Marginal Likelihood Computation for Hidden Markov Models via Generalized Two-Filter Smoothing

BY ADAM PERSING[1] & AJAY JASRA[2]

[1]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK.

E-Mail: *a.persing11@ic.ac.uk*

[2]Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG.

E-Mail: *staja@nus.edu.sg*

Abstract

In this note we introduce an estimate for the marginal likelihood associated to hidden Markov models (HMMs) using sequential Monte Carlo (SMC) approximations of the generalized two-filter smoothing decomposition [3]. This estimate is shown to be unbiased and a central limit theorem (CLT) is established. This latter CLT also allows one to prove a CLT associated to estimates of expectations w.r.t. a marginal of the joint smoothing distribution; these form some of the first theoretical results associated to the SMC approximation of the generalized two-filter smoothing decomposition. The new estimate and its application is investigated from a numerical perspective.

Key Words: Marginal Likelihood, Sequential Monte Carlo, Generalized Two-Filter Smoothing

1 Introduction

Hidden Markov models provide a flexible description of a wide variety of real-life phenomena; see [4]. An HMM is a pair of discrete-time stochastic processes, $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 1}$, where $X_n \in \mathbb{R}^{d_x}$ is an unobserved process and $y_n \in \mathbb{R}^{d_y}$ is observed. The hidden process $\{X_n\}_{n\geq 0}$ is a Markov chain with initial density δ_{x_0} at time 0 and transition density $f_\theta(x_n|x_{n-1})$, with $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ i.e. $\mathbb{P}_\theta(X_0 \in A) = \delta_{x_0}(A)$ and $\mathbb{P}_\theta(X_n \in A|X_{n-1} = x_{n-1}) = \int_A f_\theta(x_n|x_{n-1})dx_n$ $n \geq 1$ where $\mathbb{P}_\theta$ denotes probability, $A \subseteq \mathbb{R}^{d_x}$, δ_{x_0} is the Dirac measure with mass at x_0, and dx_n an assumed dominating measure. In addition, the observations $\{Y_n\}_{n\geq 1}$ conditioned upon $\{X_n\}_{n\geq 0}$ are statistically independent and have marginal density $g_\theta(y_n|x_n)$, i.e.$\mathbb{P}_\theta(Y_n \in B|\{X_k\}_{k\geq 0} = \{x_k\}_{k\geq 1}) = \int_B g_\theta(y_n|x_n)dy_n$ $n \geq 1$ with $B \subseteq \mathbb{R}^{d_y}$ and dy_n the dominating measure. The HMM described above is often referred to in the literature as a state-space model. Here θ is a static parameter, which is fixed throughout and we shall only be concerned with scenario that one observes a batch data set $y_{1:T} := (y_1, \ldots, y_T)$. The joint density of the observations $p_\theta(y_{1:T})$ is termed the marginal likelihood. For most models of practical interest, this quantity cannot be evaluated exactly. A popular collection of approximation techniques for HMMs, which can estimate the marginal likelihood are SMC methods.

SMC techniques simulate a collection of N samples in parallel, sequentially in time and combine importance sampling and resampling to approximate a sequence of probability distributions of increasing state-space known up-to an additve constant; see [9] for an introduction. These techniques provide a natural estimate of the marginal likelihood of HMMs (as well as for normalizing constants of Feynman-Kac representations; see [6]). The estimate is quite well understood and is known to be unbiased [6] and the relative variance is known to increase linearly with T [5, 12]. However, the standard SMC estimate is not the only alternative one can consider. A relatively recent procedure designed for smoothing, is based upon the *generalized* two-filter decomposition (see e.g. [2] for the two-filter smoothing decomposition). Roughly, the idea is to run two independent SMC algorithms, one forwards (as before) and one backwards (which approximates a collection of appropriately defined target distributions) and for them to 'meet' at some point. Using this procedure, one can yield more efficient schemes for smoothing, relative to standard SMC procedures. In the following note we:

1. Introduce a new estimate, costing $\mathcal{O}(N)$, of the marginal likelihood using the generalized two-filter smoothing decomposition.
2. Establish that this estimate is unbiased and prove a CLT, under some assumptions.
3. Numerically investigate the estimate.

It is remarked that via 2. we can also establish a CLT for an estimate of expectations w.r.t. a marginal of the joint smoothing distribution.

This note is in two halves; the first focuses on the idea from a methodological perspective. The second is the proof of our results in point 2. The note is structured as follows: in Section 2 we discuss the estimate and our main

result. In Section 3 some simulations investigating the new estimate are given; in particular, some comparisons to the forward filtering backward simulation (FFBSi) algorithm in [8]. The proofs of our results are housed in the appendix.

2 SMC and Generalized Two-Filter Smoothing

2.1 SMC Algorithm

We consider the joint smoothing distribution, with θ fixed:

$$\pi_\theta(x_{1:T}|y_{1:T}) = \frac{\prod_{n=1}^T g_\theta(y_n|x_n) f_\theta(x_n|x_{n-1})}{\int_{\mathbb{R}^{Td_x}} \prod_{n=1}^T g_\theta(y_n|x_n) f_\theta(x_n|x_{n-1}) dx_{1:T}} \tag{1}$$

the denominator is denoted $p_\theta(y_{1:T})$; this is the marginal likelihood. We remark that throughout, the transition and observation densities can be time-inhomogeneous, but we omit this from our notation. One can construct an SMC algorithm to sample sequentially from $\pi_\theta(x_1|y_1), \ldots, \pi_\theta(x_{1:T}|y_{1:T})$. The idea is to use a collection of particles, simulated in parallel, which are written $(\overrightarrow{X}^i_{1:n})_{i\in\{1,\ldots,N\}}$ to denote samples forward in time, the reason for the notation will become apparent below. We will sometimes denote the index of a particle at time n by $\overrightarrow{a}^i_n$, and we adopt the notation $\overrightarrow{x}_n^{\overrightarrow{a}^i_n} = \overrightarrow{x}_n^{a(i)}$.

- Step 1: For $i \in \{1, \ldots, N\}$ sample $\overrightarrow{X}^i_1 \sim q_{1,\theta}(\cdot)$ and compute the un-normalized weight:

$$\overrightarrow{W}^i_1 = \frac{g_\theta(y_1|\overrightarrow{x}^i_1) f_\theta(\overrightarrow{x}^i_1|x_0)}{q_{1,\theta}(\overrightarrow{x}^i_1)}.$$

  For $i \in \{1, \ldots, N\}$ sample $\overrightarrow{a}^i_1 \in \{1, \ldots, N\}$ from a discrete distribution on $\{1, \ldots, N\}$ with jth probability $\overrightarrow{w}^j_1 = \overrightarrow{W}^j_1 / \sum_{l=1}^N \overrightarrow{W}^l_1$ these represent the resampled particles. Set $n = 2$.

- Step 2: If $n = T + 1$ stop. Otherwise, for $i \in \{1, \ldots, N\}$ sample $\overrightarrow{X}^i_n | \overrightarrow{x}^{a(i)}_{n-1} \sim q_{n,\theta}(\cdot|\overrightarrow{x}^{a(i)}_{n-1})$ and compute the un-normalized weight:

$$\overrightarrow{W}^i_n = \frac{g_\theta(y_n|\overrightarrow{x}^i_n) f_\theta(\overrightarrow{x}^i_n|\overrightarrow{x}^{a(i)}_{n-1})}{q_{n,\theta}(\overrightarrow{x}^i_n|\overrightarrow{x}^{a(i)}_{n-1})}.$$

  For $i \in \{1, \ldots, N\}$ sample $\overrightarrow{a}^i_n \in \{1, \ldots, N\}$ from a discrete distribution on $\{1, \ldots, N\}$ with jth probability $\overrightarrow{w}^j_n = \overrightarrow{W}^j_n / \sum_{l=1}^N \overrightarrow{W}^l_n$. Set $n = n + 1$ and return to the start of step 2.

