Central limit theorem for sequential Monte Carlo methods and its applications to Bayesian inference

Nicolas Chopin†

Laboratoire de Statistique, CREST, Paris, France.

Summary. The terms 'Sequential Monte Carlo methods', or similarly 'particle filters', refer to a general class of iterative algorithms which perform Monte Carlo approximations of a given sequence of distributions of interest (π_t). Their use is usually justified by first-order asymptotics, i.e. it is shown that computed estimates converge almost surely as the number of 'particles' (simulated values) tends towards infinity. In this paper, we establish a central limit theorem for these estimates. This result holds under minimal assumptions on the distributions π_t, and apply in a general framework which encompasses most of sequential Monte Carlo methods that have been considered in the literature, including the resample-move algorithm of Gilks and Berzuini (2001) or the residual resampling scheme of Liu and Chen (1998). The corresponding asymptotic variances provide a convenient measurement of the precision of a given particle filter. We study in particular in some typical examples of Bayesian applications whether and to which rate these asymptotic variances diverge in time, in order to assess the long term reliability of the considered algorithm.

Keywords: Convergence of sequential Monte Carlo methods; Markov Chain Monte Carlo; Particle filter; State-space models

Résumé. Les expressions 'méthodes de Monte Carlo séquentielles', ou 'filtres particulaires', font référence à une classe générale d'algorithmes itératifs effectuant des approximations de Monte Carlo d'une suite donnée de distributions d'intérêt (π_t). Leur utilisation est justifiée par une asymptotique d'ordre un, i.e. on démontre que les estimateurs calculés convergent presque sûrement lorsque le nombre de 'particules' (valeurs simulées) tend vers l'infini. Dans cet article, nous établissons un théorème central limite pour ces estimateurs. Ce résultat requiert des hypothèses très faibles sur les distributions (π_t), et s'applique dans un cadre général qui comprend la plupart des méthodes de Monte Carlo séquentielles considérées dans la littérature, y compris l'algorithme 'resample-move' de Gilks et Berzuini (2001), ou le ré-échantillonnage résiduel de Liu et Chen (1998). Les variances asymptotiques correspondantes permettent de mesurer de façon simple la précision d'un filtre particulaire donné. Nous étudions en particulier dans quelques exemples typiques d'applications Bayésiennes la vitesse de divergence de ces variances asymptotiques, afin d'évaluer la stabilité de long terme de l'algorithme considéré.

Mots-clés: Convergence des méthodes de Monte Carlo séquentielles; Filtre particulaire; Markov Chain Monte Carlo; Modèle à espace d'état

†*Address for correspondence:* INSEE, Timbre F410, 18 Boulevard Auguste Pinard, 75675 Paris Cedex, France.
E-mail: chopin@ensae.fr

1. Introduction

Sequential Monte Carlo methods form an emerging yet already very active branch of the Monte Carlo paradigm. Their growing popularity comes in part from the fact that they are often the only viable computing techniques in those situations where data must be processed sequentially. Their range of applicability is consequently very wide, and includes non exclusively signal processing, financial modelling, speech recognition, computer vision, neural networks, molecular biology and genetics, target tracking, geophysics, among others. A very good introduction to the field has been written by Künsch (2001), while the edited volume of Doucet et al. (2001) provides an interesting coverage of recent developments in theory and applications.

Specifically, sequential Monte Carlo methods (which are often termed as 'particle filters') are useful in any problem which involves a large sequence of distributions of interest $\pi_t(d\theta_t)$. In a sequential Bayesian framework, $\pi_t(d\theta_t)$ will usually represent the posterior distribution of parameter θ_t given the t first observations. The term 'parameter' must be understood here in a broad sense, in that θ_t may include any unknown quantity which may be inferred from the t first observations, and is not necessarily of constant dimension. We denote Θ_t the support of $\pi_t(d\theta_t)$.

