}$$

where $(\gamma_t^\theta[i])_{1\le i\le p}$ are weights. To update the weights $\gamma_t^\theta[i]$, we propose to use the relation

$$\gamma_{t+1}^\theta[i] = m_{t+1}^\theta(X_{t+1}^\theta[i]) + r_{t+1}^\theta\left(X_t^\theta\left[I_t^\theta[i]\right], X_{t+1}^\theta[i]\right) + \gamma_t^\theta\left[I_t^\theta[i]\right] \tag{25}$$

Using (19) and (20), the resulting approximation to Q_{t+1}^{θ} is then given by

$$\hat{Q}_{t+1}^{\theta} = \sum_{i=1}^{p} w_{t+1}^{\theta}[i]\gamma_{t+1}^{\theta}[i] \tag{26}$$

It is clear from (25) that the couples $(X_{t+1}^{\theta}[i], \gamma_{t+1}^{\theta}[i])$ obtained with this scheme are still conditionally iid given $\mathcal{F}_t = \sigma\left(X_t^{\theta}[1:p], \gamma_t^{\theta}[1:p], (Y_t)_{t\geq 1}\right)$ and that the marginal distribution of $X_{t+1}^{\theta}[i]$ is the same as for the standard bootstrap filter.

For a set $B \in \mathcal{B}(E)$, denote by $\mathbb{I}_B$ the indicator function of the set B. First write

$$\mathbb{E}^{\theta}\left(\gamma_{t+1}^{\theta}[i]\mathbb{I}_B(X_{t+1}^{\theta}[i]) \,\middle|\, I_t^{\theta}[i], \mathcal{F}_t\right) = \\ \int_B \left\{m_{t+1}^{\theta}(x) + r_{t+1}^{\theta}\left(X_t^{\theta}\left[I_t^{\theta}[i]\right], x\right) + \gamma_t^{\theta}\left[I_t^{\theta}[i]\right]\right\} K^{\theta}\left(X_t^{\theta}\left[I_t^{\theta}[i]\right], dx\right) \tag{27}$$

And thus

$$\mathbb{E}^{\theta}\left(\gamma_{t+1}^{\theta}[i]\mathbb{I}_B(X_{t+1}^{\theta}[i]) \,\middle|\, \mathcal{F}_t\right) = \\ \sum_{i=1}^{p} w_t^{\theta}[i] \int_B \left(m_{t+1}^{\theta}(x) + r_{t+1}^{\theta}\left(X_t[i], x\right)\right) K^{\theta}\left(X_t^{\theta}[i], dx\right) + \sum_{i=1}^{p} w_t^{\theta}[i]\gamma_t^{\theta}[i] K^{\theta}\left(X_t^{\theta}[i], B\right) \tag{28}$$

(28) indeed coincides with what would be obtained by application of (19)-(18) and (12) to $\hat{\pi}_t^{\theta}$ and $\hat{\Pi}_t^{\theta}$, as defined in (21) and (24). Using standard arguments concerning the convergence of empirical measures, it is then easily shown that when p gets large,

$$\hat{\Pi}_{t+1}^{\theta} = 1/p \sum_{i=1}^{p} \gamma_{t+1}^{\theta}[i]\delta_{X_{t+1}^{\theta}[i]}$$

is a good approximation to the exact mapping (19)-(18) applied to $\hat{\Pi}_t^{\theta}$. It is of course a very weak result in itself and only constitutes a small step towards proving that $\hat{\Pi}_t^{\theta}$ is indeed a good approximation to Π_t^{θ}. Note also that results concerning the filtered approximation $\hat{\Phi}_t^{\theta} = \sum_{i=1}^{p} w_t^{\theta}[i]\gamma_t^{\theta}[i]\delta_{X_t^{\theta}[i]}$ are more difficult to obtain because of the weights $w_t^{\theta}[i]$ which couple the particles together (this is equally true for standard particle filtering).

4 Numerical experiment

For reason of space, it is not possible to present here a detailed set of numerical simulations. We however give a very simple example which illustrates some of the differences between the algorithm presented in the previous section and more conventional uses of the particle filter (such as for tracking, etc.)

We consider, the case of a first order scalar Gaussian autoregressive model observed in additive uncorrelated Gaussian white noise, for which it is easily checked that

$$\begin{aligned} k^{\theta}(x_{t-1}, x_t) &= n(x_t; \beta + \phi(x_{t-1} - \beta), \sigma^2) \\ g^{\theta}(y_t|x_t, y_{t-1}) &= n(y_t; x_t, \rho^2) \end{aligned} \tag{29}$$

where β is the mean value of the hidden chain X_t, ϕ is the AR parameter, σ^2 the innovation variance and ρ^2 is the variance of the additive noise. We further assume that X_1 has pdf $n(\bullet; \beta, \sigma^2/(1-\phi^2))$ which corresponds to the stationary distribution of the hidden chain. We focus on the computation of the gradient of the log-likelihood, that is when the functions m_t^{θ}

and r_t^θ are fixed according to (8). For this model, it is possible to carry out the computations of section 2 exactly, using Gaussian formulas – see (Charalambous and Logothetis, 1998) for details. For this simple toy example it is thus possible to contrast the results of the proposed algorithm with exact (recursive) evaluations of the gradient of the log-likelihood.