The estimate of the marginal likelihood is:

$$p^N_\theta(y_{1:T}) = \prod_{n=1}^T \left(\frac{1}{N} \sum_{l=1}^N \overrightarrow{W}^l_n \right). \tag{2}$$

2.2 Generalized Two-Filter Smoothing

It is well-known that the above SMC algorithm does not approximate the joint smoothing distribution at all well. One technique which is known to assist the simulation procedure (at least empirically) is the SMC approximation of the generalized two-filter representation [3]. The algorithm works by defining two filters. One works as the SMC algorithm above and moves forward in time. The other works backward in time, on a sequence of densities defined below. These two algorithms 'meet' at some pre-specified time $t \in \{1, \ldots, T\}$.

Define the following sequence of densities (we use the convention $\prod_\emptyset = 1$):

$$\widetilde{\pi}_\theta(x_{n:T}|y_{n:T}) \propto \xi_{n,\theta}(x_t) g_\theta(y_n|x_n) \left[\prod_{n=t+1}^T g_\theta(y_n|x_n) f_\theta(x_n|x_{n-1}) \right] \quad n \in \{t, \ldots, T\}$$

where, at this stage, $\xi_{n,\theta}$ are a sequence of (essentially) arbitrary density functions w.r.t. dx_n. In practice, the $\xi_{n,\theta}$ are critical to the efficiency of the algorithm and we return to this point in Section 3. We write the normalizing constant as $\tilde{p}_\theta(y_{n:T})$. One can use SMC to approximate this sequence of densities. We will sometimes denote the index of a particle at time n by $\overleftarrow{a}^i_n$, and we adopt the notation $\overleftarrow{x}_n^{\overleftarrow{a}^i_n} = \overleftarrow{x}_n^{a(i)}$.

- Step 1: For $i \in \{1, \ldots, N\}$ sample $\overleftarrow{X}_T^i \sim q_{T,\theta}(\cdot)$ and compute the un-normalized weight:

$$\overleftarrow{W}_T^i = \frac{\xi_{T,\theta}(\overleftarrow{x}_T^i) g_\theta(y_T | \overleftarrow{x}_T^i)}{q_{T,\theta}(\overleftarrow{x}_1^i)}.$$

  For $i \in \{1, \ldots, N\}$ sample $\overleftarrow{a}_T^i \in \{1, \ldots, N\}$ from a discrete distribution on $\{1, \ldots, N\}$ with jth probability $\overleftarrow{w}_T^j = \overleftarrow{W}_T^j / \sum_{l=1}^N \overleftarrow{W}_T^l$ these represent the resampled particles. Set $n = T - 1$.

- Step 2: If $n = t - 1$ stop. Otherwise For $i \in \{1, \ldots, N\}$ sample $\overleftarrow{X}_n^i | \overleftarrow{x}_{n+1}^{a(i)} \sim q_{n,\theta}(\cdot | \overleftarrow{x}_{n+1}^{a(i)})$ and compute the un-normalized weight:

$$\overleftarrow{W}_n^i = \frac{\xi_{n,\theta}(\overleftarrow{x}_n^i) g_\theta(y_n | \overleftarrow{x}_n^i) f_\theta(\overleftarrow{x}_{n+1}^{a(i)} | \overleftarrow{x}_n^i)}{\xi_{n+1,\theta}(\overleftarrow{x}_{n+1}^{a(i)}) q_{n,\theta}(\overrightarrow{x}_n^i | \overleftarrow{x}_{n+1}^{a(i)})}.$$

  For $i \in \{1, \ldots, N\}$ sample $\overleftarrow{a}_n^i \in \{1, \ldots, N\}$ from a discrete distribution on $\{1, \ldots, N\}$ with jth probability $\overleftarrow{w}_n^j = \overleftarrow{W}_n^j / \sum_{l=1}^N \overleftarrow{W}_n^l$. Set $n = n - 1$ and return to the start of step 2.

One can estimate the normalizing constant $\tilde{p}_\theta(y_{t+1:T})$ by using a similar expression to (2); this estimate is denoted $\tilde{p}_\theta^N(y_{t+1:T})$.

2.3 Two Estimates of the Marginal Likelihood

The objective here is to consider how one can use generalized two-filter smoothing to estimate the marginal likelihood. One can consider [3, Proposition 3] which states that

$$p_\theta(y_{1:T}) = \int \pi_\theta(x_{t-1}, y_{1:t-1}) \widetilde{\pi}_\theta(x_t, y_{t:T}) \frac{f_\theta(x_t | x_{t-1})}{\xi_{t,\theta}(x_t)} dx_{t-1:t}.$$

After some standard calculations, one has

$$\begin{aligned} p_\theta(y_{1:T}) &= p_\theta(y_{1:t-2}) \widetilde{p}_\theta(y_{t+1:T}) \int q_{t-1,\theta}(x_{t-1} | x_{t-2}) q_{t,\theta}(x_t | x_{t+1}) \overrightarrow{W}_{t-1}(x_{t-2:t-1}) \overleftarrow{W}_t(x_{t:t+1}) \times \\ &\quad \pi_\theta(x_{t-2} | y_{1:t-2}) \widetilde{\pi}_\theta(x_{t+1} | y_{t+1:T}) \frac{f(x_t | x_{t-1})}{\xi_{t,\theta}(x_t)} dx_{t-2:t+1} \end{aligned}$$

whence an SMC estimate, one filter run up-to time $t-1$ forward and the other run backward to time t with no resampling at the final time step only, of the marginal likelihood is

$$p_\theta^N(y_{1:T}) = p_\theta^N(y_{1:t-2}) \widetilde{p}_\theta^N(y_{t+1:T}) \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \overrightarrow{W}_{t-1}(\overrightarrow{x}_{t-2:t-1}^i) \overleftarrow{W}_t(\overleftarrow{x}_{t:t+1}^j) \frac{f(\overleftarrow{x}_t^j | \overrightarrow{x}_{t-1}^i)}{\xi_{t,\theta}(\overleftarrow{x}_t^j)}$$

This estimate is perhaps slightly undesirable as it has a computational cost of $\mathcal{O}(N^2)$.

An alternative approach is to use the slightly modified representation

$$p_\theta(y_{1:T}) = p_\theta(y_{1:t-1}) \widetilde{p}_\theta(y_{t+1:T}) \int \pi_\theta(x_{t-1} | y_{1:t-1}) \widetilde{\pi}_\theta(x_{t+1} | y_{t+1:T}) \frac{f_\theta(x_t | x_{t-1}) f_\theta(x_{t+1} | x_t)}{\xi_{t+1,\theta}(x_{t+1})} g_\theta(y_t | x_t) dx_{t-1:t+1}.$$

Again, after some simple manipulations, one arrives at the formula

$$\begin{aligned} p_\theta(y_{1:T}) &= p_\theta(y_{1:t-2}) \widetilde{p}_\theta(y_{t+2:T}) \int \pi_\theta(x_{t-2} | y_{1:t-2}) \widetilde{\pi}_\theta(x_{t+2} | y_{t+2:T}) \overrightarrow{W}_{t-1}(x_{t-2:t-1}) \overleftarrow{W}_{t+1}(x_{t+1:t+2}) \times \\ &\quad q_{t-1,\theta}(x_{t-1} | x_{t-2}) q_{t+1,\theta}(x_{t+1} | x_t) \frac{f_\theta(x_t | x_{t-1}) f_\theta(x_{t+1} | x_t)}{\xi_{t+1,\theta}(x_{t+1})} g_\theta(y_t | x_t) dx_{t-2:t+2}. \end{aligned}$$