Due to their Monte Carlo nature, the justification of sequential Monte Carlo methods usually takes the form of a law of large numbers, that is, computed estimates are shown to converge almost surely to the quantity of interest as the number of 'particles' (simulated values) tends towards infinity. This result however gives little insight on the rate of convergence of these estimates. To complete this first convergence theorem, we establish in this paper a central limit theorem for these estimates. This result holds under weak assumptions, and applies in a general framework which encompasses most algorithms found in the literature, see §2.2 for references. We will see for instance that it applies to the resample-move algorithm of Gilks and Berzuini (2001), in which particles are 'moved' from time to time through a given MCMC (Monte Carlo Markov Chain) transition, in order to introduce new particle values.

To our knowledge, there have been previously two attempts in establishing a central limit theorem for particle filters. First, the theorem of Del Moral and Guionnet (1999) only applies to a basic sequential Monte Carlo method which does not incorporate any 'move' mechanism. Furthermore, their proof is specific to the nature of distributions π_t, namely it is assumed that π_t is the posterior density of a state-space model which fulfils given conditions. In contrast, we do not make any assumption on π_t. The central limit theorem of Gilks and Berzuini (2001) is closer in spirit to our work, since the possibility of 'moving' particles is considered. But as pointed out by the authors themselves the considered mode of convergence (as $H_1, H_2, ..., H_t \to +\infty$, recursively, where H_t is the number of particles at iteration t) is not very realistic, since the number of particles usually remains identical through iterations. In this respect one major point of our work is that convergence is established as $H = H_1 = H_2 = \ldots$ goes towards infinity. This complicates the proof but is certainly more relevant to the actual behaviour of particle filters. Furthermore, both papers cited above only consider the multinomial resampling scheme of Gordon et al. (1993), while our theorem equally applies to algorithms which resort to the residual resampling scheme of Liu and Chen (1998), which is a more efficient alternative.

We believe that our central limit theorem can become a precious tool in the study of theoretical aspects of particle filtering. By way of illustration, we formalize some principles that have been stated previously on a heuristic basis. For instance, the residual resampling

scheme is indeed more efficient than the multinomial resampling scheme, since it is shown to lead to a smaller asymptotic variance. Similarly, the 'Rao-Blackwellization' technique of Doucet et al. (2000), which consists in integrating out when possible some dimensions of θ_t, is shown to reduce the asymptotic variance.

The most promising application of our central limit theorem is the possibility to assess the stability of a given particle filter (in terms of precision of computed estimates) through the time behaviour of the corresponding asymptotic variances. This is a critical issue since it is well known that sequential Monte Carlo methods tend to degenerate in numerous cases, sometimes at a very fast rate. We consider in this paper some typical Bayesian problems, such as the sequential analysis of state-space models. In this particular case, $\pi_t(d\theta_t)$ will stand for the posterior distribution of the sequence of state variables $x_1, \ldots, x_t$, possibly completed by a fixed parameter θ, and θ_t will be therefore of increasing dimension. We will show that under some conditions stability can be achieved at least for 'filtering' the states, that is for approximating the marginal posterior density $\pi_t(x_t)$.

We will also provide some interesting results for problems where π_t more simply represents the posterior distribution of a constant parameter $\theta_t = \theta$. While overlooked in the literature, this case has many practical applications. Firstly, there are some examples of dynamical models such that the marginal posterior density of the fixed parameter θ (that is to say integrated over latent variables $x_1, \ldots, x_t$) is tractable. It is then possible to apply the aforementioned 'Rao-Blackwellization' technique and consider the sequence of these marginal densities instead of the joint densities of θ and $x_1, \ldots, x_t$. Related algorithms, see Doucet et al. (2000), Chen and Liu (2000), and Chopin (2001), therefore fall in this 'constant parameter' category. Secondly, particle filtering allows more generally for sequentially estimating any parametric model that does not involve a hidden process. Note this sequential framework may be purely instrumental, in that the observations $y_1, \ldots, y_T$ to be taken into account may be immediately available, $\pi(d\theta|y_1, \ldots, y_T)$ being the only posterior distribution of interest. In such a case only the final output of the algorithm is considered. Chopin (2002) argues that particle filtering is an efficient computational strategy for the Bayesian inference of 'static' parametric models, provided that the sample size is important, since accesses to data are reduced.