Note that since we expect the maximum likelihood estimator to be consistent for this model, it implies that the gradient of the log-likelihood satisfies

$$\lim_{T\to\infty} \frac{1}{T} \nabla_\theta \log p^\theta(Y_{1:T}) \to L(\theta,\theta_*) \qquad (\text{in } \mathbb{P}^{\theta_*} \text{ probability}) \tag{30}$$

where L is a deterministic function which depends both on the test parameter value θ and on the actual parameter value θ_* under which the observations $Y_{1:T}$ are distributed. Eq. (30) shows that normalization by T is indeed necessary if we want to compare the results obtained on different time horizons T.

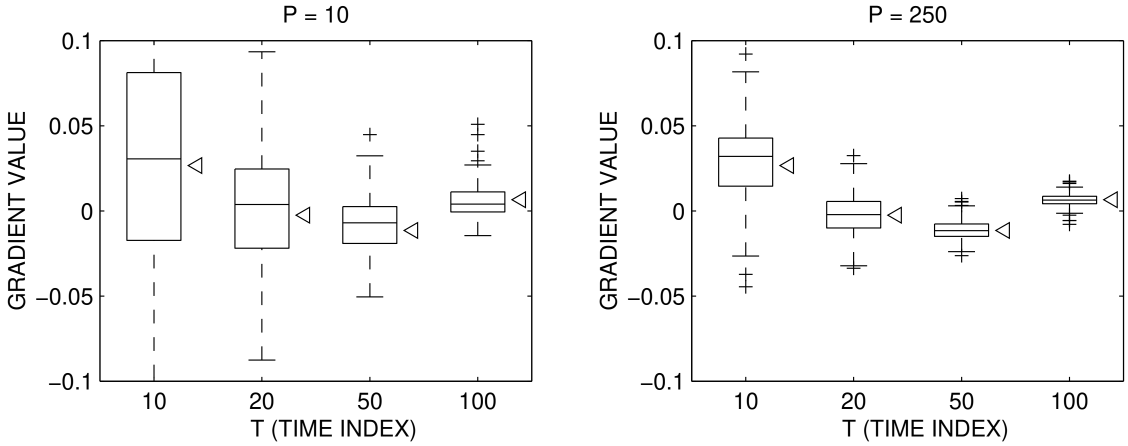

Figure 1: Box and whiskers plots summarizing 200 independent runs of the proposed algorithm compared with exact computations (triangles) , for different combination of p and T.

In the case of figure 1, the actual parameter vector is set to $(\beta,\phi,\sigma^2,\rho^2) = (1, 0.9, 0.05, 0.01)$ and the test parameter vector is $(\beta,\phi,\sigma^2,\rho^2) = (0.8, 0.8, 0.06, 0.015)$. The number of particles p is 10 for the left plot, and 250 for the right plot. The time horizon T varies, in each plot, from 10 to 100, and all simulations use the same observation sequence. As explained above, the quantity displayed is $\frac{1}{T}\nabla_\theta \log p^\theta(Y_{1:T}) - L(\theta,\theta_*)$, as computed by implementing the recursion in section 2 (triangles), and as approximated with the proposed algorithm (box and whiskers plots). $L(\theta,\theta_*)$ is determined empirically by running the exact gradient recursions for up to 100 000 observations. Note that only the component of the gradient corresponding to the first parameter (β) is shown since the situation is comparable for the other components of the gradient. The box and whiskers plots correspond to 200 independent Monte Carlo runs of the proposed algorithm and thus give and idea of the stochastic variability due to the particle approximation.

Comparing the left and right plots in figure 1, clearly shows that the precision of the approximation of the gradient improves when augmenting the number of particles. Since the square root of p is increased by a factor 5 in the right plot compared to the left one, the reduction in variance appears to be compatible with the results obtained for the standard bootstrap filter on some models (for which a CLT with $\sqrt{p}$ normalization was shown to hold) (Del Moral and Miclo, 2000). In this application however, increasing the number of particles is not the only source of stochastic averaging: When looking at any of the two plots for different values of the time horizon T, one clearly sees that the normalized gradient gets closer to $L(\theta,\theta_*)$ as T increases (which means that, on figure 1, the triangles gets closer to zero with increasing observation sizes), which is expected from (30). There is thus an interplay between the number of particles p and the time horizon T which makes fixing the number of particles for practical applications a challenging issue. One final difference with more standard uses of particle filters is the fact that the observations are not distributed under $\mathbb{P}^{\theta}$ (the test parameter value) but under $\mathbb{P}^{\theta_*}$. This point is crucial because we need to be able to reliably approximate the gradient of the log-likelihood for different values θ of the parameter. Figure 1, suggests that the proposed algorithm is reasonably efficient in this respect (since in this example θ is indeed different from θ_*) but robustness wrt large deviations of θ from θ_* is certainly an aspect which deserves more investigations.