Now, if one runs the two forward and backward SMC algorithms up-to times $t-1$ and $t+1$ respectively, not resampling at the very final time steps, one has the approximation:

$$\begin{aligned} p_\theta^N(y_{1:T}) &= p_\theta^N(y_{1:t-2}) \widetilde{p}_\theta^N(y_{t+2:T}) \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \overrightarrow{W}_{t-1}(\overrightarrow{x}_{t-2:t-1}^i) \overleftarrow{W}_{t+1}(\overleftarrow{x}_{t+1:t+2}^j) \times \\ &\quad \int \frac{f_\theta(x_t | \overrightarrow{x}_{t-1}^i) f_\theta(\overleftarrow{x}_{t+1}^j | x_t)}{\xi_{t+1,\theta}(\overleftarrow{x}_{t+1}^j)} g_\theta(y_t | x_t) dx_t. \end{aligned}$$

This quantity can be approximated using the following procedure in [10]. Consider a conditional density $q_{t,\theta}(x_t|x_{t-1},x_{t+1})$ and two probabilities $\overrightarrow{\beta}^i_{t-1}$, $\overleftarrow{\beta}^j_{t+1}$, $i,j\in\{1,\ldots,N\}$ $\sum_{i=1}^N \overrightarrow{\beta}^i_{t-1}=1$, $\sum_{j=1}^N \overleftarrow{\beta}^j_{t+1}=1$. Sample $i(1),j(1),\ldots,i(N),j(N)$ using the $\overrightarrow{\beta}^i_{t-1}$, $\overleftarrow{\beta}^j_{t+1}$ and then, for each pair $i(l),j(l)$ sample $X_t^l|x_{t-1}^{i(l)},x_{t+1}^{j(l)}$ from the distribution induced by $q_{t,\theta}(\cdot|x_{t-1}^{i(l)},x_{t+1}^{j(l)})$, which leads to the estimate, which only costs $\mathcal{O}(N)$:

$$\begin{aligned} p_\theta^N(y_{1:T}) &= p_\theta^N(y_{1:t-2})\widetilde{p}_\theta^N(y_{t+2:T})\frac{1}{N}\sum_{l=1}^N \frac{1}{N^2}\overrightarrow{W}_{t-1}(\overrightarrow{x}^{i(l)}_{t-2:t-1})\overleftarrow{W}_{t+1}(\overleftarrow{x}^{j(l)}_{t+1:t+2})\times \\ &\frac{f_\theta(x_t^l|\overrightarrow{x}^{i(l)}_{t-1})f_\theta(\overleftarrow{x}^{j(l)}_{t+1}|x_t^l)}{\xi_{t+1,\theta}(\overleftarrow{x}^{j}_{t+1})\overrightarrow{\beta}^{i(l)}_{t-1}\overleftarrow{\beta}^{j(l)}_{t+1}q_{t,\theta}(x_t^l|x_{t-1}^{i(l)},x_{t+1}^{j(l)})}g_\theta(y_t|x_t^l). \end{aligned} \tag{3}$$

2.4 Unbiasedness and Central Limit Theorem

We will give some analysis of the estimate (3); we denote this estimate $p_\theta^N(y_{1:T})$. We make an assumption (A1) which is detailed in the appendix. In addition, the notations for the expression of the asymptotic variance are also defined in the appendix. For a function $\varphi:\mathbb{R}^d\to\mathbb{R}$ such that $\sup_{x\in\mathbb{R}^d}|\varphi(x)|<+\infty$, we write $\varphi\in\mathcal{B}_b(\mathbb{R}^d)$. $\mathcal{N}_d(\mu,\Sigma)$ denotes a $d-$dimensional normal distribution with mean μ and covariance Σ; if $d=1$ the subscript d is omitted.

Theorem 2.1. *We have*

$$\mathbb{E}[p_\theta^N(y_{1:T})]=p_\theta(y_{1:T})\quad\forall\theta\in\Theta.$$

In addition, assume (A1). Then for fixed $T>2$, $t\in\{3,\ldots,T-2\}$ and any $\theta\in\Theta$ we have that

$$\sqrt{N}(p_\theta^N(y_{1:T})-p_\theta(y_{1:T}))\Rightarrow Z_\theta$$

where $Z_\theta\sim\mathcal{N}(0,\sigma^2_{t,T}(\theta))$ with

$$\sigma^2_{t,T}(\theta) = \sigma^2_{\overrightarrow{\gamma}_{t-1,\theta}}\big(\overrightarrow{W}_{t-1}\overleftarrow{\gamma}_{t+1,\theta}[\overleftarrow{W}^\xi_{t+1}I_{gf}(.,\cdot)]\big)+\sigma^2_{\overleftarrow{\gamma}_{t+1,\theta}}\big(\overleftarrow{W}^\xi_{t+1}\overrightarrow{\gamma}_{t-1,\theta}(\overrightarrow{W}_{t-1}I_{gf}(\cdot,.))\big)$$

where, $\varphi\in\mathcal{B}_b(\mathbb{R}^{2d_x})$

$$\begin{aligned} \sigma^2_{\overrightarrow{\gamma}_{t-1,\theta}}(\varphi) &= \sum_{q=1}^{t-1}\overrightarrow{\gamma}_{q,\theta}(1)^2\overrightarrow{\eta}_{q,\theta}\bigg(\Big[\overrightarrow{Q}_{q,t-1}(\varphi)-\overrightarrow{\eta}_{q,\theta}(\overrightarrow{Q}_{q,t-1}(\varphi))\Big]^2\bigg) \\ \sigma^2_{\overleftarrow{\gamma}_{t,\theta}}(\varphi) &= \sum_{q=0}^{T-t-1}\overleftarrow{\gamma}_{T-q,\theta}(1)^2\overleftarrow{\eta}_{T-q,\theta}\bigg(\Big[\overleftarrow{Q}_{T-q,t+1}(\varphi)-\overleftarrow{\eta}_{T-q,\theta}\overleftarrow{Q}_{T-q,t+1}(\varphi)\Big]^2\bigg). \end{aligned}$$

Remark 2.1. *Under some additional mixing conditions, one may establish that the asymptotic variance $\sigma^2_{t,T}(\theta)$ when divided by $p_\theta(y_{1:T})^2$ (i.e. the asymptotic variance associated to a normalized estimate) obeys the following inequality: $\sigma^2_{t,T}(\theta)/p_\theta(y_{1:T})^2\leq C_1(\theta)(t-1)+C_2(\theta)(T-t)$ where the first term is the error from the forward filter and the second from the backward filter. Unfortunately, this provides little intuition on how to select t and it simply implies that if the forward algorithm works better, one should choose t large and vice versa.*

Remark 2.2. *Let $\varphi:\mathbb{R}^{d_x}\to\mathbb{R}$, $\varphi\in\mathcal{B}_b(\mathbb{R}^{d_x})$, and consider $\mathbb{E}_\theta[\varphi(X_t)|y_{1:T}]$ where $3\leq t\leq T-2$ and the expectation is w.r.t. the joint smoothing distribution, with density (1). Using the ideas in Section 2.3 one can show that an estimator of $\mathbb{E}_\theta[\varphi(X_t)|y_{1:T}]$ is*

$$\frac{1}{p_\theta^N(y_{1:T})}p_\theta^N(y_{1:t-2})\widetilde{p}_\theta^N(y_{t+2:T})\frac{1}{N^3}\sum_{l=1}^N\overrightarrow{W}_{t-1}(\overrightarrow{x}^{i(l)}_{t-2:t-1})\overleftarrow{W}_{t+1}(\overleftarrow{x}^{j(l)}_{t+1:t+2})\frac{\varphi(x_t^l)f_\theta(x_t^l|\overrightarrow{x}^{i(l)}_{t-1})f_\theta(\overleftarrow{x}^{j(l)}_{t+1}|x_t^l)g_\theta(y_t|x_t^l)}{\xi_{t+1,\theta}(\overleftarrow{x}^{j}_{t+1})\overrightarrow{\beta}^{i(l)}_{t-1}\overleftarrow{\beta}^{j(l)}_{t+1}q_{t,\theta}(x_t^l|x_{t-1}^{i(l)},x_{t+1}^{j(l)})}.$$