The paper is organized as follows. Section 2 proposes a generic description of particle filters, establishes in this general framework a central limit theorem for computed estimates, and draws some conclusions from this result. Section 3 discusses the stability of particle filters through the time behaviour of the asymptotic variances provided by the central limit theorem. Section 4 proposes several directions for further research. Proofs of theorems are put in the Appendix.

2. Central limit theorem for particle filters

2.1. General formulation of particle filters

In full generality, a particle system is a triangular array of random variables in $\Theta \times \mathbb{R}^+$,

$$(\theta^{(j,H)}, w^{(j,H)})_{j \leq H},$$

where Θ is some space of interest. The variables $\theta^{(j,H)}$ are usually called 'particles', and their contribution to the sample may vary according to their weight $w^{(j,H)}$. We will say

that this particle system *targets* a given distribution π defined on Θ if and only if

$$\frac{\sum_{j=1}^{H} w^{(j,H)} \varphi(\theta^{(j,H)})}{\sum_{j=1}^{H} w^{(j,H)}} \to \mathbb{E}_{\pi}(\varphi) \tag{1}$$

holds almost surely as $H \to +\infty$ for any measurable function φ such that the expectation above exists. A first example of particle system is a denumerable set of independent draws from π, with unit weights, which obviously targets π. In this simple case, particles and weights do not depend on H, and the particle system is rather a sequence than a triangular array. This is not the case in general however, and, while cumbersome, the dependence in H will be maintained in notations to allow for a rigourous mathematical treatment.

Now assume a sequence $(\pi_t)_{t\in\mathbb{N}}$ of distributions defined on a sequence of probabilized spaces (Θ_t). In most if not all applications, Θ_t will be a power of the real line or some subset of it, and henceforth $\pi_t(\cdot)$ will also denote the density of π_t with respect to an appropriate version of Lebesgue measure. A sequential Monte Carlo algorithm (or particle filter) is a method for producing a particle system whose target evolves in time: at iteration t of the algorithm, the particle system targets π_t, and therefore allows for Monte Carlo approximations of the distribution of (current) interest π_t. Clearly enough particle filters do not operate in practice on infinite triangular arrays but rather manipulate particle vectors of pre-chosen size H. One must keep in mind however that the justification of such methods is essentially asymptotic, and therefore justifies this abstract framework.

The structure of a particle filter can be decomposed in three basic iterative operations, that we will refer to hereafter as mutation, correction and selection steps. At the beginning of iteration t, consider a particle system $(\theta_{t-1}^{(j,H)}, 1)_{j\leq H}$, that is with unit weights, currently targeting π_{t-1}. The mutation step consists in producing new particles drawn from

$$\theta_t^{(j,H)} \sim k_t(\theta_{t-1}^{(j,H)}, d\theta_t),$$

where k_t is a transition kernel which maps Θ_{t-1} into the set $\mathcal{P}(\Theta_t)$ of probability measures over Θ_t. The 'mutated' particles (with unit weights) targets the new distribution $\widetilde{\pi}_t = \int \pi_{t-1}(\theta_{t-1}) k_t(\theta_{t-1}, \cdot) d\theta_{t-1}$. This distribution $\widetilde{\pi}_t$ is usually not relevant to the considered application, but rather serves as an intermediary stage for practical reasons. To shift the target to distribution of interest π_t, particles are assigned weights proportional to

$$w_t^{(j,H)} \propto \upsilon_t(\theta_t^{(j,H)}), \text{ with } \upsilon_t(\theta_t) = \pi_t(\theta_t)/\widetilde{\pi}_t(\theta_t).$$

This is the correction step. The particle system $(\theta_t^{(j,H)}, w_t^{(j,H)})$ targets π_t. The function υ_t is referred to as the weight function. Note the normalizing constants of densitiesπ_t and $\widetilde{\pi}_t$ are intractable in most applications. This is why weights are defined up to a multiplicative constant, which has no bearings anyway on the estimates produced by the algorithm, since they are weighted averages.