5 Conclusion

We recall the complete algorithm for recursively approximating general Q_T^{θ} functionals: First, initialize the recursion with,

$$\begin{aligned}
X_1^{\theta}[i] &\sim \pi_1^{\theta} \\
w_1^{\theta}[i] &= \frac{h^{\theta}(Y_1|X_1^{\theta}[i])}{\sum_{i=1}^{p} h^{\theta}(Y_1|X_1^{\theta}[i])} \\
\gamma_1^{\theta}[i] &= m_1^{\theta}(X_1^{\theta}[i])
\end{aligned} \tag{31}$$

for $i = 1, \ldots p$ (with independent draws for $X_1^{\theta}[i]$). Then, for $t \geq 1$,

$$\begin{aligned}
I_t^{\theta}[i] &\sim \text{Discrete}(w_t^{\theta}[1], \ldots, w_t^{\theta}[p])) \\
X_{t+1}^{p} &\sim K^{\theta}_{X_t^{\theta}[I_t^{\theta}[i]]} \\
w_{t+1}^{\theta}[i] &= \frac{g^{\theta}(Y_{t+1}|X_{t+1}^{\theta}[i], Y_t)}{\sum_{j=1}^{p} g^{\theta}(Y_{t+1}|X_{t+1}^{\theta}[j], Y_t)} \\
\gamma_{t+1}^{\theta}[i] &= \gamma_t^{\theta}\left[I_t^{\theta}[i]\right] + r_{t+1}^{\theta}\left(X_t^{\theta}\left[I_t^{\theta}[i]\right], X_{t+1}^{\theta}[i]\right) + m_{t+1}^{\theta}(X_{t+1}^{\theta}[i])
\end{aligned} \tag{32}$$

for $i = 1, \ldots p$ (where the random draws are independent and conditionally independent from previous draws). For any time index t, the partial smoothed functional may be evaluated by computing (26).

The above algorithm provides an efficient solution for computing recursively the functionals that are needed for likelihood-based estimation of partially observed Markovian processes in general settings. Because of its resemblance with the standard bootstrap filter, it is expected that the finite horizon behavior of the method may be analyzed using the tools developed by (Del Moral and Miclo, 2000). The long-term behavior (large values of t) of the particle approximation is probably different however from the case considered by (Del Moral and Miclo, 2000) because of the limiting behavior of Q_t^{θ} – see (Douc et al., 2000) for recent results on that point. Finally, dependence in the parameter value θ is also an important issue.

Acknowledgment

The author wishes to thank his colleague Eric Moulines for his help and comments.

References

Brockwell, P. J. and Davis, R. A. (1991). *Time series: Theory and methods.* Springer.

Cappé, O., Buchoux, V., and Moulines, E. (1998). Quasi-newton method for maximum likelihood estimation of hidden markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Seattle.

Cappé, O., Doucet, A., Lavielle, M., and Moulines, E. (1999). Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing*, 73(1-2):3–25.

Charalambous, C. D. and Logothetis, A. (1998). New finite dimensional filters for the log-likelihood gradient, Hessian and Fisher information matrices: the discrete time case. In *Proc. IEEE Conf. Decision Control*, Tampa.

Collings, I. B. and Ryden, T. (1998). A New Maximum Likelihood Gradient Algorithm for On-Line Hidden Markov Model Identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Seattle.

De Jong, P. (1989). Smoothing and interpolation with the state space model. *J. Amer. Statist. Assoc.*, 84:1085–1088.

Del Moral, P. and Miclo, L. (2000). *Branching and Interacting Particle Systems Approximations of Feynman-Kac Formulae with Applications to Non-Linear Filtering*, volume 1729 of *Lecture Notes in Mathematics*. Springer, to appear.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–38 (with discussion).

Douc, R., Moulines, E., and Rydén, T. (2000). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. preprint.

Doucet, A., de Freitas, N., and Gordon, N., editors (2000). *Sequential Monte Carlo Methods in Practice*. Springer, to appear.

Elliott, R. J. and Krishnamurthy, V. (1999). New finite dimensional filters for estimation of linear gauss-markov models. *IEEE Trans. Automatic Control*, 44(5):938–951.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hürzeler, M. and Kunsch, H. R. (2000). Approximating and maximising the likelihood for a general state space model. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*. Springer.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility, likelihood inference and comparison with arch models. *Review of Economic Studies*, 65:361–394.

LeGland, F. and Mevel, L. (1997). Recursive estimation in HMMs. In *Proc. IEEE Conf. on Decision and Control*, San Diego.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov models and other models for discrete-valued time series*. Chapman & Hall.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filter. *J. Amer. Statist. Assoc.*, 94:590–599.

Zeitouni, O. and Dembo, A. (1988). Exact filters for the estimation of the number of transitions of finite-state continuous-time markov processes. *IEEE Trans. Inform. Theory*, 34(4).