Denote the estimate as $p^N_{\theta,t}(\varphi)/p^N_\theta(y_{1:T})$ and set $\mathbb{E}_\theta[\varphi(X_t)|y_{1:T}]=p_{\theta,t}(\varphi)/p_\theta(y_{1:T})$. Standard calculations reveal (e.g. [6, pp. 301]) that

$$\frac{p^N_{\theta,t}(\varphi)}{p^N_\theta(y_{1:T})}-\frac{p_{\theta,t}(\varphi)}{p_\theta(y_{1:T})}=\frac{p_\theta(y_{1:T})}{p^N_\theta(y_{1:T})}p^N_{\theta,t}\bigg(\frac{1}{p_\theta(y_{1:T})}\Big[\varphi-\frac{p_{\theta,t}(\varphi)}{p_\theta(y_{1:T})}\Big]\bigg).$$

Now, upon inspection of the proofs in the appendix, one can easily deduce:

- $p_\theta(y_{1:T})/p_\theta^N(y_{1:T})$ *will converge in probability to 1.*
- *Let* $\widetilde{\varphi} = 1/p_\theta(y_{1:T})[\varphi - p_{\theta,t}(\varphi)/p_\theta(y_{1:T})]$*, then*

$$\sqrt{N}p_{\theta,t}^N\left(\frac{1}{p_\theta(y_{1:T})}\left[\varphi - \frac{p_{\theta,t}(\varphi)}{p_\theta(y_{1:T})}\right]\right) \Rightarrow Z_\theta(\widetilde{\varphi})$$

where $Z_\theta(\widetilde{\varphi}) \sim \mathcal{N}(0, \sigma_{t,T}^2(\widetilde{\varphi}))$*,*

$$\sigma_{t,T}^2(\theta) = \sigma^2_{\overrightarrow{\gamma}_{t-1,\theta}}\left(\overrightarrow{W}_{t-1}\overleftarrow{\gamma}_{t+1,\theta}[\overleftarrow{W}^{\xi}_{t+1}I_{gf\widetilde{\varphi}}(.,\cdot)]\right) + \sigma^2_{\overleftarrow{\gamma}_{t+1,\theta}}\left(\overleftarrow{W}^{\xi}_{t+1}\overrightarrow{\gamma}_{t-1,\theta}(\overrightarrow{W}_{t-1}I_{gf\widetilde{\varphi}}(\cdot,.))\right)$$

and for $(\tilde{x}_{t-1}, \tilde{x}_{t+1}) \in \mathbb{R}^{2d_x}$

$$I_{gf\widetilde{\varphi}}(\tilde{x}_{t-1}, \tilde{x}_{t+1}) = \int_{\mathbb{R}^{d_x}} g_\theta(y_t|x_t)\widetilde{\varphi}(x_t)f_\theta(\tilde{x}_{t+1}|x_t)f_\theta(x_t|\tilde{x}_{t-1})dx_t.$$

See the appendix for further definitions of the notations.

Hence, on using Slutsky's Lemma, one has a univariate CLT for an approximation of $\mathbb{E}_\theta[\varphi(X_t)|y_{1:T}]$*. One can follow the ideas of [6, pp. 301-302] to prove a multivariate CLT. It may be possible to compare this estimate (through the asymptotic variance) relative to the one produced by the forward filtering backward smoothing algorithm; see [7, 8].*

Remark 2.3. *The unbiased property allows one to use the SMC approximation of the generalized two-filter representation within a particle Markov chain Monte Carlo [1] algorithm. In [11], we have established an appropriate target distribution in this context.*

3 Numerical Examples

3.1 Measuring the New Estimate's Sensitivity to t

Consider the linear Gaussian model provided in Section 4 of [10]: $X_0 \sim \mathcal{N}_2(\mu_0, \Sigma_0)$, $X_{n+1} \mid (X_{1:n} = x_{1:n}, Y_{1:n} = y_{1:n}) \sim \mathcal{N}_2(Fx_n, Q)$, $Y_n \mid (X_{1:n} = x_{1:n}, Y_{1:n-1} = y_{1:n-1}) \sim \mathcal{N}(Gx_n, R)$, with

$$G = (1, 0) \qquad R = \tau^2$$

$$F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad Q = \nu^2 \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

We ran the two-filter SMC algorithm to calculate the marginal likelihood via (3) for the instance where $T = 300$. Our objective is to observe how the choice of t affects the accuracy and precision of the new estimate. A Kalman filter is used to allow us to choose $\xi_{n,\theta}(x_n) = \pi_\theta(x_n|y_{1:n-1})$ (the predictor); this corresponds to an extremely favourable choice (indeed one recovers the FFBS procedure, when considering the smoother). In all simulations, we used the optimal importance distributions and β resampling weights as in Appendix A of [10]. We set $N = 300$. We ran nine versions of the two-filter algorithm, with $t \in \{T/10, 2T/10, \ldots, 9T/10\}$. In each case, we plotted the variability of the estimate and compared (3) to the maximum likelihood estimate provided by the Kalman filter. We ran many simulations for different pairs of values of the state noise, ν^2, and the observation noise, τ^2; specifically, we looked at 225 possible pairings where ν^2 and τ^2 each ranged from 1 to 98. The results are displayed in Figure 3.2.

We found the same phenomenon across all pairs of values of ν^2 and τ^2. There is an increase in the variance of (3) as t approaches T (i.e., when the new estimate relies more on the forward filter and less on the backward filter). Furthermore, we see the accuracy of the estimate fall as t approaches T. These results are in accordance with the degeneracy measures of the two filters. The forward filter's effective sample size (ESS) stays around 200, while the backward filter's ESS *never* drops below $N = 300$ (due to the choice of $\xi_{n,\theta}$ and the various proposals adopted). The choice for $\xi_{n,\theta}$ ensures the backward filter's consistently strong performance. This suggests that a better algorithm may result from removing any dependence on the forward filter and running a backward filter (with $\xi_{n,\theta}(x_n) = \pi_\theta(x_n|y_{1:n-1})$) from time T to 1; this point is discussed in Section 4. Note that, in comparison to the standard SMC estimate, the results for the new estimate (for this model and the current settings) were superior w.r.t. the variability of the estimate (results not shown).

3.2 Comparing the Two-filter Decomposition to FFBSi

Remark 2.2 above parallels a similar result shown in [8] for another $\mathcal{O}(N)$ SMC smoothing approximation based on FFBS. To explore this point further, we used the same example from [10] to compare the two-filter SMC algorithm to the FFBSi algorithm in [8]. We used both algorithms to calculate the expected value of the state of the hidden process given $y_{1:T=300}$ at time $t \in \{T/10, \ldots, 9T/10\}$. Both algorithms utilized $N = 300$ particles. Note that FFBSi relies on rejection sampling, and so due to its stochastic running time, it is difficult to exactly match the computation times. Again, 50 simulations per algorithm per (τ^2, ν^2) pair for 225 pairs are run; see Figure 3.2. We found that the two algorithms gave very similar results, although the two-filter decomposition yielded estimates of lower variance (see Figure 3.2). This is especially true at lower values of t, where, as above, the backward filter has more influence on the two-filter estimate.