Finally, the selection step consists in replacing the current vector of particles by a new, uniformly weighted vector $(\widehat{\theta}_t^{(j,H)}, 1)_{j\leq H}$ which contains a number n_j of replicates of particle $\theta_t^{(j,H)}$, $n_j \geq 0$. The n_j's are random variables such that $\sum n_j = H$ and $\mathbb{E}(n_j) = w^{(j,H)} / \sum_j w_t^{(j,H)}$. In this way, particles with too small a weight are discarded, while particles with important weight serve as a multiple starting point for the next mutation step. They are various ways for generating the n_j's. Denote ρ_j the normalized

weights,

$$\rho_j = w_t^{(j,H)} / \sum_{j=1}^{H} w_t^{(j,H)}$$

where dependencies in H and t are omitted for convenience. Multinomial resampling (Gordon et al., 1993) amounts to drawing independently the H new particles from the multinomial distribution which produces $\theta_t^{(j,H)}$ with probability ρ_j. Residual resampling (Liu and Chen, 1998) consists in reproducing $\lfloor H\rho_j \rfloor$ times each particle $\theta_t^{(j,H)}$, where $\lfloor \cdot \rfloor$ stands for the integer part. The particle vector is completed by $H^r = H - \sum_j \lfloor H\rho_j \rfloor$ independent draws from the multinomial distribution which produces $\theta_t^{(j,H)}$ with probability $(H\rho_j - \lfloor H\rho_j \rfloor)/H^r$. Another interesting selection scheme has been proposed by Whitley (1994), for which the number of replicates n_j is ensured to differ from $H\rho_j$ by at most one. It has been rediscovered and christened 'systematic sampling' by Carpenter et al. (1999). We failed however to extend our results to this third selection scheme.

The structure of a particle filter can be summarized as follows.

---

1. Mutation: Draw for $j = 1, \ldots, H$,
$$\theta_t^{(j,H)} \sim k_t(\theta_{t-1}^{(j,H)}, d\theta_t),$$
where $k_t : \Theta_{t-1} \to \mathcal{P}(\Theta_t)$.
2. Correction: Assign weights to particles so that, for $j = 1, \ldots, H$,
$$w_t^{(j,H)} \propto \upsilon_t(\theta_t^{(j,H)}) = \pi_t(\theta_t^{(j,H)}) / \widetilde{\pi}_t(\theta_t^{(j,H)}),$$
where $\widetilde{\pi}_t(\cdot) = \int \pi_{t-1}(\theta_{t-1}) k_t(\theta_{t-1}, \cdot)\, d\theta_{t-1}$.
3. Selection: resample, according to a given selection scheme,
$$(\theta_t^{(j,H)}, w_t^{(j,H)})_{j \leq H} \to (\widehat{\theta}_t^{(j,H)}, 1)_{j \leq H}.$$

The first mutation step, $t = 0$, is assumed to create initial particles by drawing independently from some instrumental distribution $\widetilde{\pi}_0$.

