Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models

Jimmy Olsson, Olivier Cappé, Randal Douc, Eric Moulines

SEQUENTIAL MONTE CARLO SMOOTHING WITH APPLICATION TO PARAMETER ESTIMATION IN NON-LINEAR STATE SPACE MODELS

JIMMY OLSSON, OLIVIER CAPPÉ, RANDAL DOUC, AND ÉRIC MOULINES

Abstract. This paper concerns the use of Sequential Monte Carlo methods (SMC) for smoothing in general state space models. A well known problem when applying the standard SMC technique in the smoothing mode is that the resampling mechanism introduces degeneracy of the approximation in the path-space. However, when performing maximum likelihood estimation via the EM algorithm, all involved functionals will be of additive form for a large subclass of models. To cope with the problem in this case, a modification, relying on forgetting properties of the filtering dynamics, of the standard method is proposed. In this setting, the quality of the produced estimates is investigated both theoretically and through simulations.

1. Introduction

This paper is devoted to the study of sequential Monte Carlo methods for smoothing in non-linear state space models. We consider a bivariate process $\{(X_k, Y_k); k \geq 0\}$, where $\{X_k\}$ is a Markov chain on a state space X. Conditional on $\{X_k\}$, $\{Y_k\}$ is a sequence of independent random variables on the space Y such that the distribution of Y_k is governed by X_k only. In this framework, $\{X_k\}$ is not observed, and measurements on the system have to be made through the observed process $\{Y_k\}$. Put, for $i \leq j$, $\boldsymbol{Y}_{i:j} \triangleq (Y_i, \ldots, Y_j)$; similar vector notation will be used for other quantities. In the following, assume that we are given a set $\boldsymbol{y}_{0:n}$ of observations of $\boldsymbol{Y}_{0:n}$. Operating on state space models, a constantly recurring problem is to compute expectation values of form $\mathbb{E}[t_n(\boldsymbol{X}_{0:n})|\boldsymbol{Y}_{0:n}]$, where t_n is a real-valued, measurable function. In this report we focus on the case of t_n being an *additive functional* t_n given by

$$t_n(\boldsymbol{x}_{0:n}) = \sum_{k=0}^{n-1} s_k(\boldsymbol{x}_{k:k+1}) \ , \tag{1.1}$$

where $\{s_k; k \geq 0\}$ is a sequence of measurable functions (which may depend on the observed values $\boldsymbol{y}_{0:n}$).

Key words and phrases. EM algorithm, exponential family, particle filters, sequential Monte Carlo methods, state space models, stochastic volatility model.

This work was supported by a grant from the Swedish Foundation for Strategic Research, a French government oversea student grant, and a grant from the French National Agency for Research (ANR-2005 ADAP'MC project).

As an example of when smoothing of such *additive functionals* is important, consider the case of maximum likelihood estimation via the EM algorithm. Let p_θ be a generic symbol for densities, where the index θ is a parameter vector. Then, the complete data log-likelihood function is given by

$$\log p_\theta(\boldsymbol{x}_{0:n}, \boldsymbol{y}_{0:n}) = \sum_{k=0}^{n-1} \log p_\theta(x_{k+1}|x_k) + \sum_{k=0}^{n} \log p_\theta(y_k|x_k) + \log p_\theta(x_0) \ ,$$

yielding the following intermediate quantity of the EM algorithm:

$$\begin{aligned}\mathcal{Q}(\theta;\theta') &\triangleq \mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{X}_{0:n}, \boldsymbol{Y}_{0:n}) \middle| \boldsymbol{Y}_{0:n}\right] \\ &= \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n-1} \log p_\theta(X_{k+1}|X_k) \middle| \boldsymbol{Y}_{0:n}\right] + \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n} \log p_\theta(Y_k|X_k) \middle| \boldsymbol{Y}_{0:n}\right] \\ &\quad + \mathbb{E}_{\theta'}\left[\log p_\theta(X_0) \middle| \boldsymbol{Y}_{0:n}\right] \ ,\end{aligned}$$

where $\mathbb{E}_{\theta'}$ denotes expectation under θ'.

Having an initial estimate θ' of the parameter vector at hand, an improved estimate is obtained by means of computation and maximization of $\mathcal{Q}(\theta;\theta')$ with respect to θ. This procedure is recursively repeated in order to obtain convergence to a stationary point $\theta_\star$ of the log-likelihood function $\ell_n(\theta) \triangleq p_\theta(\boldsymbol{y}_{0:n})$.

The computation of smoothed sum functionals of the form above will also be the crucial matter when considering direct maximum likelihood estimation via the score function $\log \nabla_\theta \ell_n(\theta)$. Under appropriate differentiability assumptions, the *Fisher identity* (see Louis, 1982) states that

$$\begin{aligned}\nabla_\theta \ell_n(\theta) &= \mathbb{E}_\theta\left[\sum_{k=0}^{n-1} \nabla_\theta \log p_\theta(X_{k+1}|X_k) \middle| \boldsymbol{Y}_{0:n}\right] + \mathbb{E}_\theta\left[\sum_{k=0}^{n} \nabla_\theta \log p_\theta(Y_k|X_k) \middle| \boldsymbol{Y}_{0:n}\right] \\ &\quad + \mathbb{E}_\theta\left[\nabla_\theta \log p_\theta(X_0) \middle| \boldsymbol{Y}_{0:n}\right] \ ,\end{aligned} \tag{1.2}$$

yielding an expression which is closely related to that of the intermediate quantity of EM.

By applying Bayes' formula it is (see, e.g., Cappé et al., 2005) straightforward to derive recursive formulas not only for the smoothing and filter densities, that is, $p_\theta(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})$ and $p_\theta(x_k|\boldsymbol{y}_{0:k})$, respectively, but also for expectations of the form $\mathbb{E}_\theta\left[t_k(\boldsymbol{X}_{0:k})\middle|\boldsymbol{Y}_{0:k}\right]$ discussed above. However, tractable closed form solutions are available only if the state space X is finite or the model is linear and Gaussian. *Particle filtering methods* (often alternatively termed *sequential Monte Carlo methods*) constitute a class of algorithms that are well suited for providing approximative solutions of the smoothing and filtering recursions. In recent years, sequential Monte Carlo methods have been put in use, sometimes very successfully in many different fields (see Doucet et al. (2001b) and Ristic et al. (2004) and the references therein). For shorter introductions we refer to the articles by Doucet et al. (2001a) and Künsch

(2001). In particle filter algorithms, a set of weighted simulations, the *particles*, is recursively updated using importance sampling and resampling techniques. An estimate of the joint smoothing distribution is given by the empirical measure associated with the particle trajectories.

A well known problem when applying sequential Monte Carlo methods to sample the joint smoothing distribution is that the resampling mechanism of the particle filter introduces degeneracy of the approximation. To cope with the situation, Doucet et al. (2004) suggest a nonlinear, non-Gaussian counterpart of the forward-filtering backward-smoothing procedure used for linear Gaussian state-space models. After a standard forward particle filtering pass, a new set of weighted trajectories, based on particles originating from the first run, are generated through a series of multinomial resampling steps in the time-reversed direction. In this way the degeneracy of the particle paths is avoided, providing a robustification of the particle smoothing approximation. Since the method requires an additional backward simulation sweep, this is obtained at the cost of computational work. Thus, the algorithm is well fitted to sample from the joint smoothing distribution. Nevertheless, it appears (perhaps unnecessarily) complex to approximate the smoothing functionals of the form (1.1).

In this contribution, we study a SMC technique to smooth additive functionals initially advocated in Kitagawa and Sato (2001). The method exploits the *forgetting properties* of the conditional hidden chain and is not affected by the degeneracy of the particle trajectories. Compared to Doucet et al. (2004), it is computationally efficient. Furthermore, we perform, under suitable regularity assumptions of the latent chain, a theoretical analysis of the behavior of the obtained estimates. It turns out that the L^p error is roughly $O(N^{-1/2} n \log n)$ and the bias $O(N^{-1} n \log n)$, where N denotes the number of particles and n the number of observations.

For a comparison, applying the results of Del Moral and Doucet (2003, Theorem 4) to a functional of type (1.1) provides an $O(N^{-1/2} n^2)$ bound of the L^p error for the standard trajectory-based particle filtering smoother. Finally, we apply, for a noisily observed autoregressive model and the stochastic volatility model proposed by Hull and White (1987), the technique to the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1991). In this setting we investigate, both theoretically and through simulations, the quality of the produced estimates.

The paper is organized as follows. In Section 2 we describe the state space framework and introduce basic notation and assumptions. Some important concepts, such as the smoothing recursion, are also recalled. In Section 3, particle filtering is briefly described, and a method, based on forgetting ideas, for approximating additive functionals is presented. In the theoretical part, Section 4, a number of results describing the convergence of the approximations to the exact quantities are derived. The first results are valid for any fixed sequence of observed values, but are easily, under additional assumptions on the model, extended to randomly varying

observations. As a preparation for this section, we recall the uniform forgetting property of the conditional hidden Markov chain. In Section 5 we use the proposed method for estimating parameters in a noisily observed autoregressive model and a stochastic volatility model via the EM algorithm. This is done in the light of the theory developed in Section 4.

2. Basic notation and concepts

2.1. **Model description.** Let the hidden process $X \triangleq \{X_k; k \geq 0\}$ be a homogeneous discrete time Markov chain taking values in a state space $(\mathsf{X}, \mathcal{X})$. Let $(Q_\theta, \theta \in \Theta \subseteq \mathbb{R}^d)$ and ν denote the Markov transition kernel and the initial distribution of X, respectively. The family $\{Q_\theta(x, \cdot),\ x \in \mathsf{X},\ \theta \in \Theta\}$, is assumed to be dominated by the probability measure μ on X and we denote by $q_\theta(x, \cdot)$, the corresponding Radon-Nikodým derivatives. The observations $\{Y_k; k \geq 0\}$ are random variables taking values in a measurable space $(\mathsf{Y}, \mathcal{Y})$. These variables are conditionally independent given the sequence $\{X_k; k \geq 0\}$ of hidden states, and the conditional distribution of Y_k depends on X_k only. Furthermore, there exists, for all $x \in \mathsf{X}$ and $\theta \in \Theta$, a density function $y \mapsto g(x, y; \theta)$ and a measure λ on $(\mathsf{Y}, \mathcal{Y})$ such that, for $k \geq 0$,

$$\mathbb{P}_\theta \left(Y_k \in A | X_k = x \right) = \int_A g(x, y; \theta)\, \lambda(\mathrm{d}y) \,, \quad \text{for all } A \in \mathcal{Y}.$$

We denote by $\mathbb{P}_{\theta,\nu}$ the joint distribution of X and Y, which is indexed by both the parameter θ and the initial distribution ν; however, many conditional probabilities and expectations in this paper (like the one of the previous display) do not depend on the initial distribution, and in those cases ν is expunged from the notation. In addition, denote by $\mathcal{G}_k$ the σ-algebra generated by the observed process from time 0 to k.

2.2. **The smoothing recursion.** The *joint smoothing distribution* $\phi_{\nu,0:n|n}$ is the probability defined, for $A \in \mathcal{X}^{\otimes(n+1)}$, by

$$\phi_{\nu,0:n|n}[\boldsymbol{y}_{0:n}](A; \theta) \triangleq \mathbb{P}_{\theta,\nu} \left((X_0, \ldots, X_n) \in A | \, \boldsymbol{Y}_{0:n} = \boldsymbol{y}_{0:n} \right) \,.$$

Under the assumptions above, the joint smoothing distribution has a density (for which we will use the same symbol) with respect to $\mu^{\otimes(n+1)}$ satisfying the recursion

$$\begin{aligned} &\phi_{\nu,0:k+1|k+1}[\boldsymbol{y}_{0:k+1}](\boldsymbol{x}_{0:k+1}; \theta) \\ &\quad = \frac{\mathrm{L}_{\nu,k}(\theta; \boldsymbol{y}_{0:k})}{\mathrm{L}_{\nu,k+1}(\theta; \boldsymbol{y}_{0:k+1})} q_\theta(x_k, x_{k+1}) g_{k+1}(x_{k+1}, y_{k+1}; \theta) \phi_{\nu,0:k|k}[\boldsymbol{y}_{0:k}](\boldsymbol{x}_{0:k}; \theta) \,, \end{aligned} \tag{2.1}$$

where $\mathrm{L}_{\nu,k}(\theta; \boldsymbol{y}_{0:k})$ is the *likelihood function* given by

$$\mathrm{L}_{\nu,k}(\theta; \boldsymbol{y}_{0:k}) \triangleq \int_{\mathsf{X}} \cdots \int_{\mathsf{X}} g_0(x_0, y_0; \theta) \nu(x_0) \prod_{l=1}^{k} q_\theta(x_{l-1}, x_l) g_l(x_l, y_l; \theta) \mu^{\otimes(k+1)}(\mathrm{d}\boldsymbol{x}_{0:k}) \,.$$