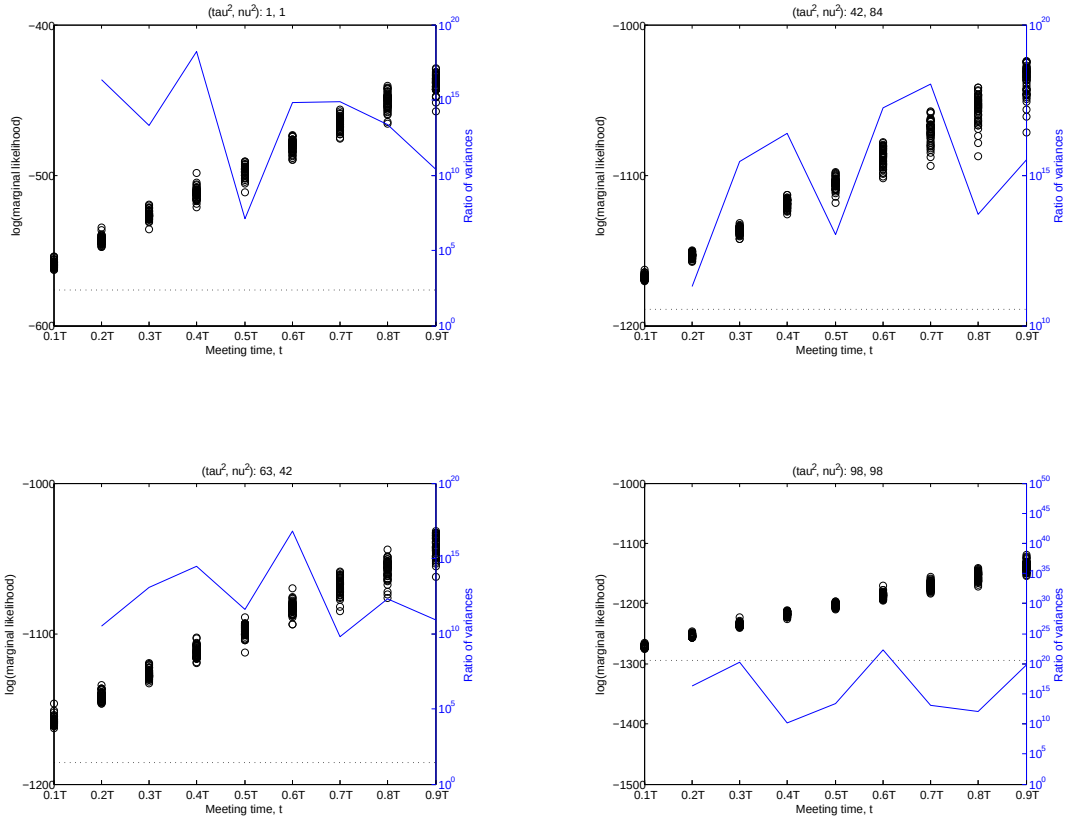

Figure 1: We present the output for some pairs of ν^2 and τ^2. The circles, whose scale is on the left, give the 50 simulated values of the logarithm of the marginal likelihood per time point. The solid line, whose scale is on the right, measures the ratio of the variance of these 50 values at each time point to the variance at the previous time point. The dotted line gives the logarithm of the marginal likelihood as provided by the Kalman filter.

4 Discussion

In this note, we introduced a new $\mathcal{O}(N)$ estimate of the marginal likelihood using the generalized two-filter decomposition. We established that this estimate is unbiased and proved a CLT, under some assumptions. Numerical examples suggested that the new estimate is sensitive to changes in the meeting point of the forward and backward filters. When choosing $\xi_{n,\theta}(x_n) = \pi_\theta(x_n|y_{1:T})$, the backward filter significantly outperforms the forward filter and it may be beneficial to remove the forward filter from the estimation procedure. However, one can seldom make this choice for the $\xi_{n,\theta}$, and so we would like to approximate them. In joint work with Prof. A. Doucet, we

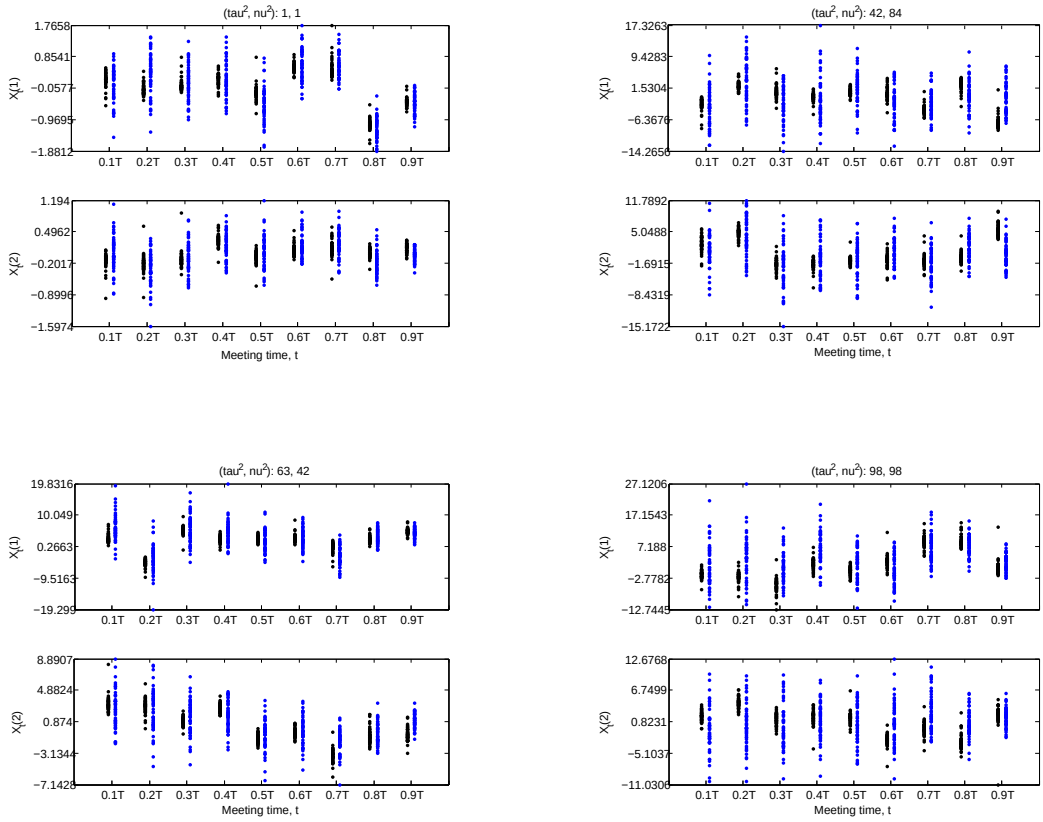

Figure 2: We present the output for some pairs of ν^2 and τ^2. At each time point, the black dots (left) represent 50 simulated expected values from the two-filter algorithm and the blue dots (right) represent 50 estimates from FFBSi. The first component of $\mathbb{E}[X_t|y_{1:T}]$ is on top, and the second component of $\mathbb{E}[X_t|y_{1:T}]$ is on the bottom.

are exploring a smoothing algorithm where one introduces a discrete valued auxiliary variable $J \in \{1, ..., N\}$ and considers the sequence of extended backward targets (where we condition upon the particles from a forward SMC algorithm) $\widetilde{\pi}_\theta(j, x_{n:T}|y_{1:T}) \propto f_\theta\left(\overrightarrow{x}_n \mid \overrightarrow{x}^j_{n-1}\right) g_\theta(y_n|x_n)\left[\prod_{n=t+1}^T g_\theta(y_n|x_n) f_\theta(x_n|x_{n-1})\right]$ $n \in \{t, \ldots, T\}$. The idea is to approximate the $\xi_{n,\theta}$ that are used above, via the forward filter.