It is shown without difficulty that the particle system produced by this generic algorithm does target iteratively the distributions of interest, that is the following convergences hold almost surely,

$$\begin{aligned}
H^{-1} \sum_{j=1}^{H} \varphi(\theta_t^{(j,H)}) &\to \mathbb{E}_{\widetilde{\pi}_t}(\varphi) \\
\frac{\sum_{j=1}^{H} w_t^{(j,H)} \varphi(\theta_t^{(j,H)})}{\sum_{j=1}^{H} w_t^{(j,H)}} &\to \mathbb{E}_{\pi_t}(\varphi) \\
H^{-1} \sum_{j=1}^{H} \varphi(\widehat{\theta}_t^{(j,H)}) &\to \mathbb{E}_{\pi_t}(\varphi)
\end{aligned}$$

provided the corresponding expectations exist. These convergences will be referred to as the law of large number for particle filters. This law of large number ensures the consistency of the weighed averages above, but gives little insight on their precision. This calls for a central limit theorem that completes these consistency results.

2.2. Some examples of particle filters

The general formulation given in the previous section encompasses most sequential Monte Carlo methods described in the literature. By way of illustration, assume first that distributions π_t are defined on a common space $\Theta_t = \Theta$. Such a situation arises for instance in a sequential setting where the only unknown quantity is a 'fixed' parameter θ. In a Bayesian framework, π_t will be the posterior density of θ, given the t first observations, $\pi_t(\theta) = \pi(\theta|y_{1:t})$, where $y_{1:t}$ denotes the sequence of the observations $y_1, \ldots, y_t$. If particles are not mutated, k_t being the 'identity kernel' $k_t(\theta, \cdot) = \delta_\theta$, we have $\widetilde{\pi}_t = \pi_{t-1}$ for $t > 0$, and our generic particle filter becomes one of the variations of the sequential importance resampling algorithm (Rubin, 1988; Gordon et al., 1993; Liu and Chen, 1998). The weight function simplifies to

$$\upsilon_t(\theta) = \pi(\theta|y_{1:t})/\pi(\theta|y_{1:t-1}) \propto p(y_t|y_{1:t-1}, \theta)$$

in a Bayesian model, where $p(y_t|y_{1:t-1}, \theta)$ is the conditional likelihood of y_t, given the parameter θ and previous observations.

In this setting, the support of π_t is expected to concentrate progressively on a certain region of the space Θ, and to eventually converge to a Dirac mass centred at some θ_0. In the Bayesian framework, this phenomenon represents the accumulation of information on θ as more and more data is taken into account. If k_t is set to the identity kernel, the set of plausible values for particles is generated once and for all from $\widetilde{\pi}_0$ at the first step of the algorithm. Less and less of these values should consequently contribute to a correct representation of π_t as t grows, the others being assigned a very little weight or even discarded through the successive selection steps. To counter this degeneracy effect and introduce new values in the particle sample, Gilks and Berzuini (2001) proposed to mutate particles through a given kernel k_t which admits π_{t-1} as an invariant density. In that case, we still have $\widetilde{\pi}_t = \pi_{t-1}$, and the expression of weight functions is unchanged. Such a kernel is usually built through MCMC (Markov Chain Monte Carlo) methodology (see Robert and Casella, 1999, for an authoritative presentation). The ability of such a kernel to 'rejuvenate' the particle system seems to be related to its mixing properties, in that a strong dependence in previous state θ_{t-1} should prevent any reduction of degeneracy.

Now consider the case where π_t is defined on a space of increasing dimension of the form $\Theta_t = \mathcal{X}^t$. A typical application is the sequential inference of a dynamical model which involves a latent process (x_t), and π_t stands then for density $\pi(x_{1:t}|y_{1:t})$. Assume k_t can be decomposed in

$$k_t(x_{1:t-1}, dx_{1:t}) = \kappa_t(x_{1:t-1}, dx_{1:t-1}) q_t(x_t|x_{1:t-1})\, dx_t,$$

where $\kappa_t : \mathcal{X}^{t-1} \to \mathcal{P}(\mathcal{X}^{t-1})$ is a transition kernel, and $q_t(\cdot|\cdot)$ is a conditional density. If κ_t admits π_{t-1} as an invariant density, the weight function verifies

$$\upsilon_t(x_{1:t}) = \frac{\pi_t(x_{1:t})}{\pi_{t-1}(x_{1:t-1}) q_t(x_t|x_{1:t-1})}. \tag{2}$$

Again, the case where κ_t is the identity kernel corresponds to some version of the sequential importance resampling algorithm, while setting κ_t to a given MCMC transition kernel with invariant density π_{t-1} leads to the resample-move algorithm of Gilks and Berzuini (2001).