For notational conciseness, we will in the following omit the explicit dependence on the observations $\boldsymbol{y}_{0:k}$ of the quantities defined above from the notation, and we replace $\phi_{\nu,0:k|k}(\cdot\,;\theta)$, $g_k(\cdot\,;\theta)$, and $\mathrm{L}_{\nu,k}(\theta)$ for $\phi_{\nu,0:k|k}[\boldsymbol{y}_{0:k}](\cdot\,;\theta)$, $g(\cdot, y_k;\theta)$, and $\mathrm{L}_{\nu,k}(\theta;\boldsymbol{y}_{0:k})$, respectively. By integrating (2.1) with respect to the first k coordinates, a similar iterative formula for the filtering distributions, that is, the marginals $\phi_{\nu,k}(\cdot\,;\theta) \triangleq \mathbb{P}_{\theta,\nu}(X_k \in \cdot\,|\mathcal{G}_k)$, is obtained. A recursive formula for the expected value of additive functionals $\{t_n\}$ of the form given in (1.1) follows as a direct consequence of (2.1): For a fixed functional such that all expectations $\mathbb{E}_{\theta,\nu}[|t_n(\boldsymbol{X}_{0:n})||\mathcal{G}_n]$ are finite, define a family of signed measures $\{\tau_n; n \geq 0\}$ on $(\mathsf{X}, \mathcal{X})$ by

$$\tau_n(f) \triangleq \int_{\mathsf{X}} \cdots \int_{\mathsf{X}} f(x_n) t_n(\boldsymbol{x}_{0:n})\, \phi_{\nu,0:n|n}(\mathrm{d}\boldsymbol{x}_{0:n};\theta)\ ,$$

for $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X})$, where for any integer m, $\mathcal{B}_{\mathrm{b}}(\mathsf{X}^m)$ denotes the Banach space of bounded measurable functions on X^m furnished with the sup norm $\|f\|_{\mathsf{X}^m,\infty} \triangleq \sup_{\boldsymbol{x}\in\mathsf{X}^m} |f(\boldsymbol{x})|$. Plugging this formula into the recursion (2.1) and rewriting, again using Bayes' formula, the ratio of the likelihood functions yields, for $k \geq 0$,

$$\tau_{k+1}(f) = \frac{1}{\phi_{\nu,k}(Q_\theta g_{k+1};\theta)} \int_{\mathsf{X}}\int_{\mathsf{X}} f(x_{k+1}) Q_\theta(x_k, dx_{k+1}) g_{k+1}(x_{k+1};\theta) \times [\tau_k(\mathrm{d}x_k) + \phi_{\nu,k}(\mathrm{d}x_k;\theta) s_k(\boldsymbol{x}_{k:k+1})]\ , \quad (2.2)$$

where s_k is given by (1.1). The procedure is initialized by

$$\tau_0(f) = \frac{1}{\nu g_0} \int_{\mathsf{X}} f(x_0) t_0(x_0) g_0(x_0;\theta) \nu(\mathrm{d}x_0)\ ,$$

and at each time index k, the expected value $\mathbb{E}_{\theta,\nu}[t_k(\boldsymbol{X}_{0:k})|\mathcal{G}_k]$ may be obtained by evaluating $\tau_k(\mathsf{X})$. Since the previous formula contains the filter distribution $\phi_{\nu,k}$, the usage of (2.2) requires that the filtering equations are computed in a parallel manner.

However, as mentioned in the introduction, the simplicity of the smoothing recursion above is treacherous, since we cannot achieve closed form solutions of the likelihood function $\mathrm{L}_{\nu,k}(\theta)$ and the starting distribution $\phi_{\nu,0}(x_0;\theta)$ if the model contains nonlinear/non-Gaussian model elements.

3. Particle Approximation of Additive Functionals

Particle filtering, in its most basic form, consists of approximating the exact smoothing relations by propagating particle trajectories in the state space of the hidden chain. Given a fixed sequence of observations, this is done by following the scheme below. In order to keep the notation simple, we fix the model parameters and omit θ from the notation.

At time 0, a number N of particles $\{\xi_0^{N,i}(0); 1 \leq i \leq N\}$ are drawn from a common probability measure ς such that $\nu \ll \varsigma$. These *initial particles* are assigned the *importance weights* $\omega_0^{N,i} \triangleq W_0[\xi_0^{N,i}(0)]$, $i = 1, \ldots, N$, where, for

$x \in \mathsf{X}$, $W_0(x) \triangleq g_0(x)\,\mathrm{d}\nu/\mathrm{d}\varsigma(x)$, providing $\sum_{i=1}^N \omega_0^{N,i} f[\xi_0^{N,i}(0)]/\sum_{i=1}^N \omega_0^{N,i}$ as an importance sampling estimate of $\phi_{\nu,0}f$, for $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X})$. We define the σ-algebra $\mathcal{F}_0^N \triangleq \sigma[\xi_0^{N,1}(0), \ldots, \xi_0^{N,N}(0)]$. Henceforth, the particle paths $\boldsymbol{\xi}_m^{N,i} \triangleq [\xi_m^{N,i}(0), \ldots, \xi_m^{N,i}(m)]$, $1 \le i \le N$, are recursively updated according to the following procedure. Assume that we at time k have at hand a set $\{(\boldsymbol{\xi}_k^{N,i}, \omega_k^{N,i}); 1 \le i \le N\}$ of weighted particles approximating $\phi_{\nu,0:k|k}$, in the sense that

$$\frac{1}{\Omega_k^N} \sum_{i=1}^N \omega_k^{N,i} f(\boldsymbol{\xi}_k^{N,i}) \;,$$

with $\Omega_k^N \triangleq \sum_{i=1}^N \omega_k^{N,i}$ and $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+1})$, is an estimate of the expectation $\phi_{\nu,0:k|k}f$. Then, an updated weighted sample $\{(\boldsymbol{\xi}_{k+1}^{N,i}, \omega_{k+1}^{N,i}); 1 \le i \le N\}$, approximating the distribution $\phi_{\nu,0:k+1|k+1}$, is obtained by, firstly, simulating $\boldsymbol{\xi}_{k+1}^{N,i} \sim R_k^{\mathrm{p}}(\boldsymbol{\xi}_k^{N,i}, \cdot)$, where the path-wise proposal kernel R_k^{p} is of type

$$R_k^{\mathrm{p}}(\boldsymbol{x}_{0:k}, f) = \int_{\mathsf{X}} R_k(x_k, \mathrm{d}x_{k+1}) f(\boldsymbol{x}_{0:k}, x_{k+1}) \;,$$

with $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+2})$ and each R_k being a Markov transition kernel. The new particles are simulated independently of each other, and the special form of R_k^{p} implies that past particle trajectories are kept unchanged throughout this *mutation step*. A popular choice is to set $R_k \equiv Q$, yielding the so-called *bootstrap filter*; more sophisticated techniques involve proposals depending on the new observation y_{k+1} (see Example 5.2). Secondly, when the new observation is available, the importance weights are updated according to the formula $\omega_{k+1}^{N,i} = \omega_k^{N,i} W_{k+1}[\boldsymbol{\xi}_{k+1}^{N,i}(k : k+1)]$ where, for $(x, x') \in \mathsf{X}^2$, $W_k(x, x') \triangleq g_k(x')\,\mathrm{d}Q(x, \cdot)/\mathrm{d}R_{k-1}(x, \cdot)(x')$. Furthermore, we define $\mathcal{F}_{k+1}^N \triangleq \mathcal{F}_k^N \vee \sigma(\boldsymbol{\xi}_{k+1}^{N,1}, \ldots, \boldsymbol{\xi}_{k+1}^{N,N})$. Now, for $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+2})$, the self-normalized estimate

$$\phi_{\nu,0:k+1|k+1}^N f \triangleq \frac{1}{\Omega_{k+1}^N} \sum_{j=1}^N \omega_{k+1}^{N,j} f(\boldsymbol{\xi}_{k+1}^{N,j})$$

provides an approximation of the expectation $\phi_{\nu,0:k+1|k+1}f$.

As it is well established, the previous scheme fails because the distribution of the importance weights becomes more and more skewed as k increases. To prevent degeneracy, a *selection mechanism* should be introduced. In its simpler form, this mechanism amounts to resample, when needed, the propagated particles by drawing, conditionally independently, indices $I_k^{N,1}, \ldots, I_k^{N,N}$ in the set $\{1, \ldots, N\}$ multinomially with respect to the normalized weights, that is,

$$\mathbb{P}\left(I_k^{N,i} = j \,\middle|\, \mathcal{F}_k^N \vee \mathcal{G}_k\right) = \frac{\omega_k^{N,j}}{\Omega_k^N} \;, \quad j \in \{1, \ldots, N\} \;.$$

Now, a new particle cloud $\{\widehat{\boldsymbol{\xi}}_k^{N,i}; 1 \le i \le N\}$ is formed by setting $\widehat{\boldsymbol{\xi}}_k^{N,j} = \boldsymbol{\xi}_k^{N,I_k^{N,j}}$. After the resampling procedure, the weights are all reset $\omega_k^{N,i} = 1/N$, yielding

another estimate

$$\widehat{\phi}^N_{\nu,0:k|k} f \triangleq \frac{1}{N} \sum_{i=1}^{N} f\big(\widehat{\boldsymbol{\xi}}_k^{N,i}\big)$$

of $\phi_{\nu,0:k|k} f$. Note that the resampling mechanism might modify the whole trajectory of a certain particle, implying that in general, for $m \leq n$, $\xi_n^{N,i}(m) \neq \xi_{n+1}^{N,i}(m)$.

The multinomial resampling method is not the only conceivable way to carry out the selection step; see Section 5. The collection Doucet et al. (2001b) contains several suggestions for how to improve the algorithm above in general, as well as a great variety of examples of adjustments of the technique to specific applications.

Having a particle filtering device at our disposal, an approximation of $\gamma_n \triangleq \tau_n(\mathsf{X}) = \mathbb{E}_\nu[t_n(\boldsymbol{X}_{0:n})|\mathcal{G}_n]$ is obtained by propagating a system of particle trajectories and associated weights $\{(\boldsymbol{\xi}_k^{N,j}, \omega_k^{N,j}); 1 \leq j \leq N, 0 \leq k \leq n\}$, and using the estimators

$$\gamma_n^N = \frac{1}{\Omega_n^N} \sum_{j=1}^{N} \omega_n^{N,j} t_n(\boldsymbol{\xi}_n^{N,j}) \quad \text{or} \quad \widehat{\gamma}_n^N = \frac{1}{N} \sum_{j=1}^{N} t_n\big(\widehat{\boldsymbol{\xi}}_n^{N,j}\big) \,. \tag{3.1}$$

When the functional $\{t_n\}$ has the form given in (1.1), it is straightforward to verify that storing the whole particle trajectories is indeed not required to evaluate (3.1): upon defining $t_k^{N,i} \triangleq t_k(\boldsymbol{\xi}_k^{N,i})$ we have, for $k \geq 1$,

$$t_{k+1}^{N,i} = \begin{cases} t_k^{N,i} + s_k\left[\boldsymbol{\xi}_{k+1}^{N,i}(k:k+1)\right] \,, & \text{if no resampling;} \\ t_k^{N,I_{k+1}^i} + s_k\left[\widehat{\boldsymbol{\xi}}_{k+1}^{N,i}(k:k+1)\right] \,, & \text{if resampling occurs.} \end{cases} \tag{3.2}$$

The recursion is initialized by $t_1^{N,i} = t_1(\boldsymbol{\xi}_1^{N,i})$. In accordance with (3.1), γ_n^N is obtained as $\sum_{i=1}^N \omega_n^{N,i} t_n^{N,i} / \Omega_n^N$. Hence, for each particle $\boldsymbol{\xi}_k^{N,i}$ we only need to store its current position $\xi_k^{N,i}(k)$, weight $\omega_k^{N,i}$ and associated functional value $t_k^{N,i}$. Thus, the method necessitates only minor adaptations once the particle filter has been implemented.

As illustrated in Fig. 3.1, as n increases, the path trajectories system collapse, and the estimators (3.1) are not reliable for sensible N values (see Doucet et al. (2001b), Kitagawa and Sato (2001) and Andrieu and Doucet (2003) for a discussion).

To cope with this drawback we suggest the following method, advocated first in Kitagawa and Sato (2001). By the forgetting property of the time-reversed conditional hidden chain (Theorem 4.2), we expect that, for a large enough integer $\Delta_n \leq n - k$,

$$\mathbb{E}_\nu\left[s_k(\boldsymbol{X}_{k:k+1}) \middle| \mathcal{G}_n\right] \approx \mathbb{E}_\nu\left[s_k(\boldsymbol{X}_{k:k+1}) \middle| \mathcal{G}_{k+\Delta_n}\right] \,, \tag{3.3}$$

yielding,

$$\gamma_n = \mathbb{E}_\nu\left[\sum_{k=0}^{n-1} s_k(\boldsymbol{X}_{k:k+1}) \middle| \mathcal{G}_n\right] \approx \sum_{k=0}^{n-1} \mathbb{E}_\nu\left[s_k(\boldsymbol{X}_{k:k+1}) \middle| \mathcal{G}_{(k+\Delta_n)\wedge n}\right] .$$

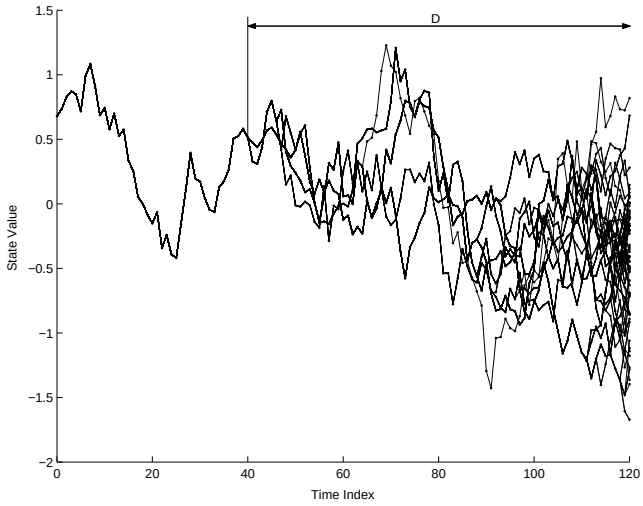

FIGURE 3.1. Typical particle trajectories for $N = 50$; see Section 5 for details regarding model and algorithm.