Acknowledgements

This project has been initialized in joint research with Arnaud Doucet and we thank him for his input which has been critical. We thank two referees and an associate editor, whose comments have greatly enhanced the article.

A Proof of the CLT

Here we describe a Feynman-Kac representation, which is used in the proof of the CLT. Let $t \in \{3, \ldots, T-2\}$, with $T > 2$ also fixed.

Define, the forward Feynman-Kac un-normalized $n-$time marginal, $n \in \{1, \ldots, t-1\}$:

$$\overrightarrow{\gamma}_{n,\theta}(dx_n) = \int \left[\prod_{p=1}^{n-1} \overrightarrow{W}_p(x_p) M_p(x_{p-1}, dx_p)\right] M_n(x_{n-1}, dx_n)$$

with $x_p = (x'_p, \tilde{x}_p) \in \mathbb{R}^{2d_x}$, $M_1(x_0, dx_0) = \delta_{x_0}(dx'_1) q_{1,\theta}(\tilde{x}_1|x'_1) d\tilde{x}_1$ and

$$M_p(x_{p-1}, dx_p) \quad = \quad \delta_{\tilde{x}_{p-1}}(dx'_p) q_{p,\theta}(\tilde{x}_p|x'_p) d\tilde{x}_p.$$

The normalized operator is

$$\overrightarrow{\eta}_{n,\theta}(dx_n) = \overrightarrow{\gamma}_{n,\theta}(dx_n)/\overrightarrow{\gamma}_{n,\theta}(1).$$

We also define the forward semi-group operator:

$$\overrightarrow{Q}_{p,n}(x_p, dx_n) = \int \prod_{q=p}^{n-1} \overrightarrow{W}_q(x_q) M_{q+1}(x_q, dx_{q+1})$$

with $1 \leq p \leq n \leq t-1$. The selection mutation operator:

$$\overrightarrow{\Phi}_q(\overrightarrow{\eta}_{q-1,\theta})(\cdot) \quad = \quad \frac{\overrightarrow{\eta}_{q-1,\theta}(\overrightarrow{W}_{q-1} M_q(\cdot))}{\overrightarrow{\eta}_{q-1,\theta}(\overrightarrow{W}_{q-1})} \quad q \in \{0, \ldots, t-1\}$$

with the conventions $\overrightarrow{\Phi}_1(\overrightarrow{\eta}_{0,\theta}) = \overrightarrow{\eta}_{1,\theta}$.

Define, the backward Feynman-Kac un-normalized $n-$time marginal, $n \in \{0, \ldots, T-t-1\}$:

$$\overleftarrow{\gamma}_{T-n,\theta}(dx_n) = \int \left[\prod_{p=0}^{T-n-1} \overleftarrow{W}_{T-p}(x_{T-p}) M_{T-p}(x_{T-p+1}, dx_{T-p}) \right] M_n(x_{n+1}, dx_n)$$

with $x_n = (x'_n, \tilde{x}_n) \in \mathbb{R}^{2d_x}$, $M_T(d\tilde{x}_T) = q_T(\tilde{x}_T) d\tilde{x}_T \delta_x(dx'_T)$, $x \in \mathbb{R}^{d_x}$ an arbitrary point

$$M_n(x_{n+1}, dx_n) \quad = \quad q_{n,\theta}(\tilde{x}_n|x'_n) dx_n \delta_{\tilde{x}_{n+1}}(dx'_n) \quad n \in \{t+1, \ldots, T-1\}.$$

The normalized operator $\overleftarrow{\eta}_{T-n,\theta} = \overleftarrow{\gamma}_{T-n,\theta}(dx_n)/\overleftarrow{\gamma}_{T-n,\theta}(1)$. Also define the semi-group operator

$$\overleftarrow{Q}_{p,n}(x_p, dx_n) = \int \prod_{s=0}^{p-n-1} \overleftarrow{W}_{p-s}(x_{p-s}) M_{p-s-1}(x_{p-s}, dx_{p-s-1})$$

with $T \geq p \geq n \geq t+1$. Also

$$\overleftarrow{\Phi}_{T-q}(\overleftarrow{\eta}_{T-q+1,\theta})(\cdot) \quad = \quad \frac{\overleftarrow{\eta}_{T-q+1,\theta}(\overleftarrow{W}_{T-q+1} M_{T-q}(\cdot))}{\overleftarrow{\eta}_{T-q+1,\theta}(\overleftarrow{W}_{T-q+1})} \quad q \in \{0, \ldots, T-t-1\}$$

and $\overleftarrow{\Phi}_T(\overleftarrow{\eta}_{T+1}) = \overleftarrow{\eta}_T$.

We will use the notation

$$\begin{aligned} I_{gf}(\tilde{x}_{t-1}, \tilde{x}_{t+1}) \quad &= \quad \int_{\mathbb{R}^{d_x}} g_\theta(y_t|x_t) f_\theta(\tilde{x}_{t+1}|x_t) f_\theta(x_t|\tilde{x}_{t-1}) dx_t \\ W^\xi_{t+1}(x_{t+1}) \quad &= \quad \frac{\overleftarrow{W}_{t+1}(x_{t+1})}{\xi_{t+1,\theta}(\tilde{x}_{t+1})} \end{aligned}$$

with

$$\begin{aligned} \mu_{t-1}(\overrightarrow{W}_{t-1} \overleftarrow{\gamma}_{t+1,\theta}[\overleftarrow{W}^\xi_{t+1} I_{gf}(.,.)]) \quad &= \quad \int \mu_{t-1}(dx_{t-1}) \overrightarrow{W}_{t-1}(x_{t-1}) [\int \overleftarrow{\gamma}_{t+1,\theta}(dx_{t+1}) \overleftarrow{W}^\xi_{t+1}(x_{t+1}) I_{gf}(\tilde{x}_{t-1}, \tilde{x}_{t+1})] \\ \mu_{t+1}(\overleftarrow{W}^\xi_{t+1} \overrightarrow{\gamma}_{t-1,\theta}(\overrightarrow{W}_{t-1} I_{gf}(\cdot,.))) \quad &= \quad \int \mu_{t+1}(dx_{t+1}) \overleftarrow{W}^\xi_{t+1}(x_{t+1}) [\int \overrightarrow{\gamma}_{t-1,\theta}(dx_{t-1}) \overrightarrow{W}_{t-1}(x_{t-1}) I_{gf}(\tilde{x}_{t-1}, \tilde{x}_{t+1})] \end{aligned}$$

for $\sigma-$finite measures μ_{t-1}, μ_{t+1}.

Using the above notations, we can write

$$p^N_\theta(y_{1:T}) = \overrightarrow{\gamma}^N_{t-1}(1) \overleftarrow{\gamma}^N_{t+2}(1) \frac{1}{N} \sum_{l=1}^N \frac{\overrightarrow{W}_{t-1}(\overrightarrow{x}^{i(l)}_{t-1}) \overleftarrow{W}_{t+1}(\overleftarrow{x}^{j(l)}_{t+1}) f_\theta(x^l_t|\overrightarrow{x}^{i(l)}_{t-1}) f_\theta(\overleftarrow{x}^{j(l)}_{t+1}|x^l_t)}{N^2 \xi_{t+1,\theta}(\overleftarrow{x}^{j}_{t+1}) \overrightarrow{\beta}^{i(l)}_{t-1} \overleftarrow{\beta}^{j(l)}_{t+1} q_{t,\theta}(x^l_t|\overrightarrow{x}^{i(l)}_{t-1}, \overleftarrow{x}^{j(l)}_{t+1})} g_\theta(y_t|x^l_t)$$

with

$$\begin{aligned}\overrightarrow{\gamma}_{t-1}^N(1) &= \prod_{p=1}^{t-2}\frac{1}{N}\sum_{i=1}^N \overrightarrow{W}_p(\overrightarrow{x}_p^i)\\ \overleftarrow{\gamma}_{t+2}^N(1) &= \prod_{p=0}^{T-t-2}\frac{1}{N}\sum_{i=1}^N \overrightarrow{W}_{T-p}(\overrightarrow{x}_{T-p}^i).\end{aligned}$$

To prove the central limit theorem (CLT), we make use of the following assumption, which is similar to $(H)_m$ ($m = 2$) of [5]. It is used to control remainder terms, when constructing a CLT. It implies that the backward Markov proposal kernels mix very quickly.