This second setting is more involved. In certain applications, one is primarily interested in inferring the last state x_t (Bayesian filtering), and the distribution of interest is the

marginal density of π_t in x_t rather than π_t itself. It may be sufficient in that case to resort to the sequential importance resampling algorithm, provided that the conditional density $q_t(x_t|x_{1:t-1})$ has the ability to rejuvenate the particle sample by 'forgetting' the previous states, according to a principle similar to the MCMC rejuvenation presented above. The density q_t must be chosen so that the weight function $\nu_t = \widetilde{\pi}_t/\pi_t$ is not too distorted. The simplest solution is to set q_t to the conditional prior density of x_t, given $x_{1:t-1}$, as suggested originally by Gordon et al. (1993). It is usually more efficient however to take into account in some way the information carried by y_t, in order to simulate more values compatible with the observations, as exemplified by the algorithm of Pitt and Shephard (1999). Whatever q_t, the sequential resampling algorithm is not an efficient method for inferring the whole trajectory $x_{1:t}$ (Bayesian smoothing). These assertions on the behaviour of the sequential importance resampling algorithm will be formalized more properly in the second part of this paper.

These two previous cases can be combined into one, by considering a dynamic model which features in the same time a fixed parameter θ and a sequence of latent variables (x_t), so that $\Theta_t = \Theta \times \mathcal{X}^t$, and π_t stands for the joint posterior density $\pi(\theta, x_{1:t}|y_{1:t})$. This does not change much the structure of the algorithm described in the second case, but without an efficient MCMC rejuvenation strategy depletion in parameter values must be expected for essentially the same reasons as in the first case.

Note finally that there exists an ever simpler particle filter algorithm which does not pertain to our general formulation. This is the sequential importance sampling which alternates mutation and correction steps, but does not perform any selection step. Weights are consequently not initialized to one at each iteration, and are rather updated through

$$w_t^{(j)} \propto w_{t-1}^{(j)} \upsilon_t(\theta_t^{(j)}).$$

We suppress any notational dependence in H since it is meaningless in such a case. Due to its specific nature, this algorithm will be treated separately, see §3.1.

2.3. Central limit theorem

We define the following variance-like quantities, which will play the role of asymptotic variances in our central limit theorem. Let, for any measurable $\varphi : \Theta_0 \to \mathbb{R}^d$, $\widetilde{V}_0(\varphi) = \mathbb{V}_{\widetilde{\pi}_0}(\varphi)$, and by induction, for any $\varphi : \Theta_t \to \mathbb{R}^d$,

$$\widetilde{V}_t(\varphi) = \widehat{V}_{t-1}\{\mathbb{E}_{k_t}(\varphi)\} + \mathbb{E}_{\pi_{t-1}}\{\mathbb{V}_{k_t}(\varphi)\}, \quad t > 0, \tag{3}$$

$$V_t(\varphi) = \widetilde{V}_t\left[\upsilon_t.(\varphi - \mathbb{E}_{\pi_t}\varphi)\right], \quad t \geq 0, \tag{4}$$

$$\widehat{V}_t(\varphi) = V_t(\varphi) + \mathbb{V}_{\pi_t}(\varphi), \quad t \geq 0. \tag{5}$$