The relation above suggests that waiting for all the trajectories to collapse—as (3.2) implies—is not an optimal simulation principle. Instead, when the particle population N is sufficiently large, so that (3.3) is valid for a lag Δ_n which may be far smaller than typical collapsing time. This yields the two approximations

$$\gamma_n^{N,\Delta_n} \triangleq \sum_{k=0}^{n-1} \sum_{j=1}^{N} \frac{\omega_{k(\Delta_n)}^{N,j}}{\Omega_{k(\Delta_n)}^{N}} s_k \left[\boldsymbol{\xi}_{k(\Delta_n)}^{N,j}(k:k+1) \right] , \tag{3.4}$$

$$\widehat{\gamma}_n^{N,\Delta_n} \triangleq \frac{1}{N} \sum_{k=0}^{n-1} \sum_{j=1}^{N} s_k \left[\widehat{\boldsymbol{\xi}}_{k(\Delta_n)}^{N,j}(k:k+1) \right] , \tag{3.5}$$

of γ_n, where $k(\Delta_n) \triangleq (k + \Delta_n) \wedge n$. Although somewhat more involved than the standard approximation (3.1), the lag-based approximation above may be updated recursively by maintaining a cache of the recent history of the particles as well as the cumulated contribution of terms that will not get updated anymore.

Thus, apart from increased storage requirements, computing the lag-based approximation $\widehat{\gamma}_n^{N,\Delta_n}$ is clearly not, from a computational point of view, more demanding than computing $\widehat{\gamma}_n^{N}$.

4. Theoretical evaluation of the fixed-lag technique

To accomplish the robustification above, we need to specify the lag Δ_n and how this lag should depend on n. This is done by examining the quality of the estimates produced by the algorithm in terms of bias and L^p error. Of particular interest is how these errors are affected by the lag and whether it makes their dependence on n and N more favorable in comparison with the standard trajectory-based approach.

The validity of (3.3), is based on the assumption that the conditional hidden chains—in the forward as well as the backward directions—have *forgetting properties*

that is, the distributions of two versions of each chain starting at different initial distributions approach each other as time increases. This property depends on the following uniform ergodicity conditions on the model, which imply that forgetting occurs at a *geometrical* rate.

Assumption 4.1.

(i) $\sigma_- \triangleq \inf_{\theta\in\Theta} \inf_{x,x'\in\mathsf{X}} q_\theta(x,x') > 0$, $\sigma_+ \triangleq \sup_{\theta\in\Theta} \sup_{x,x'\in\mathsf{X}} q_\theta(x,x') < \infty$.

(ii) For all $y \in \mathsf{Y}$, $\sup_{\theta\in\Theta} \|g(\cdot, y\,;\theta)\|_{\mathsf{X},\infty} < \infty$ *and* $\inf_{\theta\in\Theta} \int_{\mathsf{X}} g(x,y;\theta)\,\mu(\mathrm{d}x) > 0$.

We now define the Markov transition kernels that generate the conditional hidden chains. Introduce, for $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{k+2})$ and $\boldsymbol{x}_{0:k} \in \mathsf{X}^{k+1}$, the un-normalized path-wise transition kernel

$$L_k(\boldsymbol{x}_{0:k}, f;\theta) \triangleq \int_{\mathsf{X}} f(\boldsymbol{x}_{0:k+1}) g_{k+1}(x_{k+1};\theta)\, Q_\theta(x_k, \mathrm{d}x_{k+1}) \,.$$

Assumption 4.1 makes this integral well defined for all $k \geq 0$. It is easily seen that for all $k \leq m$, the function $L_k \cdots L_m(\boldsymbol{x}_{0:k}, \mathsf{X}^{m+2};\theta)$ depends only on x_k. Thus, a version of this function comprising only the last component is well defined, and we write $L_k \cdots L_m(x_k, \mathsf{X}^{m+2};\theta)$ in this case. For $k > m$, we set $L_k \cdots L_m \equiv \mathrm{Id}$. Using this notation and given $n \geq 0$, the *forward smoothing kernels* given by, for $k \geq 0$, $x_k \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\mathrm{F}_{k|n}(x_k, A;\theta) \triangleq \mathbb{P}_\theta\left(X_{k+1} \in A \middle| X_k = x_k, \mathcal{G}_n\right) \,,$$

can, for indices $0 \leq k < n$, be written as

$$\mathrm{F}_{k|n}(x_k, A;\theta) = \int_A \frac{g_{k+1}(x_{k+1};\theta) L_{k+1} \cdots L_{n-1}(x_{k+1}, \mathsf{X}^{n+1};\theta)\, Q_\theta(x_k, \mathrm{d}x_{k+1})}{L_k \cdots L_{n-1}(x_k, \mathsf{X}^{n+1};\theta)} \,. \quad (4.1)$$

For $k \geq n$ we simply have $\mathrm{F}_{k|n}(x_k, A;\theta) = Q_\theta(x_k, A)$.

Analogously, for the time-reversed conditional hidden chain we consider the *backward smoothing kernels* defined by, for a given $n \geq 0$,

$$\mathrm{B}_{\nu,k|n}(x_{k+1}, A;\theta) \triangleq \mathbb{P}_{\theta,\nu}\left(X_k \in A \middle| X_{k+1} = x_{k+1}, \mathcal{G}_n\right) \,, \quad (4.2)$$

where $k \geq 0$, $x_{k+1} \in \mathsf{X}$ and $A \in \mathcal{X}$. Note that since the probability (4.2) depends on the initial distribution of the latent chain, ν is included in the notation of the backward kernel. A straightforward application of Bayes' formula shows that $\mathrm{B}_{\nu,k|n}$ can, for indices $k \leq n$, be expressed as

$$\mathrm{B}_{\nu,k|n}(x_{k+1}, A;\theta) = \frac{\int_A q_\theta(x_k, x_{k+1})\, \phi_{\nu,k}(\mathrm{d}x_k;\theta)}{\int_{\mathsf{X}} q_\theta(x_k, x_{k+1})\, \phi_{\nu,k}(\mathrm{d}x_k;\theta)} \,. \quad (4.3)$$

In addition, for $k > n$, that is, for time indices outside the observed region, we have

$$\mathrm{B}_{\nu,k|n}(x_{k+1}, A;\theta) = \frac{\int_A \int_{\mathsf{X}} q_\theta(x_k, x_{k+1}) q_\theta^{k-n}(x_n, x_k)\, \phi_{\nu,n}(\mathrm{d}x_n;\theta)\, \mu(\mathrm{d}x_k)}{\int_{\mathsf{X}} q_\theta^{k-n+1}(x_n, x_{k+1})\, \phi_{\nu,n}(\mathrm{d}x_n;\theta)} \,, \quad (4.4)$$

where, for $m \geq 1$, q_θ^m denotes the density of the m-step kernel Q_θ^m. For any two probability measures η_1 and η_2 we define the *total variation distance* $\|\eta_1 - \eta_2\|_{\mathrm{TV}} = \sup_A |\eta_1(A) - \eta_2(A)|$, and for measurable functions f we recall the identity $\sup_{f:\|f\|_\infty \leq 1} |\eta_1 f - \eta_2 f| = 2\|\eta_1 - \eta_2\|_{\mathrm{TV}}$. Part (i) of Assumption 4.1 implies, for all $x \in \mathsf{X}$, $\theta \in \Theta$ and $A \in \mathcal{X}$, the bound $Q_\theta(x, A) \geq \sigma_- \mu(A)$, saying that the hidden chain allows X as a *1-small set* (see, e.g., Meyn and Tweedie, 1993). Under this condition, for any two initial distributions ν_1, ν_2 and any parameter $\theta \in \Theta$, $\|\nu_1 Q_\theta^n - \nu_2 Q_\theta^n\|_{\mathrm{TV}} \to 0$ geometrically fast as n goes to infinity. See Lindvall (1992), Sections III.9–11.

Assumption 4.1 has an important impact on the conditional hidden chains in both directions. Firstly, we consider the forward smoothing kernel. Plugging the uniform upper and lower bounds of q_θ into the formula (4.1) provides straightforwardly, for $x_k \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\mathrm{F}_{k|n}(x_k, A; \theta) \geq (1 - \rho)\kappa_k(A; \boldsymbol{y}_{k+1:n}, \theta) \;, \tag{4.5}$$

where

$$\kappa_k(A; \boldsymbol{y}_{k+1:n}, \theta) \triangleq \frac{\int_A g_{k+1}(x_{k+1}; \theta) L_{k+1} \cdots L_{n-1}(x_{k+1}, \mathsf{X}^{n+1}; \theta)\, \mu(\mathrm{d}x_{k+1})}{\int_{\mathsf{X}} g_{k+1}(x_{k+1}; \theta) L_{k+1} \cdots L_{n-1}(x_{k+1}, \mathsf{X}^{n+1}; \theta)} \;,$$

and $\rho \triangleq 1 - \sigma_-/\sigma_+$. For indices after the last observation, that is, $k \geq n$, it is easily checked that the bound (4.5) holds true with $\kappa_k(A; \boldsymbol{y}_{k+1:n}, \theta) \triangleq \mu(A)$. Thus, we conclude that X is a 1-small set also for the conditional hidden chain in the forward direction, with minorization constant $1 - \rho$. The backward direction works analogously; in fact, using (4.3) and (4.4),

$$\mathrm{B}_{\nu,k|n}(x_{k+1}, A; \theta) \geq (1 - \rho)\widehat{\kappa}_{\nu,k}(A; \boldsymbol{y}_{0:k\wedge n}, \theta) \;, \tag{4.6}$$

where, for $k \leq n$, $\widehat{\kappa}_{\nu,k}(A; \boldsymbol{y}_{0:k}, \theta) \triangleq \phi_{\nu,k}(A; \theta)$ and, for $k > n$,

$$\widehat{\kappa}_{\nu,k}(A; \boldsymbol{y}_{0:n}, \theta) \triangleq \frac{\int_A \int_{\mathsf{X}} q_\theta^{k-n}(x_n, x_k)\, \phi_{\nu,n}(\mathrm{d}x_n; \theta)\, \mu(\mathrm{d}x_k)}{\int_{\mathsf{X}^2} q_\theta^{k-n}(x_n, x_k)\, \phi_{\nu,n}(\mathrm{d}x_n; \theta)\, \mu(\mathrm{d}x_k)} \;.$$

The following theorem (see Del Moral, 2004, p.143) shows that the forward and backward kernels are geometrically ergodic.

Theorem 4.2. *Assume 4.1 and let $\theta \in \Theta$. Then, for all $k \geq m \geq 0$, all probability measures ν_1 and ν_2 on $\mathcal{X}$ and all $\boldsymbol{y}_{0:n}$,*

$$\left\|\nu_1 \mathrm{F}_{m|n} \cdots \mathrm{F}_{k|n}(\cdot\,; \theta) - \nu_2 \mathrm{F}_{m|n} \cdots \mathrm{F}_{k|n}(\cdot\,; \theta)\right\|_{\mathrm{TV}} \leq \rho^{k-m+1} \;,$$
$$\left\|\nu_1 \mathrm{B}_{\nu,k|n} \cdots \mathrm{B}_{\nu,m|n}(\cdot\,; \theta) - \nu_2 \mathrm{B}_{\nu,k|n} \cdots \mathrm{B}_{\nu,m|n}(\cdot\,; \theta)\right\|_{\mathrm{TV}} \leq \rho^{k-m+1} \;.$$

Assumption 4.1 typically requires that X is a compact set, and one has at present time failed to significantly weaken this, rather strong, restriction (see Chigansky and Lipster (2004), Doucet and Tadić (2005), Del Moral (2004), and Cappé et al. (2005) for some attempts in this direction).

4.1. **Main results.** For any stochastic variable Z, sub-σ-algebra $\mathcal{H}$ of $\mathcal{F}$ and integer $p \geq 1$, we define the conditional L^p norm $\|Z\|_{p|\mathcal{H}} \triangleq \mathbb{E}^{1/p}[|Z|^p|\mathcal{H}]$. In most cases, we will let $\mathcal{H}$ be an element of the filtration generated by the observed process; studying the particle filter mechanism under the norm $\|\cdot\|_{p|\mathcal{G}_n}$ implies the assumption that all randomness of the system is concentrated to the evolution of the particle swarm only, while the observations are fixed. For notational brevity, we assume in this section that resampling is applied at every iteration; for the sake of simplicity, we have only considered the case where multinomial resampling is used (see Chopin (2004) and Douc and Moulines (2005) for analysis of the case of residual resampling).

Assumption 4.3. *For all $k \geq 1$, $\|W_k\|_{\mathsf{X}^2,\infty} < \infty$; in addition, $\|W_0\|_{\mathsf{X},\infty} < \infty$.*

Remark 4.4. In case of the bootstrap particle filter, for which $R_k \equiv Q$, Assumption 4.3 is implied by Assumption 4.1. The same is true for the so-called *optimal kernel* used in Example 5.2.