(**A1**) 1. The incremental weights all satisfy:

$$\forall 1 \leq n \leq t-1 \quad \delta_\theta = \sup_{x,y}\frac{\overrightarrow{W}_n(x)}{\overrightarrow{W}_n(y)} < \infty \quad \forall t+1 \leq n \leq T \quad \delta_\theta = \sup_{x,y}\frac{\overleftarrow{W}_n(x)}{\overleftarrow{W}_n(y)} < \infty$$

For each $\theta \in \Theta$ there exist $0 < \underline{C}_\theta < \overline{C}_\theta < \infty$ such that for every $x, x' \in \mathbb{R}^{d_x}$, $n \in \{1, \ldots, T\}$, $y_n \in \mathbb{R}^{d_y}$

$$\underline{C}_\theta \leq f_\theta(x'|x) \leq \overline{C}_\theta \quad \underline{C}_\theta \leq \xi_{n,\theta}(x) \leq \overline{C}_\theta \quad \underline{C}_\theta \leq g_\theta(y_n|x) \leq \overline{C}_\theta.$$

In addition, for each $\theta \in \Theta$, there exist $0 < \underline{C}_\theta < \overline{C}_\theta < \infty$ as above, such that for each $x_t, x_{t-1}, x_t \in \mathbb{R}^{d_x}$, $i \in \{1, \ldots, N\}$

$$\underline{C}_\theta \leq q_{t,\theta}(x_t|x_{t-1}, x_{t+1}) \leq \overline{C}_\theta \quad \underline{C}_\theta \leq \overrightarrow{\beta}_{t-1}^i \leq \overline{C}_\theta \quad \underline{C}_\theta \leq \overleftarrow{\beta}_{t+1}^i \leq \overline{C}_\theta.$$

2. For $m = 2$ and some sequence of numbers $\omega_p^{(m)} \in [1, \infty)$ such that for each $p \in \{-1, \ldots, T-t-m\}$ and any $(x, x') \in \mathbb{R}^{2d_x}$ we have

$$M_{T-p,T-p-m}(x, dy) \leq \omega_p^{(m)} M_{T-p,T-p-m}(x', dy)$$

where $M_{p,q} = M_{p-1} \ldots M_q$, $p \geq q$.

Proof of Theorem 2.1. We have that:

$$\mathbb{E}[p_\theta^N(y_{1:T})|\overrightarrow{\mathscr{F}}_{t-1}^N \otimes \overleftarrow{\mathscr{F}}_{t+1}^N] = \overrightarrow{\gamma}_{t-1}^N \otimes \overleftarrow{\gamma}_{t+1}^N(\overrightarrow{W}_{t-1}\overleftarrow{W}_{t+1}^\xi I_{gf})$$

where $\overrightarrow{\mathscr{F}}_{t-1}^N$ and $\overleftarrow{\mathscr{F}}_{t+1}^N$ are the filtrations generated by the forward and backward particle systems up-to time $t-1$ and $t+1$ respectively. We can use the decomposition of [6] to obtain the following formula:

$$\mathbb{E}[p_\theta^N(y_{1:T})|\overrightarrow{\mathscr{F}}_{t-1}^N \otimes \overleftarrow{\mathscr{F}}_{t+1}^N] - p_\theta(y_{1:T}) = \alpha(N) + \beta(N) + R(N)$$

where

$$\begin{aligned}\alpha(N) &= \sum_{q=1}^{t-1}\overrightarrow{\gamma}_q^N(1)[\overrightarrow{\eta}_q^N - \overrightarrow{\Phi}_q(\overrightarrow{\eta}_{q-1}^N)](\overrightarrow{Q}_{q,t-1}[\overrightarrow{W}_{t-1}\overleftarrow{\gamma}_{t+1}(\overleftarrow{W}_{t+1}^\xi I_{gf}(.,\cdot))])\\ \beta(N) &= \sum_{q=0}^{T-t-1}\overleftarrow{\gamma}_{T-q}^N(1)[\overleftarrow{\eta}_{T-q}^N - \overleftarrow{\Phi}_{T-q}(\overleftarrow{\eta}_{T-q-1}^N)](\overleftarrow{Q}_{T-q,t}[W_{t+1}^\xi\overrightarrow{\gamma}_{t-1}(W_{t-1}I_{gf}(\cdot,.))])\\ R(N) &= \sum_{q=1}^{t-1}\overrightarrow{\gamma}_q^N(1)[\overrightarrow{\eta}_q^N - \overrightarrow{\Phi}_q(\overrightarrow{\eta}_{q-1}^N)](\overrightarrow{Q}_{q,t-1}[\overrightarrow{W}_{t-1}[\overleftarrow{\gamma}_{t+1}^N - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}_{t+1}^\xi I_{gf}(.,\cdot))]).\end{aligned}$$

It is straightforward to verify that the expectation of this quantity is exactly zero, which establishes the unbiased property.

By using the Marcinicwiez-Zygmund inequality

$$\mathbb{E}[|\sqrt{N}(p_\theta^N(y_{1:T}) - \mathbb{E}[p_\theta^N(y_{1:T})|\overrightarrow{\mathscr{F}}_{t-1}^N \otimes \overleftarrow{\mathscr{F}}_{t+1}^N])|] \leq \frac{C_\theta}{N^2}\mathbb{E}[|\overrightarrow{\gamma}_{t-1}^N(1)\overleftarrow{\gamma}_{t+2}^N(1)|]$$

for some $C_\theta < +\infty$. For any fixed t, T, $\sup_{N\geq 1}\mathbb{E}[\overrightarrow{\gamma}^N_{t-1}(1)^2]^{1/2} < \infty$ and $\sup_{N\geq 1}\mathbb{E}[\overleftarrow{\gamma}^N_{t+2}(1)^2]^{1/2} < \infty$ (see the proof of Lemma A.1), thus, via Cauchy-Schwarz, we can deduce that (note that $\to_{\mathbb{P}}$ denotes convergence in probability)

$$\sqrt{N}(p^N_\theta(y_{1:T}) - \mathbb{E}[p^N_\theta(y_{1:T})|\overrightarrow{\mathscr{F}}^N_{t-1} \otimes \overleftarrow{\mathscr{F}}^N_{t+1}]) \to_{\mathbb{P}} 0.$$

The weak convergence of $\sqrt{N}\alpha(N)$ and $\sqrt{N}\beta(N)$ can be obtained by the independence of the terms and [6, Proposition 9.4.1]. By Lemma A.1 the remainder $\sqrt{N}R(N)$ converges to zero in probability and we can conclude the result. □