Notations $\mathbb{E}_{k_t}(\varphi)$ and $\mathbb{V}_{k_t}(\varphi)$ are short-hands for, respectively, functions $\mu(\theta_{t-1}) = \mathbb{E}_{k_t(\theta_{t-1},\cdot)}\{\varphi(\cdot)\}$ and $\Sigma(\theta_{t-1}) = \mathbb{V}_{k_t(\theta_{t-1},\cdot)}\{\varphi(\cdot)\}$. Note these equations do not necessary lead to definite quantities for any φ. We now precise the classes of functions for which the central limit theorem enunciated below will hold, and in particular for which these asymptotic variances exist. Define recursively $\Phi_t^{(d)}$ to be the set of measurable $\varphi : \Theta_t \to \mathbb{R}^d$ such that for some $\delta > 0$,

$$\mathbb{E}_{\widetilde{\pi}_t}\|\upsilon_t.\varphi\|^{2+\delta} < +\infty, \tag{6}$$

and that function $\theta_{t-1} \mapsto \mathbb{E}_{k_t(\theta_{t-1},\cdot)}\{\upsilon_t(\cdot)\varphi(\cdot)\}$ is in Φ_{t-1}^d. The initial set $\Phi_0^{(d)}$ contains all measurable functions whose moments of order two over $\widetilde{\pi}_0$ are finite.

THEOREM 1. *If the selection step consists in multinomial resampling, and provided that the unit function* $\theta_t \mapsto 1$ *belongs to* $\Phi_t^{(1)}$ *for every* t*, then for any* $\varphi \in \Phi_t^{(d)}$*,* $\mathbb{E}_{\pi_t}(\varphi)$*,* $V_t(\varphi)$ *and* $\widehat{V}_t(\varphi)$ *are finite quantities, and the following convergences in distribution hold as* $H \to +\infty$*,*

$$H^{1/2}\left\{\frac{\sum_{j=1}^H w_t^{(j,H)}\varphi(\theta_t^{(j,H)})}{\sum_{j=1}^H w_t^{(j,H)}} - \mathbb{E}_{\pi_t}(\varphi)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, V_t(\varphi)\},$$

$$H^{1/2}\left\{H^{-1}\sum_{j=1}^H \varphi(\widehat{\theta}_t^{(j,H)}) - \mathbb{E}_{\pi_t}(\varphi)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widehat{V}_t(\varphi)\}.$$

A proof is given in the Appendix. In the course of this proof an additional central limit theorem is established for the unweighted particle system $(\theta_t^{(j,H)}, 1)$ produced by the mutation step, which targets $\widetilde{\pi}_t$. This result is not given here however, for it holds for a slightly different class of functions, and is of less practical interest. The condition that function $\theta_t \mapsto 1$ belongs to $\Phi_t^{(1)}$ implies that the weight function υ_t has finite moment of order $2+\delta$ over $\widetilde{\pi}_t$, for some $\delta > 0$, and therefore restricts somehow the dispersion of particle weights. Note this condition also ensures that Φ_t^d contains all bounded functions.

A central limit theorem also holds when the selection step follows the residual sampling scheme of Liu and Chen (1998), but this imposes some change in the expression of the asymptotic variances. The new expression of $\widehat{V}_t(\varphi)$ is

$$\widehat{V}_t(\varphi) = V_t(\varphi) + R_t(\varphi), \tag{7}$$

where

$$R_t(\varphi) = \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\varphi\varphi'\} - \frac{1}{\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\}}\left[\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\varphi\}\right]\left[\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\varphi\}\right]', \tag{8}$$

and $r(x)$ is x minus its integer part.