Theorem 4.5. *Under assumptions 4.1, 4.3, for $n \geq 0$, the following holds true for all $\Delta_n \geq 0$ and $N \geq 1$.*

(i) For all $p \geq 2$,

$$\left\|\widehat{\gamma}_n^{N,\Delta_n} - \gamma_n\right\|_{p|\mathcal{G}_n} \leq 2\rho^{\Delta_n} \sum_{k=0}^{n-\Delta_n} \|s_k\|_{\mathsf{X}^2,\infty} + \frac{B_p}{\sqrt{N}(1-\rho)} \sum_{k=0}^{n-1} \|s_k\|_{\mathsf{X}^2,\infty} \left(\frac{1}{\sigma_-} \sum_{m=1}^{(k+\Delta_n)\wedge n} \frac{\|W_m\|_{\mathsf{X}^2,\infty} \rho^{0\vee(k-m)}}{\mu g_m} + \frac{\|W_0\|_{\mathsf{X},\infty}}{\nu g_0} + 1 \right) ,$$

(ii)

$$\left|\mathbb{E}\left[\widehat{\gamma}_n^{N,\Delta_n} \middle| \mathcal{G}_n\right] - \gamma_n\right| \leq 2\rho^{\Delta_n} \sum_{k=0}^{n-\Delta_n} \|s_k\|_{\mathsf{X}^2,\infty} + \frac{B}{N(1-\rho)^2} \sum_{k=0}^{n-1} \|s_k\|_{\mathsf{X}^2,\infty} \left(\frac{1}{\sigma_-^2} \sum_{m=1}^{(k+\Delta_n)\wedge n} \frac{\|W_m\|_{\mathsf{X}^2,\infty}^2 \rho^{0\vee(k-m)}}{(\mu g_m)^2} + \frac{\|W_0\|_{\mathsf{X},\infty}^2}{(\nu g_0)^2} + 1 \right) .$$

Here B_p and B are universal constants such that B_p depends on p only.

For the purpose of illustrating these bounds, assume that all $\|s_k\|_{\mathsf{X}^2,\infty}$ and all fractions $\|W_k\|_{\mathsf{X}^2,\infty}/\mu g_k$ are uniformly bounded in k. We then draw the conclusion that if the lag is increased with n at a rate of $\log n$ then the error is then dominated by the variability due to the particle filter—the second term of 4.5(i)—which is of order $O(N^{-1/2} n \log n)$. In contrast, setting $\Delta_n = n$, that is, using the direct full-path approximation, would result in a stochastic error of order $O(N^{-1/2} n^2)$.

4.2. **Extension to randomly varying observations.** As mentioned, all results presented above concern smoothing distribution approximations produced by the particle filter algorithm *given a fixed sequence of observations*. However, when studying asymptotic properties—consistency and asymptotic normality—of the approximative particle filter MLE (see Olsson and Rydén, 2005) it is a necessity to extend these results to the case of a randomly varying sequence of observations.

For the bounds presented in Theorem 4.5, the conditioning on $\mathcal{G}_n$ can be removed by introducing additional model assumptions. Denote by $\mathring{Q}$, $\mathring{g}$ and $\mathring{\nu}$ the kernel, measurement and initial densities, respectively, of the state space model generating the observations used by the particle filter (we stress that $\mathring{Q}$ and $\mathring{g}$ are not assumed to belong to the parametric families $\{(Q_\theta, g_\theta); \theta \in \Theta\}$). In addition, we let $\mathring{\mathbb{P}}$ be the law of the bivariate Markov associated to $(\mathring{Q},\mathring{g},\mathring{\nu})$, $\mathring{\mathbb{E}}$ the corresponding expectation, and $\|\cdot\|_{p,\circ}$ the L^p norm under $\mathring{\mathbb{P}}$. Using these observed values as input, the evolution of the particle cloud follows the usual dynamics $(Q_\theta,g_\theta,\nu,\, \theta \in \Theta)$. The objective is to *a priori*, that is, before the observations $\boldsymbol{y}_{0:n}$ are available, form an idea of how well the particle filter will approximate γ_n.

Assumption 4.6. *Let t_n be given by (1.1). For $p \geq 2$ and $\ell \geq 1$ there exists a constant $a_{p,\ell}(t_n) \in \mathbb{R}$ such that*

$$\sup_{\substack{0\leq k\leq n\\ 0\leq i\leq n-1}} \mathring{\mathbb{E}}\left[\frac{\|W_k\|_{\mathsf{X},\infty}^p \, \|s_i\|_{\mathsf{X}^2,\infty}^\ell}{(\mu g_k)^p}\right] \vee \mathring{\mathbb{E}}\left[\|s_i\|_{\mathsf{X}^{n+1},\infty}^\ell\right] \leq a_{p,\ell}(t_n) \; .$$

Proposition 4.7. *Assume 4.1 and 4.3. Then, the following holds true for all $N \geq 1$.*

(i) If Assumption 4.6 is satisfied for $\ell = p \geq 2$, then

$$\left\|\widehat{\gamma}_n^{N,\Delta_n} - \gamma_n\right\|_{p,\circ} \leq 2b_p(t_n)\rho^{\Delta_n}(n-\Delta_n+1) + \frac{B_p b_p(t_n)}{\sqrt{N}(1-\rho)}\left\{\frac{\Delta_n n + \Delta_n}{\sigma_-} + n\left[\frac{1}{\sigma_-(1-\rho)} + \frac{1}{\inf_{x\in\mathsf{X}}\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right]\right\} ,$$

where $b_p(t_n) \triangleq [a_{p,p}(t_n)]^{1/p}$;

(ii) if Assumption 4.6 is satisfied for $p = 2$, $\ell = 1$, then

$$\left|\mathring{\mathbb{E}}\left[\widehat{\gamma}_n^{N,\Delta_n} - \gamma_n\right]\right| \leq 2a_{2,1}(t_n)\rho^{\Delta_n}(n-\Delta_n+1) + \frac{B_p a_{2,1}(t_n)}{N(1-\rho)^2}\left\{\frac{\Delta_n n + \Delta_n}{\sigma_-^2} + n\left[\frac{1}{\sigma_-^2(1-\rho)} + \frac{1}{\inf_{x\in\mathsf{X}}\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right]\right\} .$$

The proof of this result is given in Section 7.2.

Remark 4.8. In the case of a compact state space X, Assumption 4.6 implies only limited additional restrictions on the state space model. In fact, for a large class of models, Assumption 4.6 follows as a direct consequence of Assumption 4.1.

5. Applications to maximum likelihood estimation

We now return to the computation of maximum likelihood estimator. In the following we consider models for which the set of complete data log-likelihood functions, parameterized by the parameter vector, constitute an *exponential family*; that is, for all $\theta \in \Theta$ and $n \geq 0$,

$$p_\theta(\boldsymbol{x}_{0:n}, \boldsymbol{y}_{0:n}) = \exp\left[\langle \boldsymbol{\psi}(\theta), \boldsymbol{S}_n(\boldsymbol{x}_{0:n})\rangle - c(\theta)\right] h(\boldsymbol{x}_{0:n}) \; . \tag{5.1}$$

Here $\boldsymbol{\psi}$ and the sufficient statistics $\boldsymbol{S}_n$ are $\mathbb{R}^{d_s}$-valued functions on Θ and X^{n+1} , respectively, and c is a real-valued function on θ and h is a real-valued non-negative function on X^{n+1}. By $\langle \cdot\,, \cdot \rangle$ we denote the scalar product. All these functions may depend on the observed values $\boldsymbol{y}_{0:n}$, even though this is expunged from the notation.

If the complete data log-likelihood function is of the particular form (5.1) and the expectation $\phi_{\nu,0:n|n}(\boldsymbol{S}_n; \theta)$ is finite for all $\theta \in \Theta$, the intermediate quantity of EM can be written as

$$\mathcal{Q}(\theta; \theta') = \left\langle \boldsymbol{\psi}(\theta), \phi_{\nu,0:n|n}(\boldsymbol{S}_n; \theta') \right\rangle - c(\theta) + \phi_{\nu,0:n|n}\left(\log h; \theta'\right) \; .$$

Since the last term does not depend on θ, it has no effect on the updated parameter value. Thus, we will in the following exclude this term from the expression of $\mathcal{Q}$. We will throughout this section assume that maximization of $\langle \boldsymbol{\psi}(\theta), \boldsymbol{s}\rangle - c(\theta)$ with respect to the parameter is feasible for all possible values $\boldsymbol{s}$ of the sufficient statistics. Note finally that, as mentioned in the introduction, a typical element $S_{n,m}(\boldsymbol{x}_{0:n})$ of the vector $\boldsymbol{S}_n(\boldsymbol{x}_{0:n})$ is an additive functional $S_{n,m}(\boldsymbol{x}_{0:n}) = \sum_{k=0}^{n-1} s_{n,m}^{(k)}(\boldsymbol{x}_{k:k+1})$ so that $\phi_{\nu,0:n|n}(\boldsymbol{S}_n; \theta')$ can be estimated using either (3.4) or (3.5). Denoting by $\widehat{\boldsymbol{S}}_n$ such estimator, we may approximate the intermediate quantity by

$$\widehat{\mathcal{Q}}^N(\theta; \theta') = \left\langle \boldsymbol{\psi}(\theta), \widehat{\boldsymbol{S}}_n \right\rangle - c(\theta) \; . \tag{5.2}$$

In the next step—referred to as the M-step—$\widehat{\mathcal{Q}}^N(\theta; \theta')$ is maximized with respect to θ, providing a new parameter estimate. This procedure is repeated recursively given an initial guess $\widehat{\theta}_0$. Proceeding in this way, we obtain a Monte-Carlo version of the EM algorithm (see Tanner, 1993; Fort and Moulines, 2003). We summarize the algorithm below.

For $i = 1, 2, \ldots$

Simulation step: Run, using N_i particles, the particle filter n steps under the parameter $\widehat{\theta}^{i-1}$, providing an approximation $\widehat{\boldsymbol{S}}_n^i$ of the smoothed quantity $\phi_{\nu,0:n|n}(\boldsymbol{S}_n; \widehat{\theta}^{i-1})$. Use this to obtain $\widehat{\mathcal{Q}}^{N_i}(\theta; \widehat{\theta}^{i-1})$ via (5.2).

M-step: Next, choose the new estimate $\widehat{\theta}^i$ to be the (or any, if several exist) value of $\theta \in \Theta$ which maximizes $\widehat{\mathcal{Q}}^{N_i}(\theta; \widehat{\theta}^{i-1})$.

As an illustration, we consider the problem of inference in a noisily observed AR(1) model and the stochastic volatility (SV) model. For the first model, it is possible to carry out exact computation of the MLE, which is interesting from a comparative point of view. Since the noisily observed AR(1) model belongs to the class of linear/Gaussian models, for which ergodicity of the conditional hidden chain is established, in the sense that (3.3) holds true, the application of the fixed-lag technique is well-founded in this case. However, similar theoretical results are not available for the SV model, and here we instead plead *empirical evidence* that this model also exhibits such forgetting properties. Thus, both the following examples will be analyzed in the light of the theory developed in Section 4.

Example 5.1 (SMCEM for noisily observed AR(1) model)**.** We consider the state space model

$$X_{k+1} = aX_k + \sigma_w W_{k+1} \ ,$$
$$Y_k = X_k + \sigma_v V_k \ ,$$

the variables $\{W_k; k \geq 1\}$ and $\{V_k; k \geq 0\}$ being independent standard normal distributed variables. We let the initial distribution ν be a diffuse prior; in particular this means that the initial filter distribution $\phi_{\nu,0}$ is $\mathcal{N}(y_0, \sigma_v)$. Throughout this example we consider an observation data set of length $n = 10000$, produced by simulation under the parameters $a^{\mathrm{sim}} = 0.98$, $\sigma_w^{\mathrm{sim}} = 0.2$ and $\sigma_v^{\mathrm{sim}} = 1$. This noisily observed AR(1) model dovetails into the framework of exponential families, with

$$\boldsymbol{\psi}(\theta) = \left(\frac{1}{2\sigma_w^2}, -\frac{a}{\sigma_w^2}, \frac{a^2}{2\sigma_w^2}, \frac{1}{2\sigma_v^2} \right) \ ,$$

and components of the $\mathbb{R}^4$-valued function $\boldsymbol{x}_{0:n} \mapsto \boldsymbol{S}_n(\boldsymbol{x}_{0:n})$ given by

$$S_{n,1}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=1}^{n-1} x_k^2 \ , \qquad S_{n,2}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=0}^{n-1} x_k x_{k+1} \ ,$$
$$S_{n,3}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=0}^{n} x_k^2 \ , \qquad S_{n,4}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=0}^{n} (y_k - x_k)^2 \ .$$

Furthermore, up to terms not depending on parameters,

$$c(\theta) = \frac{n}{2} \log \sigma_w^2 + \frac{n+1}{2} \log \sigma_v^2 \ .$$

In this setting, one step of the MCEM algorithm is carried out in the following way. Having produced an estimate $\widehat{\theta}^{i-1}$ of the parameters $\theta = (a, \sigma_w^2, \sigma_v^2)$ at the previous iteration, we compute an approximation $\widehat{\boldsymbol{S}}_n = (\widehat{S}_{n,1}, \widehat{S}_{n,2}, \widehat{S}_{n,3}, \widehat{S}_{n,4})$ of $\phi_{\nu,0:n|n}(\boldsymbol{S}_n; \widehat{\theta}^{i-1})$ using the particle filter and update the parameters according to

$$\widehat{a}^i = \frac{\widehat{S}_{n,2}}{\widehat{S}_{n,1}} \ , \quad \left(\widehat{\sigma}_w^i\right)^2 = \frac{1}{n} \left(\widehat{S}_{n,3} - \widehat{a}^i \widehat{S}_{n,2} \right) \ , \quad \left(\widehat{\sigma}_v^i\right)^2 = \frac{\widehat{S}_{n,4}}{n+1} \ . \tag{5.3}$$

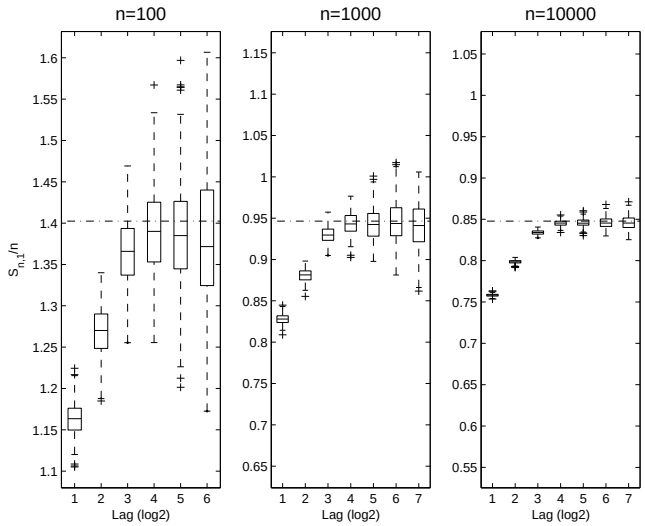

FIGURE 5.1. Boxplots of estimates of $\phi_{\nu,0:n|n}S_{n,1}/n$, produced with the fixed-lag technique, for the noisily observed AR(1) model in Example 5.1.