Lemma A.1. *Assume (A1). Then for fixed $T > 2$, $t \in \{3, \ldots, T-2\}$ we have that*

$$\sqrt{N}R(N) = \sqrt{N}\sum_{q=1}^{t-1}\overrightarrow{\gamma}^N_q(1)[\overrightarrow{\eta}^N_q - \overrightarrow{\Phi}_q(\overrightarrow{\eta}^N_{q-1})](\overrightarrow{Q}_{q,t-1}[\overrightarrow{W}_{t-1}[\overleftarrow{\gamma}^N_{t+1} - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}^\xi_{t+1}I_{gf}(.,\cdot))]) \to_{\mathbb{P}} 0.$$

Proof. To shorten the subsequent notations, define:

$$\begin{aligned}\xi^N_{q,t-1}(x) &= \overrightarrow{Q}_{q,t-1}[\overrightarrow{W}_{t-1}[\overleftarrow{\gamma}^N_{t+1} - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}^\xi_{t+1}I_{gf}(.,\cdot)](x)\\ \bar{\xi}^N_{q,t-1} &= \sup_x \overrightarrow{W}_{t-1}(x)\sup_x \overrightarrow{Q}_{q,t-1}(|[\overleftarrow{\gamma}^N_{t+1} - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}^\xi_{t+1}I_{gf}(.,\cdot)|)(x).\end{aligned}$$

It is remarked that for any bounded function φ, $\sup_x \overrightarrow{Q}_{q,t-1}(\varphi)(x) < \infty$ by assumption.

We will now show that $\sqrt{N}R(N)$ will go-to zero in $\mathbb{L}_1$. To that end, we can consider the expectation of each summand in the series for $R(N)$. We have

$$\mathbb{E}\Big[\Big|\overrightarrow{\gamma}^N_q(1)[\overrightarrow{\eta}^N_q - \overrightarrow{\Phi}_q(\overrightarrow{\eta}^N_{q-1})]\Big(\frac{\xi^N_{q,t-1}}{\bar{\xi}^N_{q,t-1}}\Big)\bar{\xi}^N_{q,t-1}\Big|\Big] \leq \mathbb{E}\Big[\Big|\overrightarrow{\gamma}^N_q(1)[\overrightarrow{\eta}^N_q - \overrightarrow{\Phi}_q(\overrightarrow{\eta}^N_{q-1})]\Big(\frac{\xi^N_{q,t-1}}{\bar{\xi}^N_{q,t-1}}\Big)\Big|^2\Big]^{1/2}\mathbb{E}[(\bar{\xi}^N_{q,t-1})^2]^{1/2}$$

where we have used Cauchy-Schwarz. For the first expectation on the R.H.S. one can condition on $\overrightarrow{\mathscr{F}}^N_{q-1} \otimes \overleftarrow{\mathscr{F}}^N_{t+1}$ and apply the Marcinicwiez-Zygmund inequality (noting that $\sup_x |\xi^N_{q,t-1}(x)|/\bar{\xi}^N_{q,t-1}$ is upper-bounded by a finite deterministic constant) to obtain that

$$\mathbb{E}\Big[\Big|\overrightarrow{\gamma}^N_q(1)[\overrightarrow{\eta}^N_q - \overrightarrow{\Phi}_q(\overrightarrow{\eta}^N_{q-1})]\Big(\frac{\xi^N_{q,t-1}}{\bar{\xi}^N_{q,t-1}}\Big)\Big|^2\Big]^{1/2} \leq \frac{C}{\sqrt{N}}\mathbb{E}[\overrightarrow{\gamma}^N_q(1)^2]^{1/2}.$$

Note that for each q, $\mathbb{E}[\overrightarrow{\gamma}^N_q(1)^2]^{1/2} < \infty$ (e.g. [5, Corollary 5.2], or by using the upper-bound on the $\overrightarrow{W}_n$).

Now, we move onto the expression $\mathbb{E}[(\bar{\xi}^N_{q,t-1})^2]^{1/2}$. From the definition of $\bar{\xi}^N_{q,t-1}$, we have that

$$\mathbb{E}[(\bar{\xi}^N_{q,t-1})^2]^{1/2} = \sup_x \overrightarrow{W}_{t-1}(x)\sup_x \overrightarrow{Q}_{q,t-1}(1)(x)\mathbb{E}[\overrightarrow{\overline{Q}}_{q,t-1}(|[\overleftarrow{\gamma}^N_{t+1} - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}^\xi_t I_{gf}(.,\cdot)|)^2]^{1/2}$$

where $\overrightarrow{\overline{Q}}_{q,t-1}(\cdot)(x) := \sup_x \overrightarrow{Q}_{q,t-1}(\cdot)(x)/\sup_x \overrightarrow{Q}_{q,t-1}(\cdot)(x)$. Application of Jensen's inequality and Fubini leads to

$$\mathbb{E}[(\bar{\xi}^N_{q,t-1})^2]^{1/2} \leq \sup_x \overrightarrow{W}_{t-1}(x)\sup_x \overrightarrow{Q}_{q,t-1}(1)(x)\overrightarrow{\overline{Q}}_{q,t-1}\Big(\mathbb{E}[|[\overleftarrow{\gamma}^N_{t+1} - \overleftarrow{\gamma}_{t+1,\theta}](\overleftarrow{W}^\xi_t I_{gf}(.,\cdot))|^2]\Big)^{1/2}.$$

Then by [5, Theorem 5.1, Corollary 5.2] (it is remarked that the corollary of that paper can be adapted to deal when $\overleftarrow{\gamma}_{t+1}$ integrates a bounded function), it follows for N large enough relative to $T-t$ (we will take N to infinity and $T-t$ is fixed) there exist some finite constant $C(T,t)$ that depends upon T, t but not q or x_{t-1} such that

$$\mathbb{E}[(\bar{\xi}^N_{q,t-1})^2]^{1/2} \leq \sup_x G_{t-1}(x)\sup_x Q_{q,t-1}(1)(x)\frac{C(T,t)}{\sqrt{N}}.$$

Hence we have that:

$$\sqrt{N}\mathbb{E}[|R(N)|] \leq \frac{C(T,t)}{\sqrt{N}}$$

where $C(T,t)$ is some finite constant that may grow with T. We thus conclude as $T < \infty$. □

References

[1] Andrieu, C., Doucet, A. & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. Ser. B*, **72**, 269–342.

[2] Bresler, Y. (1986). Two-filter formula for discrete-time non-linear Bayesian smoothing. *Intl. J. Control*, **43**, 629–641.

[3] Briers, M. & Doucet, A. & Maskell, S. (2010). Smoothing algorithms for state-space models. *Ann. Inst. Statist. Math.*, **62**, 61–89.

[4] Cappé, O., Ryden, T, & Moulines, É. (2005). *Inference in Hidden Markov Models*. Springer: New York.

[5] Cérou, F., Del Moral, P. & Guyader, A. (2011). A non-asymptotic variance theorem for un-normalized Feynman-Kac particle models. *Ann. Inst. Henri Poincare*, **47**, 629–649.

[6] Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

[7] Del Moral, P., Doucet, A. & Singh, S. S. (2010). A backward interpretation of Feynman-Kac formulae. *M2AN*, **44**, 947–975.

[8] Douc, P., Garivier, A., Moulines, E. & Olsson, J. (2011). On the forward filtering backward smoothing particle approximations of the smoothing distribution in general state space models. *Ann. Appl. Probab.*, **21**, 2109–2145.

[9] Doucet, A., Godsill, S. & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.*, **10**, 197–208.

[10] Fearnhead, P., Wyncoll, D. & Tawn, J. (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika*, **97**, 2, 447–464.

[11] Persing, A. (2012). *Eighteen Month PhD. Report*. Imperial College London.

[12] Whiteley, N., Kantas, N. & Jasra, A. (2012). Linear variance bounds for particle approximations of time homogeneous Feynman-Kac formulae. *Stoch. Proc. Appl.*, **122**, 1840–1865.