We simulated, for each $n = 100, 1000, 10000$ observations, 1000 sequential Monte Carlo estimates of $\phi_{\nu,0:n|n}S_1$ using the fixed-lag smoothing technique for the parameter values $a = 0.8$, $\sigma_w = 0.5$ and $\sigma_v = 2$. Here the standard bootstrap particle filter with systematic resampling was used, implying that, for all $k \geq 0$, $R_k \equiv Q$. The dotted lines indicate the exact expected values, obtained by means of disturbance smoothing. To study the bias-variance trade-off—discussed in detail in the previous section—of the method, we used six different lags for each n and a constant particle population size $N = 1000$. The result is displayed in Figure 5.1, from which it is evident that the bias is controlled for a size of the lag that increases approximately logarithmically with n: In particular, from the plot we deduce that optimal outcome is gained when lags of size 2^4, 2^4, and 2^5 are used for n being 100, 1000, and 10000, respectively.

When the lag is sufficiently large for ignoring the term of the bias which is deduced from forgetting arguments—being roughly of magnitude $n\rho^{\Delta_n}$—, increasing the lag further exclusively leads to an increase of variance as well as bias of the estimates; compare the two last boxes of each plot. This is completely in accordance with the theoretical results of Section 3. Note that the scale on the y-axis is the same for the three panels although the y-axis has been shifted in each panel due to the fact that the value of the normalized smoothed statistic evolves as the number of observations increases.

In Figure 5.2, we again report the cases $n = 100, 1000, 10000$ observations, and compare the basic approximation strategy (3.2) with the one based on fixed-lag smoothing with suitable lags. Guided by the plots of Figure 5.1 and the theory developed in the previous section, we choose the lags 2^4, 2^4, and 2^5, respectively.

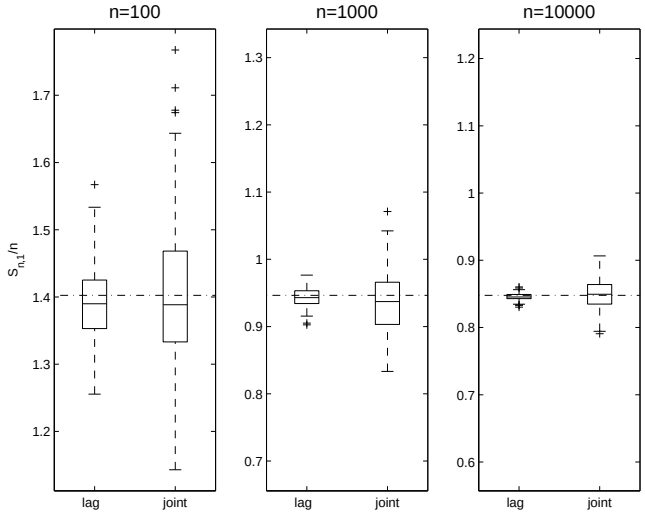

FIGURE 5.2. Boxplots of estimates of $\phi_{\nu,0:n|n} S_{n,1}/n$, produced by means of both the fixed-lag technique and standard trajectory-based smoothing, for the noisily observed AR(1) model in Example 5.1. Each box is based on 200 estimates, and the size of the particle population was $N = 1000$ for all cases.

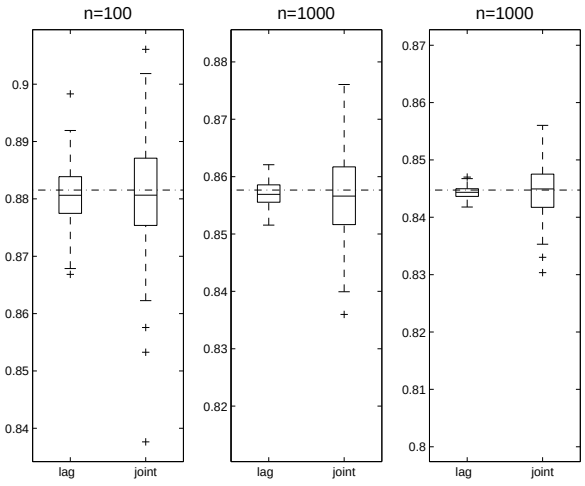

FIGURE 5.3. Boxplots of Monte Carlo EM updates of a based on the estimates of Figure 5.2.

The number of particles was set to 1000 for all n. It is obvious that fixed-lag smoothing drastically reduces the variance without significantly raising the bias. As in the previous figure, dotted lines indicate exact values. As expected, the bias of the two techniques increases with n, since the number of particles is held constant.

Finally, we display in Figure 5.3 the resulting parameter estimates obtained by plugging the additive functional estimates of Figure 5.3 into the EM updating formulas (5.3), resulting in one iteration step of the Monte Carlo EM algorithm. Exact updates are indicated by dotted lines. As for the additive functional estimates, it

is clear that introducing the lag implies an obvious improvement. As the updating formulas (5.3) involve n, the dependence on n of the variance of the parameter estimates is more involved here than in the functional case.

Example 5.2 (SMCEM for the stochastic volatility (SV) model)**.** In the discrete time case, the canonical version of the SV model (Hull and White, 1987; Jacquier et al., 1994) is given by the two relations

$$X_{k+1} = \alpha X_k + \sigma \epsilon_{k+1} \ , \quad Y_k = \beta \exp\left(\frac{X_k}{2}\right) \varepsilon_k \ ,$$

where $\{\epsilon_k; k \geq 1\}$ and $\{\varepsilon_k; k \geq 0\}$ are independent standard normal distributed variables.

To use the SV model in practice, we need to estimate the parameters $\theta = (\beta, \alpha, \sigma)$. Throughout this example we will use a sequence of data obtained by simulation under the parameters $\beta^{\text{sim}} = 0.63$, $\alpha^{\text{sim}} = 0.975$ and $\sigma^{\text{sim}} = 0.16$. These parameters are consistent with empirical estimates estimates for daily equity return series and often used in simulation studies. In conformity with Example 5.1, we assume that the latent chain is initialized by an improper diffuse prior. The SV model is within the scope of exponential families, with

$$\boldsymbol{\psi}(\theta) = \left(-\frac{\alpha^2}{2\sigma^2}, -\frac{1}{2\sigma^2}, \frac{\alpha}{\sigma^2}, -\frac{1}{2\beta^2}\right) \ ,$$

and the components of the $\mathbb{R}^4$-valued function $\boldsymbol{x}_{0:n} \mapsto \boldsymbol{S}_n(\boldsymbol{x}_{0:n})$ given by

$$S_{n,1}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=0}^{n-1} x_k^2 \ , \qquad S_{n,2}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=1}^{n} x_k^2 \ ,$$

$$S_{n,3}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=1}^{n} x_k x_{k-1} \ , \qquad S_{n,4}(\boldsymbol{x}_{0:n}) \triangleq \sum_{k=0}^{n} y_k \exp(-x_k) \ .$$

In addition, up to terms not depending on parameters,

$$c(\theta) = \frac{n+1}{2} \log \beta^2 + \frac{n+1}{2} \log \sigma^2.$$

Let $\widehat{\boldsymbol{S}}_n = (\widehat{S}_{n,1}, \widehat{S}_{n,2}, \widehat{S}_{n,3}, \widehat{S}_{n,4})$ be a particle approximation of $\phi_{\nu,0:n|n}(\boldsymbol{S}_n; \widehat{\theta}^{i-1})$. To apply the Monte Carlo EM algorithm to the SV model is not more involved than for the autoregressive model in Example 5.1. In fact, the updating formulas appear to be completely analogous:

$$\widehat{\alpha}^i = \frac{\widehat{S}_{n,3}}{\widehat{S}_{n,1}} \ , \quad \left(\widehat{\sigma}^i\right)^2 = \frac{1}{n}\left(\widehat{S}_{n,2} - \widehat{\alpha}^i \widehat{S}_{n,3}\right) \ , \quad \left(\widehat{\beta}^i\right)^2 = \frac{\widehat{S}_{n,4}}{n+1} \ .$$

In this example, it is possible and more efficient to use a proposal kernel T_k which does not reduce to the prior kernel Q. Indeed, the conditional density of X_{k+1} given *both* X_k and Y_{k+1}—so-called "optimal" proposal density following the terminology of Liu and Chen (1995)—, although known only up to a normalization constant,

is log-concave. As suggested by Pitt and Shephard (1999), it is then possible to find the mode of the optimal proposal density, for each current particle position $\{\xi_k^{N,i}(k); 1 \leq i \leq N\}$, using a fast-converging numerical search and to mimic the optimal proposal by simulating the new particle position from a t-distribution with fixed degrees of freedom but mean and variance adjusted to the determined mode and its Hessian. We adopt this approach here. Note that in this case, the proposal kernel T_k does depend on the new observation Y_{k+1}.

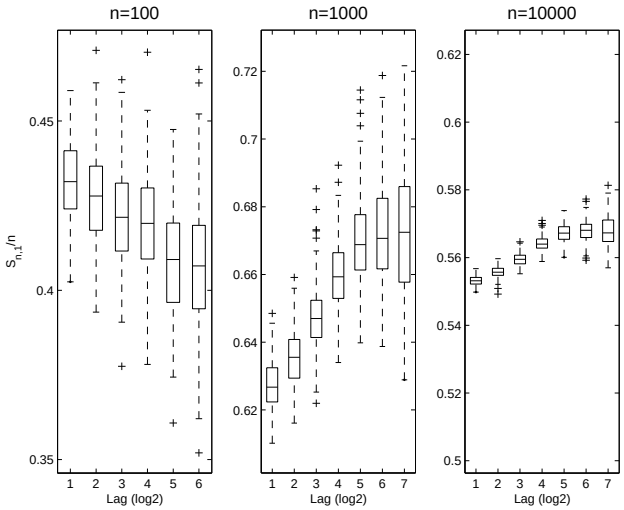

FIGURE 5.4. Boxplots of estimates of $\phi_{\nu,0:n|n}S_{n,1}/n$, produced with the fixed-lag technique, for the SV model in Example 5.2. Each box is based on 200 estimates, and the size of the particle population was set to $N = 1000$ in all cases.

Using the described proposal technique, we repeat the numerical investigations of Example 5.1. The resulting approximation of $\phi_{\nu,0:n|n}S_{n,1}$, displayed in Figure 5.4, behaves similarly. Here again, we observe that moderate values of the lag Δ are sufficient to suppress the bias, supporting our (to the best of our knowledge, yet unproved) conjecture that forgetting also occurs in the stochastic volatility model.

We finally compare the SMCEM parameter estimates obtained with the fixed-lag approximation and the standard trajectory-based approximation on a simulated dataset of length $n = 5000$. Note that for the SMCEM procedure to converges to the MLE, it is necessary to augment the number of simulations that are performed as we progress through the EM iterations. We follow the recommendation of Fort and Moulines (2003) and start by running 150 iterations of the Monte Carlo EM procedure with the number of particles set to $N = 100$. For the subsequent 100 iterations, the number of particles increases at a quadratic rate, with a final value (for the 250th Monte Carlo EM iteration) equal to $N = 1600$. The cumulative number of simulations performed during the 250 SMCEM iterations is equal to 75000 (times the length of the observation sequence) which is quite moderate for a Monte Carlo

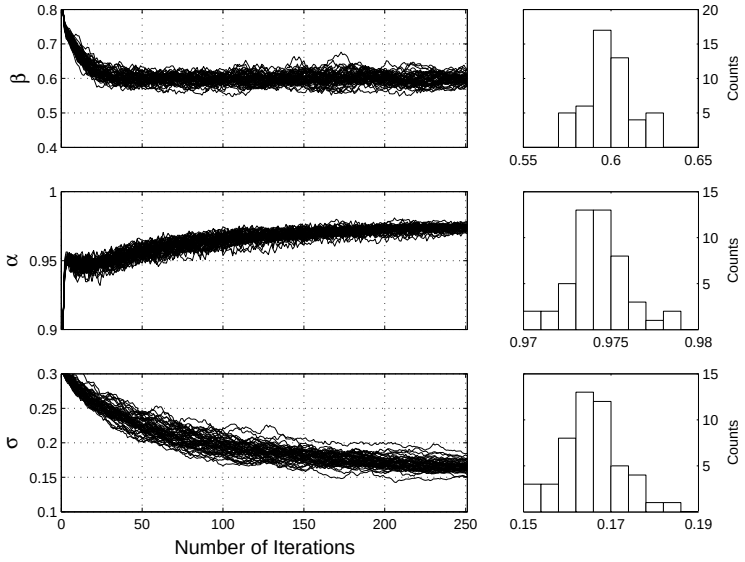

FIGURE 5.5. SMCEM parameter estimates of β, α and σ from $n = 5000$ observations using the standard trajectory-based smoothing approximation. Each plot overlays 50 realizations of the particle simulations; the histograms pertain to the final (250th) SMCEM iteration.

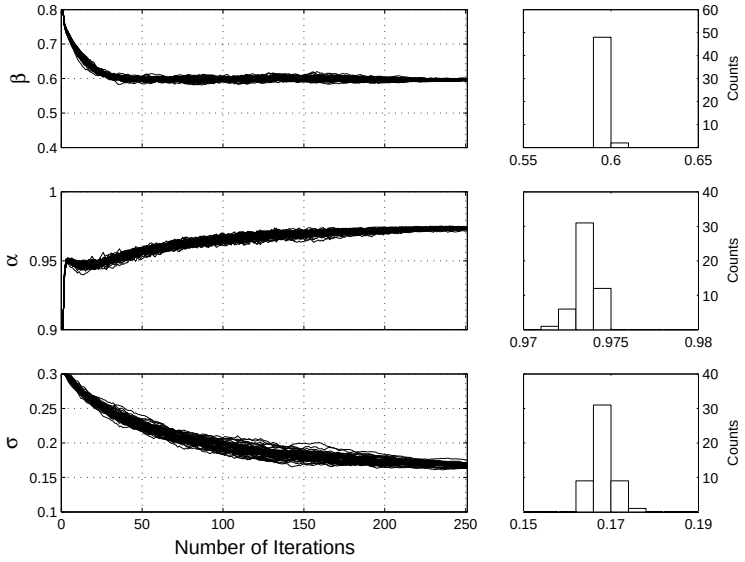

FIGURE 5.6. SMCEM parameter estimates of β, α and σ from $n = 5000$ observations using the fixed-lag smoothing approximation with $\Delta = 40$. Each plot overlays 50 realizations of the particle simulations; the histograms pertain to the final (250th) SMCEM iteration.

based optimization method. In Figures 5.5 and 5.6, we display the superimposed trajectories of parameter estimates for 50 realizations of the particles, together with histograms of the final estimates (at iteration 250) when using, respectively, the

trajectory-based approximation, in Figure 5.5, and the fixed-lag approximation with $\Delta = 40$, in Figure 5.6.

Smoothing algorithm	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\sigma}$
Trajectory-based,	0.5991	0.9742	0.1659
with 75000 total simulations	std. 0.0136	std. 0.0019	std. 0.0070
Trajectory-based,	0.5990	0.9739	0.1666
with 750000 total simulations	std. 0.0045	std. 0.0011	std. 0.0043
Fixed-lag,	0.5962	0.9735	0.1682
with 75000 total simulations	std. 0.0019	std. 0.0006	std. 0.0024

TABLE 1. Mean and standard deviation of SMCEM parameter estimates at the 250th iteration (estimated from 50 independent runs).

Not surprisingly, the fact that the particle simulations are iterated for several successive values of the parameter estimates only amplifies the differences observed so far. With the fixed-lag approximation, the standard deviation of the final SMCEM parameter estimate is divided by a factor of 7 for β, and of 3 for α and σ, which is quite impressive in the context of Monte Carlo methods : to achieve the same accuracy with the trajectory-based approximation, one would need about ten times more particles to compensate for the higher simulation variance. Table 1, shows that the fixed-lag approximation (third row) indeed remains more reliable than the trajectory-based approximation, even when the latter is computed from ten times more particles (second row). Note that for the trajectory-based approximation, multiplying the number of particles by ten does not reduce the standard deviation of the estimates as much as expected from the asymptotic theory. This is certainly due to the moderate number of particles used in the baseline setting, as we start from $N = 100$ particles during the first SMCEM iterations and terminate with $N = 1600$.

6. SUMMARY

We have discussed a modification of the standard trajectory-based sequential Monte Carlo method for smoothing in state space models, which is based on observations made by Kitagawa and Sato (2001). In addition, the mechanisms and gains of the method have been investigated by means of an exhaustive theoretical analysis of the resulting particle estimates. This is the main contribution of the paper. Since the technique is based on forgetting properties of the conditional hidden chain, this analysis had to be performed under suitable regularity conditions on the latent dynamics. An examination of the classical bias-variance-tradeoff of the procedure lead to the key observation that the lag should be increased logarithmically with n. As by-products we obtained a time-uniform $O(N^{-1})$ bound of the particle filtering

bias (follows by Proposition 7.1(ii) below, applied with $i = n$), and an extension to randomly varying observations under fairly weak additional model assumptions.

Finally, we applied, with the theory developed in Section 4 as a guideline, the fixed-lag technique within the framework of Monte Carlo EM for exponential families. In fact, the size of the optimal lag provided by the simulation study, that is, the lag yielding both a minimum bias and a minimum variance, was precisely of order $\log n$. Increasing the lag further only increased the variance. Compared to the standard trajectory-based smoothing, which corresponds to a lag of size n, the proposed fixed-lag smoothing approach dramatically reduced the Monte Carlo variance of the parameter estimates.

7. Proofs

7.1. **Proof of of Theorem 4.5.** The proof of Theorem 4.5 comprises partly the geometric ergodicity of the time-reversed conditional hidden chain (Theorem 4.2), partly the next proposition.

Proposition 7.1. *Assume 4.1, 4.3 and let $0 \leq i \leq n$. Consider a bounded and measurable function f_i, possibly depending on $\boldsymbol{Y}_{0:n}$, of the form $f_i(\boldsymbol{x}_{0:n}) = f_i(\boldsymbol{x}_{i:n})$. Then the following holds true for all $N \geq 1$.*

(i) For all $p \geq 2$,

$$\left\| \widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i \right\|_{p|\mathcal{G}_n} \leq \frac{B_p \|f_i\|_{\mathsf{X}^{n+1},\infty}}{\sqrt{N}(1-\rho)} \left[\frac{1}{\sigma_-} \sum_{k=1}^{n} \frac{\|W_k\|_{\mathsf{X}^2,\infty} \rho^{0 \vee (i-k)}}{\mu g_k} + \frac{\|W_0\|_{\mathsf{X},\infty}}{\nu g_0} + 1 \right] ;$$

(ii)

$$\left| \mathbb{E}\left[\widehat{\phi}^N_{\nu,0:n|n} f_i \middle| \mathcal{G}_n \right] - \phi_{\nu,0:n|n} f_i \right| \leq \frac{B \|f_i\|_{\mathsf{X}^{n+1},\infty}}{N(1-\rho)^2} \left[\frac{1}{\sigma_-^2} \sum_{k=1}^{n} \frac{\|W_k\|^2_{\mathsf{X}^2,\infty} \rho^{0 \vee (i-k)}}{(\mu g_k)^2} + \frac{\|W_0\|^2_{\mathsf{X},\infty}}{(\nu g_0)^2} + 1 \right] .$$

Here B_p and B are universal constants such that B_p depends on p only.

To prove Proposition 7.1 we need some preparatory lemmas and definitions. In accordance with the scheme presented in Section 3, it is valid that, for $A \in \mathcal{X}^{\otimes(k+1)}$,

$$\begin{aligned} &\mathbb{P}\left(\boldsymbol{\xi}^{N,i}_k \in A \middle| \mathcal{F}^N_{k-1} \vee \mathcal{G}_k \right) \\ &= \sum_{j=1}^{N} \mathbb{P}\left(I^{N,i}_{k-1} = j \middle| \mathcal{F}^N_{k-1} \vee \mathcal{G}_k \right) \mathbb{P}\left(\boldsymbol{\xi}^{N,i}_k \in A \middle| I^{N,i}_{k-1} = j, \mathcal{F}^N_{k-1} \vee \mathcal{G}_k \right) \\ &= \sum_{j=1}^{N} \frac{\omega^{N,j}_{k-1}}{\Omega^N_{k-1}} R^{\mathrm{p}}_{k-1}\left(\boldsymbol{\xi}^{N,j}_{k-1}, A \right) , \quad i \in \{1, \ldots, N\} . \end{aligned}$$

That is, conditional on $\mathcal{F}_{k-1}^N$, the swarm $\{\boldsymbol{\xi}_k^{N,i}; 1 \le i \le N\}$ of mutated particles at time k is obtained by sampling N independent and identically distributed particles from the measure

$$\eta_k^N \triangleq \phi_{\nu,0:k-1|k-1}^N R_{k-1}^{\mathrm{p}} \, .$$

Using this notation, define, for $A \in \mathcal{X}^{\otimes(k+1)}$,

$$\mu_{k|n}^N(A) \triangleq \frac{\int_A W_k(\boldsymbol{x}_{k-1:k}) L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1}) \, \eta_k^N(\mathrm{d}\boldsymbol{x}_{0:k})}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \, .$$

Hence $\mu_{k|n}^N \ll \eta_k^N$, and the resulting Radon-Nikodým derivative is given by

$$\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{x}_{0:k}) \triangleq \frac{W_k(\boldsymbol{x}_{k-1:k}) L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1})}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \, .$$

Lemma 7.2. *Let $f \in \mathcal{B}_{\mathrm{b}}(\mathsf{X}^{n+1})$. Then, for all $n \ge 0$,*

$$\begin{aligned} &\phi_{\nu,0:n|n}^N f - \phi_{\nu,0:n|n} f \\ &\qquad = \sum_{k=1}^n \varphi_k^N(f) + \frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n} f \, , \end{aligned}$$

where

$$\varphi_k^N(f) \triangleq \frac{\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) \widehat{f}_{k:n}(\boldsymbol{\xi}_k^{N,j})}{\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j})} - \mu_{k|n}^N \widehat{f}_{k:n} \, , \tag{7.4}$$

and the real-valued functions $\{\widehat{f}_{k:n}; 1 \le k \le n+1\}$ are, for a fixed point $\widehat{\boldsymbol{x}}_{0:k} \in \mathsf{X}^{k+1}$, defined by

$$\widehat{f}_{k:n}(\boldsymbol{x}_{0:k}) \triangleq \frac{L_k \cdots L_{n-1} f(\boldsymbol{x}_{0:k})}{L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1})} - \frac{L_k \cdots L_{n-1} f(\widehat{\boldsymbol{x}}_{0:k})}{L_k \cdots L_{n-1}(\widehat{\boldsymbol{x}}_{0:k}, \mathsf{X}^{n+1})} \, . \tag{7.5}$$

Proof. As a starting point, consider the decomposition

$$\begin{aligned} \phi_{\nu,0:n|n}^N f - \phi_{\nu,0:n|n} f &= \frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n} f \\ &\quad + \sum_{k=1}^n \left[\frac{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1} f}{\phi_{\nu,0:k|k}^N L_k \cdots L_{n-1}(\mathsf{X}^{n+1})} - \frac{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1} f}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \right] \, . \end{aligned}$$

Now, since

$$\begin{aligned} &\frac{\phi_{\nu,0:k-1|k-1}^N R_{k-1}^{\mathrm{p}}[W_k \widehat{f}_{k:n} L_k \cdots L_{n-1}(\mathsf{X}^{n+1})]}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \\ &= \int_{\mathsf{X}^k} \int_{\mathsf{X}} \frac{\mathrm{d}Q(x_{k-1}, \cdot)}{\mathrm{d}R_{k-1}(x_{k-1}, \cdot)}(x_k) \frac{g_k(x_k) \widehat{f}_{k:n}(\boldsymbol{x}_{0:k}) L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1})}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} \\ &\qquad \times R_{k-1}(x_{k-1}, \mathrm{d}x_k) \, \phi_{\nu,0:k-1|k-1}^N(\mathrm{d}\boldsymbol{x}_{0:k-1}) \\ &= \frac{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1} f}{\phi_{\nu,0:k-1|k-1}^N L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} - \frac{L_k \cdots L_{n-1} f(\widehat{\boldsymbol{x}}_{0:k})}{L_k \cdots L_{n-1}(\widehat{\boldsymbol{x}}_{0:k}, \mathsf{X}^{n+1})} \, , \end{aligned}$$

we acquire the term-wise decomposition

$$\varphi_k^N(f) = \frac{\phi^N_{\nu,0:k|k}[\widehat{f}_{k:n} L_k \cdots L_{n-1}(\mathsf{X}^{n+1})]}{\phi^N_{\nu,0:k|k} L_k \cdots L_{n-1}(\mathsf{X}^{n+1})} - \frac{\phi^N_{\nu,0:k-1|k-1} R^{\mathrm{p}}_{k-1}[W_k \widehat{f}_{k:n} L_k \cdots L_{n-1}(\mathsf{X}^{n+1})]}{\phi^N_{\nu,0:k-1|k-1} L_{k-1} \cdots L_{n-1}(\mathsf{X}^{n+1})} ,$$

which is equivalent to the one presented in the lemma. □

Conditional on $\mathcal{F}^N_{k-1}$, the first term on the right-hand-side of (7.4) is nothing but an importance sampling estimate of $\mu^N_{k|n}\widehat{f}_{k:n}$, based on N independent η^N_k-distributed variables.

Lemma 7.3. *Assume 4.1 and let, for $n \geq 0$ and $0 \leq i \leq n$, f_i be a function of the form described in Proposition 7.1. Furthermore, let, for $k \geq 0$, the real valued function $\widehat{f}_{i,k:n}$ be defined through (7.5). Then*

$$\left\| \widehat{f}_{i,k:n} \right\|_{\mathsf{X}^{k+1},\infty} \leq 2\rho^{0 \vee (i-k)} \left\| f_i \right\|_{\mathsf{X}^{n+1},\infty} .$$

Proof. For $k \geq i$ we bound $\widehat{f}_{i,k:n}$ by $2\left\| f_i \right\|_{\mathsf{X}^{n+1},\infty}$; however, for $k < i$ a geometrically decreasing bound of the function can be obtained by using the exponential forgetting property of the conditional latent chain. Hence, since

$$\begin{aligned} \frac{L_k \cdots L_{n-1} f_i(\boldsymbol{x}_{0:k})}{L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1})} &= \mathbb{E}\left[f_i(\boldsymbol{X}_{i:n}) \middle| X_k = x_k, \mathcal{G}_n \right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f_i(\boldsymbol{X}_{i:n}) \middle| X_i = x_i, \mathcal{G}_n \right] \middle| X_k = x_k, \mathcal{G}_n \right] \\ &= \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\big(x_k, f^*_{i,n}\big) , \end{aligned}$$

where, for $x \in \mathsf{X}$, $f^*_{i,n}(x) \triangleq \mathbb{E}\left[f_i(\boldsymbol{X}_{i:n}) \middle| X_i = x, \mathcal{G}_n \right]$, we can, for $k < i$, rewrite $\widehat{f}_{i,k:n}$ as

$$\widehat{f}_{i,k:n}(\boldsymbol{x}_{0:k}) = \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\big(x_k, f^*_{i,n}\big) - \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\big(\widehat{x}_k, f^*_{i,n}\big) .$$

Applying Theorem 4.2 to this difference yields

$$\begin{aligned} &\left| \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\big(x_k, f^*_{i,n}\big) - \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\big(\widehat{x}_k, f^*_{i,n}\big) \right| \\ &\leq 2 \left\| f^*_{i,n} \right\|_{\mathsf{X},\infty} \left\| \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\left(x_k, \cdot\right) - \mathrm{F}_{k|n} \cdots \mathrm{F}_{i-1|n}\left(\widehat{x}_k, \cdot\right) \right\|_{\mathrm{TV}} \\ &\leq 2\rho^{i-k} \left\| f^*_{i,n} \right\|_{\mathsf{X},\infty} \leq 2\rho^{i-k} \left\| f_i \right\|_{\mathsf{X}^{n+1},\infty} . \end{aligned}$$

□

Lemma 7.4. *Assume 4.1 and let $n \geq 0$. Then, for all $1 \leq k \leq n$, $x \in \mathsf{X}$ and $N \geq 1$,*

$$\frac{\mathrm{d}\mu^N_{k|n}}{\mathrm{d}\eta^N_k}(x) \leq \frac{\left\| W_k \right\|_{\mathsf{X}^2,\infty}}{\mu g_k (1-\rho)\sigma_-} .$$

Proof. First, write

$$\begin{aligned}
&L_k \cdots L_{n-1}(\boldsymbol{x}_{0:k}, \mathsf{X}^{n+1}) \\
&= \int_{\mathsf{X}} q(x_k, x_{k+1}) L_{k+1} \cdots L_{n-1}(\boldsymbol{x}_{0:k+1}, \mathsf{X}^{n+1}) g_{k+1}(x_{k+1}) \, \mu(\mathrm{d}x_{k+1}) \\
&\leq \sigma_+ \int_{\mathsf{X}} L_{k+1} \cdots L_{n-1}(\boldsymbol{x}_{0:k+1}, \mathsf{X}^{n+1}) g_{k+1}(x_{k+1}) \, \mu(\mathrm{d}x_{k+1}) \; . \qquad (7.6)
\end{aligned}$$

Now, since the function $L_{k+1} \cdots L_{n-1}(\cdot, \mathsf{X}^{n+1})$ is constant in all but the last component of the argument,

$$\begin{aligned}
&L_{k-1} \cdots L_{n-1}(\boldsymbol{x}_{0:k-1}, \mathsf{X}^{n+1}) \\
&= \int_{\mathsf{X}} q(x_{k-1}, x_k) g_k(x_k) \int_{\mathsf{X}} q(x_k, x_{k+1}) L_{k+1} \cdots L_{n-1}(\boldsymbol{x}_{0:k+1}, \mathsf{X}^{n+1}) \\
&\qquad\qquad \times \; g_{k+1}(x_{k+1}) \, \mu(\mathrm{d}x_{k+1}) \, \mu(\mathrm{d}x_k) \\
&\geq \mu g_k \sigma_-^2 \int_{\mathsf{X}} L_{k+1} \cdots L_{n-1}(\boldsymbol{x}_{0:k+1}, \mathsf{X}^{n+1}) g_{k+1}(x_{k+1}) \, \mu(\mathrm{d}x_{k+1}) \; . \qquad (7.7)
\end{aligned}$$

Since the integrals in (7.6) and (7.7) are equal, the bound of the lemma follows. □

Proof of Proposition 7.1. We start with (i). Since, conditional on $\mathcal{F}_n^N$, the random variables $f_i(\widehat{\boldsymbol{\xi}}_n^{N,j})$, $1 \leq j \leq N$, are independent and identically distributed with expectation

$$\mathbb{E}\left[f_i(\widehat{\boldsymbol{\xi}}_n^{N,j}) \middle| \mathcal{F}_n^N \vee \mathcal{G}_n \right] = \frac{1}{\Omega_n^N} \sum_{j=1}^N \omega_n^{N,j} f_i(\boldsymbol{\xi}_n^{N,j}) \; , \qquad (7.8)$$

applying the Marcinkiewicz-Zygmund inequality provides the bound

$$N^{p/2} \mathbb{E}\left[\left| \frac{1}{N} \sum_{j=1}^N f_i(\widehat{\boldsymbol{\xi}}_n^{N,j}) - \frac{1}{\Omega_n^N} \sum_{j=1}^N \omega_n^{N,j} f_i(\boldsymbol{\xi}_n^{N,j}) \right|^p \middle| \mathcal{F}_n^N \vee \mathcal{G}_n \right] \leq C_p \, \|f_i\|_{\mathsf{X}^{n+1},\infty}^p \; , \qquad (7.9)$$

where C_p is a universal constant depending on p only. Having control of this discrepancy, we focus instead on the L^p error associated with the weighted empirical measure $\phi_{\nu,0:n|n}^N$. We make use of the identity

$$a/b - c/d = (a/b)(1 - b/d) + (a - c)/d$$

on each term of the decomposition provided by Lemma 7.2. This, together with Minkowski's inequality, gives us the bound

$$\left\|\varphi_k^N(f_i)\right\|_{p|\mathcal{F}_k^N\vee\mathcal{G}_n} \leq \left\|\frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j})\widehat{f}_{i,k:n}(\boldsymbol{\xi}_k^{N,j}) - \mu_{k|n}^N\widehat{f}_{i,k:n}\right\|_{p|\mathcal{F}_k^N\vee\mathcal{G}_n} + \left\|\widehat{f}_{i,k:n}\right\|_{\mathsf{X}^{k+1},\infty}\left\|\frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) - 1\right\|_{p|\mathcal{F}_k^N\vee\mathcal{G}_n}. \quad (7.10)$$

Applying the Marcinkiewicz-Zygmund inequality to the first term of this bound gives

$$N^{p/2}\mathbb{E}\left[\left|\frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j})\widehat{f}_{k:n}(\boldsymbol{\xi}_k^{N,j}) - \mu_{k|n}^N\widehat{f}_{k:n}\right|^p \middle| \mathcal{F}_{k-1}^N\vee\mathcal{G}_n\right] \leq C_p\left\|\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}\right\|_{\mathsf{X}^{k+1},\infty}^p \left\|\widehat{f}_{k:n}\right\|_{\mathsf{X}^{k+1},\infty}^p, \quad (7.11)$$

and treating the second term in a similar manner yields

$$N^{p/2}\mathbb{E}\left[\left|\frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) - 1\right|^p \middle| \mathcal{F}_{k-1}^N\vee\mathcal{G}_n\right] \leq C_p\left\|\frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}\right\|_{\mathsf{X}^{k+1},\infty}^p. \quad (7.12)$$

Thus, we obtain, by inserting these bounds into (7.10) and applying Lemma 7.3 and Lemma 7.4,

$$\left\|\varphi_k^N(f_i)\right\|_{p|\mathcal{F}_k^N\vee\mathcal{G}_n} \leq 4C_p^{1/p}\rho^{0\wedge(i-k)}\frac{\|W_k\|_{\mathsf{X}^2,\infty}\|f_i\|_{\mathsf{X}^{n+1},\infty}}{\sqrt{N}\mu g_k(1-\rho)\sigma_-}. \quad (7.13)$$

For the last difference of the decomposition provided by Lemma 7.2 we have, using the same decomposition technique as in (7.10),

$$\begin{aligned}\left\|\frac{\phi_{\nu,0}^N L_0\cdots L_{n-1}f_i}{\phi_{\nu,0}^N L_0\cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n}f_i\right\|_{p|\mathcal{G}_n} &\leq \|f_i\|_{\mathsf{X}^{n+1},\infty}\left\|\frac{1}{N}\frac{\sum_{j=1}^N \omega_0^{N,j}L_0\cdots L_{n-1}(\boldsymbol{\xi}_0^{N,j},\mathsf{X}^{n+1})}{\nu\left[g_0L_0\cdots L_{n-1}(\mathsf{X}^{n+1})\right]} - 1\right\|_{p|\mathcal{G}_n} \\ &+ \left\|\frac{1}{N}\sum_{j=1}^N \omega_0^{N,j}L_0\cdots L_{n-1}(\boldsymbol{\xi}_0^{N,j},f_i) - \nu\left[g_0L_0\cdots L_{n-1}f_i\right]\right\|_{p|\mathcal{G}_n} \\ &\times \frac{1}{\nu[g_0L_0\cdots L_{n-1}(\mathsf{X}^{n+1})]}. \end{aligned} \quad (7.14)$$

Since, repeating arguments of Lemma 7.4,

$$\frac{\|W_0L_0\cdots L_{n-1}(\mathsf{X}^{n+1})\|_{\mathsf{X},\infty}}{\nu[g_0L_0\cdots L_{n-1}(\mathsf{X}^{n+1})]} \leq \frac{\|W_0\|_{\mathsf{X},\infty}}{\nu g_0(1-\rho)},$$

we obtain, by applying the Marcinkiewicz-Zygmund inequality to each term of (7.14),

$$\left\| \frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f_i}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^n)} - \phi_{\nu,0:n|n} f_i \right\|_{p|\mathcal{G}_n} \leq 2C_p^{1/p} \frac{\|W_0\|_{\mathsf{X},\infty} \|f_i\|_{\mathsf{X}^{n+1},\infty}}{\sqrt{N}\nu g_0(1-\rho)} \,. \tag{7.15}$$

Now (i) follows by a straightforward application of Minkowski's inequality together with (7.9), (7.13) and (7.15).

We turn to (ii). By means of the identity

$$a/b - c = (a/b)(1-b)^2 + (a-c)(1-b) + c(1-b) + a - c$$

applied to (7.4), we obtain the bound

$$\begin{aligned} \left|\mathbb{E}\left[\varphi_k^N(f_i)\middle|\mathcal{F}_{k-1}^N \vee \mathcal{G}_n\right]\right| \leq \left\|\widehat{f}_{i,k:n}\right\|_{\mathsf{X}^{k+1},\infty} & \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) - 1 \right\|_{2|\mathcal{F}_{k-1}^N \vee \mathcal{G}_n}^2 \\ & + \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) \widehat{f}_{i,k:n}(\boldsymbol{\xi}_k^{N,j}) - \mu_{k|n}^N \widehat{f}_{i,k:n} \right\|_{2|\mathcal{F}_{k-1}^N \vee \mathcal{G}_n} \\ & \times \left\| \frac{1}{N}\sum_{j=1}^N \frac{\mathrm{d}\mu_{k|n}^N}{\mathrm{d}\eta_k^N}(\boldsymbol{\xi}_k^{N,j}) - 1 \right\|_{2|\mathcal{F}_{k-1}^N \vee \mathcal{G}_n} . \end{aligned}$$

Thus, we get, by reusing (7.11) and (7.12),

$$\begin{aligned} \left|\mathbb{E}\left[\varphi_k^N(f_i)\middle|\mathcal{G}_n\right]\right| &\leq \mathbb{E}\left[\left|\mathbb{E}\left[\varphi_k^N(f_i)\middle|\mathcal{F}_{k-1}^N \vee \mathcal{G}_n\right]\right| \middle| \mathcal{G}_n\right] \\ &\leq 4C_2 \rho^{0\vee(i-k)} \frac{\|W_k\|_{\mathsf{X}^2,\infty}^2 \|f_i\|_{\mathsf{X}^{n+1},\infty}}{N(\mu g_k)^2(1-\rho)^2\sigma_-^2} \,, \end{aligned} \tag{7.16}$$

and treating the last term of the decomposition in a completely similar manner yields

$$\left| \mathbb{E}\left[\frac{\phi_{\nu,0}^N L_0 \cdots L_{n-1} f_i}{\phi_{\nu,0}^N L_0 \cdots L_{n-1}(\mathsf{X}^{n+1})} - \phi_{\nu,0:n|n} f_i \middle| \mathcal{G}_n \right] \right| \leq 2C_2 \frac{\|W_0\|_{\mathsf{X}^2,\infty}^2 \|f_i\|_{\mathsf{X}^{n+1},\infty}}{N(\nu g_0)^2(1-\rho)^2} \,. \tag{7.17}$$

Finally, from (7.8) we conclude that the multinomial selection mechanism does not introduce any additional bias, and consequently (ii) follows from the triangle inequality together with (7.16) and (7.17). □

We are now, having established Proposition 7.1, prepared for proceeding with the proof of Theorem 4.5.

Proof of Theorem 4.5. Decomposing the difference in question yields the bound

$$\left\|\widehat{\gamma}_n^{N,\Delta_n} - \gamma_n\right\|_{p|\mathcal{G}_n} \leq \sum_{k=0}^{n-1} \left\|\widehat{\phi}^N_{\nu,0:k(\Delta_n)|k(\Delta_n)} s_k - \phi_{\nu,0:k(\Delta_n)|k(\Delta_n)} s_k\right\|_{p|\mathcal{G}_n} + \sum_{k=0}^{n-\Delta_n} \left|\phi_{\nu,0:k+\Delta_n|k+\Delta_n} s_k - \phi_{\nu,0:n|n} s_k\right| \ , \quad (7.18)$$

where we have set $k(\Delta_n) = (k+\Delta_n) \wedge n$. By writing

$$\begin{aligned}
&\mathbb{E}_\nu\left[s_k(X_k, X_{k+1}) \middle| X_{k+\Delta_n+1} = x_{k+\Delta_n+1}, \mathcal{G}_{k+\Delta_n}\right] \\
&= \mathbb{E}_\nu\left[\mathbb{E}_\nu\left[s_k(X_k, X_{k+1}) \middle| X_{k+1} = x_{k+1}, \mathcal{G}_{k+\Delta_n}\right] \middle| X_{k+\Delta_n+1} = x_{k+\Delta_n+1}, \mathcal{G}_{k+\Delta_n}\right] \\
&= \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}\left(x_{k+\Delta_n+1}, \widehat{s}_{k|k+\Delta_n}\right) \ ,
\end{aligned}$$

with, for $x \in \mathsf{X}$,

$$\widehat{s}_{k|k+\Delta_n}(x) \triangleq \mathbb{E}_\nu\left[s_k(X_k, X_{k+1}) \middle| X_{k+1} = x, \mathcal{G}_{k+\Delta_n}\right] \ ,$$

we get that

$$\begin{aligned}
&\phi_{\nu,0:k+\Delta_n|k+\Delta_n} s_k - \phi_{\nu,0:n|n} s_k \\
&= \psi_{k+\Delta_n+1|k+\Delta_n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}\left(x_{k+\Delta_n+1}, \widehat{s}_{k|k+\Delta_n}\right) \\
&\qquad - \psi_{k+\Delta_n+1|n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}\left(x_{k+\Delta_n+1}, \widehat{s}_{k|k+\Delta_n}\right) \ .
\end{aligned}$$

where we have defined, for $\ell, m \geq 0$, $\psi_{\ell|m} \triangleq \mathbb{P}_\nu(X_\ell \in \cdot | \mathcal{G}_m)$. Hence, we obtain, using the exponential forgetting property (see Theorem 4.2) of the time-reversed conditional hidden chain,

$$\begin{aligned}
&\left|\phi_{\nu,0:k+\Delta_n|k+\Delta_n} s_k - \phi_{\nu,0:n|n} s_k\right| \\
&\leq 2\left\|\widehat{s}_{k|k+\Delta_n}\right\|_{\mathsf{X},\infty} \left\|\psi_{k+\Delta_n+1|k+\Delta_n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}\left(x_{k+\Delta_n+1}, \cdot\right)\right. \\
&\qquad \left. - \psi_{k+\Delta_n+1|k+\Delta_n} \mathrm{B}_{\nu,k+\Delta_n|k+\Delta_n} \cdots \mathrm{B}_{\nu,k+1|k+\Delta_n}\left(x_{k+\Delta_n+1}, \cdot\right)\right\|_{\mathrm{TV}} \\
&\leq 2\rho^{\Delta_n} \left\|s_k\right\|_{\mathsf{X}^2,\infty} \ . \qquad (7.19)
\end{aligned}$$

Plugging (7.19) and the bound of Proposition 7.1(i) into the decomposition (7.18) completes the proof of (i). The proof of part (ii) is entirely analogous and is omitted for brevity. □

7.2. **Proof of Proposition 4.7.**

Assumption 7.5. *Let f_i be the function of Proposition 7.1. For $p \geq 2$, $\ell \geq 1$ there exists a constant $\alpha^{(n)}_{p,\ell}(f_i) \in \mathbb{R}$ such that*

$$\sup_{0 \leq k \leq n} \mathring{\mathbb{E}}\left[\frac{\|W_k\|^p_{\mathsf{X},\infty} \left\|f_i\right\|^\ell_{\mathsf{X}^{n+1},\infty}}{(\mu g_k)^p}\right] \vee \mathring{\mathbb{E}}\left[\left\|f_i\right\|^\ell_{\mathsf{X}^{n+1},\infty}\right] \leq \alpha^{(n)}_{p,\ell}(f_i) \ .$$

Proposition 7.6. *Assume 4.1 and 4.3. Then, the following holds true for all $N \geq 1$.*

(i) If Assumption 7.5 is satisfied for $\ell = p \geq 2$, then

$$\left\|\widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i\right\|_{p,\circ} \leq \frac{B_p \beta_p^{(n)}(f_i)}{\sqrt{N}(1-\rho)} \left[\frac{1-\rho^i}{\sigma_-(1-\rho)} + \frac{n-i}{\sigma_-} + \frac{1}{\inf_{x\in\mathsf{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right],$$

where $\beta_p^{(n)}(f_i) \triangleq [\alpha_{p,p}^{(n)}(f_i)]^{1/p}$;

(ii) if Assumption 7.5 is satisfied for $p = 2$, $\ell = 1$, then

$$\left|\mathring{\mathbb{E}}\left[\widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i\right]\right| \leq \frac{B \alpha_{2,1}^{(n)}(f_i)}{N(1-\rho)^2} \left[\frac{1-\rho^i}{\sigma_-^2(1-\rho)} + \frac{n-i}{\sigma_-^2} + \frac{1}{\inf_{x\in\mathsf{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} + 1\right].$$

Proof. The proof of the first part is straightforward: combining Proposition 7.1 and Minkowski's inequality provides the bound

$$\begin{aligned}
&\left\|\widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i\right\|_{p,\circ} \\
&= \mathring{\mathbb{E}}^{1/p}\left[\left\|\widehat{\phi}^N_{\nu,0:n|n} f_i - \phi_{\nu,0:n|n} f_i\right\|^p_{p,\circ|\mathring{\mathcal{G}}_n}\right] \\
&\leq \frac{B_p}{\sqrt{N}(1-\rho)} \left\{\frac{1}{\sigma_-} \sum_{k=1}^{n} \mathring{\mathbb{E}}^{1/p}\left[\frac{\|W_k\|^p_{\mathsf{X},\infty}\, \|f_i\|^p_{\mathsf{X}^{n+1},\infty}}{(\mu g_k)^p}\right] \rho^{0\vee(i-k)} \right. \\
&\qquad \left. + \frac{1}{\inf_{x\in\mathsf{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)} \mathring{\mathbb{E}}^{1/p}\left[\frac{\|W_0\|^p_{\mathsf{X},\infty} \|f_i\|^p_{\mathsf{X}^{n+1},\infty}}{(\mu g_0)^p}\right] + \mathring{\mathbb{E}}^{1/p}\left[\|f_i\|^p_{\mathsf{X}^{n+1},\infty}\right]\right\}.
\end{aligned}$$

We finish the proof by plugging the bounds of Assumption 7.5 into the expression above and summing up. The proof of the second part follows similarly. □

Proof of Proposition 4.7. The proof of the first part follows by applying Proposition 7.6 and the bound (7.19) to the decomposition

$$\begin{aligned}
\left\|\widehat{\gamma}_n^{N,\Delta_n} - \gamma_n\right\|_{p,\circ} \leq \sum_{k=0}^{n-1} &\left\|\widehat{\phi}^N_{\nu,0:k(\Delta_n)|k(\Delta_n)} s_k - \phi_{\nu,0:k(\Delta_n)|k(\Delta_n)} s_k\right\|_{p,\circ} \\
&+ \sum_{k=0}^{n-\Delta_n} \left\|\phi_{\nu,0:k+\Delta_n|k+\Delta_n} s_k - \phi_{\nu,0:n|n} s_k\right\|_{p,\circ}.
\end{aligned}$$

The second part is proved in a similar manner. □

References

Andrieu, C. and Doucet, A. (2003). Online Expctation-Maximization type algorithms for parameter estimation in general state space models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*

Cappé, O., Moulines, E., Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer-Verlag, New York.

Cérou, N., Le Gland, F. and Newton, N. (2001). Stochastic Particle methods for linear tangent filtering equations. In *Optimal Control and PDE's—Innovations and applications, in honor of Alain's Bensoussans 60th anniversary* (eds. J.-L. Menaldi, E. Rofman and A. Sulem), 231–240. IOS Press.

Chigansky, P. and Lipster, R. (2004). Stability of nonlinear filters in nonmixing case. *Ann. Appl. Prob.* **14**, 2038–2056.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **32**, 2385–2411.

Crisan, D. (2001). Particle filters—a theoretical perspective. In *Sequential Monte Carlo methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.), 17–41. Springer-Verlag, New York.

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

Del Moral, P. and Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. Lecture Notes in Mathematics **1832**. Springer-Verlag, Berlin.

Del Moral, P. and Guionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré, Probab. et Stat.* **37**, 155–194.

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

Deylon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27**, 94–128.

Douc, R. and Mathias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7**, 381–420.

Douc, R. and Moulines, É. (2005). Limit theorems for weighted samples with applications to Sequential Monte Carlo Methods. eprint `arXiv:math.ST/0507042`.

Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32**, 2254–2304.

Doucet, A., de Freitas, N. and Gordon, N. (2001). An introduction to sequential Monte Carlo Methods. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.), 3–14. Springer-Verlag, New York.

Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, New York.

Doucet, A. and Tadić, V.B., (2003). Parameter estimation in general state space models using particle methods. *Ann. Inst. Statist. Math.* **55**, 409–422.

Doucet, A. and Tadić, V.B. (2005). Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models. *Stoch. Proc. Appl.* **115**, 1408–1436.

Doucet, A., Godsill, J. and West, M. (2004). Monte Carlo smoothing for nonlinear time series. *J. Am. Statist. Assoc.* **99**, 156–168.

Fort, G. and Moulines, É. (2003). Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.* **31**, 1220–1259.

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance* **42**, 281–300.

Jaquier, E., Polson, N.G. and Rossi, P.E. (1994). Bayesian analysis of stochastic volatility models (with discussion). In *J. Bus. Econom. Statist.* **12**, 371–417.

Kitagawa, G. and Sato, S. (2001). Monte-Carlo smoothing and self-organising state-space model. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.), 178–195. Springer-Verlag, New York.

Künsch, H. (2001). State space and hidden Markov models. In *Complex Stochastic Systems* (O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg, eds.), 109–173. Chapman and Hall.

Lindvall, T. (1992). *Lectures on the Coupling Method.* Wiley, New York.

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *J. Am. Statist. Assoc.* **430**, 567–576.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. B* **44**, 226–233.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, London.

Olsson, J. and Rydén, T. (2004). Asymptotic properties of the bootstrap particle filter maximum likelihood estimator for state space models. Technical Report 2005:18, Dept. of Math. Stat., Lund Univ.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables.* Oxford Studies in Probability, vol. 4, Oxford Science Publications, Clarendon Press, Oxford.

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

Pólya, G. and Szegö, G. (1976). *Problems and Theorems in analysis.* **2**. Springer, New York.

Ristic, B., Arulampalam, M. and Gordon, A. (2004). *Beyond the Kalman Filter: Particle Filters for Target Tracking* Atrech House Radar Library.

Sandmann, G. and Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics* **87**, 271–301.

Tanner, M. A. (1993). *Tools for Statistical Inference.* Springer, 2nd ed.

Wei, G. C. G. and Tanner, M. A. (1991). A Monte Carlo implementation of the EM algorithm and the poor man's Data Augmentation algorithms. *J. Am. Statist. Assoc.* **85**, 699–704.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103.

(J. Olsson) Center for Mathematical Sciences, Lund University, Sweden

(O. Cappé) Ecole Nationale Supérieure des Télecommunications, France

(R. Douc) CMAP, Ecole Polytechnique, France

(É. Moulines) Ecole Nationale Supérieure des Télecommunications, France