THEOREM 2. *Results of Theorem 1 still hold when the selection steps consists in residual resampling, except that the asymptotic variances are now defined by equations (3), (4) and (7).*

When the mutation step follows the multinomial scheme, a simpler alternative to iterative formulae (3) to (5) is the close-form expression

$$V_t(\varphi) = \sum_{k=0}^t \mathbb{E}_{\widetilde{\pi}_k}\left[\upsilon_k^2 \mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}'\right], \tag{9}$$

where $\mathcal{E}_t$ is the functional operator which attributes to φ the function

$$\mathcal{E}_t(\varphi) : \theta_{t-1} \mapsto \mathbb{E}_{k_t(\theta_{t-1},\cdot)}\left[\upsilon_t(\cdot)\left\{\varphi(\cdot)\right\}\right], \tag{10}$$

and $\mathcal{E}_{k+1:t}(\varphi) = \mathcal{E}_{k+1} \circ \ldots \circ \mathcal{E}_t(\varphi)$ for $k+1 \leq t$, $\mathcal{E}_{t+1:t}(\varphi) = \varphi$. A similar formula for the residual case can be obtained indirectly by deriving the variation of the asymptotic variance incurred by resorting to the residual scheme rather than the multinomial scheme, that is, for $t > 0$,

$$V_t^r(\varphi) - V_t(\varphi) = \widehat{V}_{t-1}^r(\varphi) - \widehat{V}_{t-1}(\varphi) = \sum_{k=0}^{t-1}\left[R_k\{\mathcal{E}_{k+1:t}(\varphi)\} - \mathbb{V}_{\pi_k}\{\mathcal{E}_{k+1:t}(\varphi)\}\right], \tag{11}$$

where $V_t^r(\varphi)$ and $\widehat{V}_{t-1}^r(\varphi)$ are the asymptotic variances defined in the residual case, through (3), (4) and (7). In the following, we will similarly distinguish the residual case through a r-suffix in notations.

2.4. First conclusions

A first application of this central limit theorem is to provide more rigourous justification for some heuristics that have been stated previously in the literature, see for instance Liu and Chen (1998). Inequalities in this section refer to the canonical order for symmetric matrices, that is to say $A > B$ (resp. $A \geq B$) if and only if $A - B$ is positive definite (resp. positive semidefinite).

First, it is preferable to derive any estimate before the selection step, since the immediate effect of the latter is a strict increasing of asymptotic variance: $\widehat{V}(\varphi) > V(\varphi)$ for any non constant function φ. In this respect one may wonder why selection steps should be performed. We will see that this immediate degradation of the particle system is often largely compensated by gains in precision for future inference.

Second, residual sampling outperforms multinomial resampling in every case. Let $\varphi : \Theta_t \to \mathbb{R}^d$ and $\underline{\varphi} = \varphi - \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\varphi\}$, then

$$\begin{aligned} R_t(\varphi) &= \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\underline{\varphi}\underline{\varphi}'\}, \\ &\leq \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)(\underline{\varphi} - \mathbb{E}_{\pi_t}\underline{\varphi})(\underline{\varphi} - \mathbb{E}_{\pi_t}\underline{\varphi})'\} \\ &\leq \mathbb{V}_{\pi_t}(\varphi) \end{aligned}$$

since $r(x) \leq x$. It follows from this inequality and (11) that $V_t(\varphi) \geq V_t^r(\varphi)$. Actually, a substantial gain should be expected when using the residual scheme since the majoration used above is clearly not sharp.

Our central limit theorem also provides formal justification for resorting to 'marginalized' particle filter, as explained in the following section.

2.5. Marginalized particle filters

In some specific cases, it is possible to decompose each space Θ_t in $\Xi_t \times \Lambda_t$ in such a way that the marginal density π_t^m of π_t over Ξ_t is computable (up to a positive constant). When such a structure can be exhibited, it is beneficial to implement a particle filter which tracks the marginal densities π_t^m instead of the 'complete' densities π_t, since it produces more precise estimators (in a sense that we precise below). The idea of resorting to 'marginalized' particle filters has been formalized by Doucet et al. (2000), and implemented in various settings by Chen and Liu (2000), Chopin (2001) and Andrieu and Doucet (2002), among others. Doucet et al. (2000)'s justification for resorting to 'marginalized' particle filters is that they feature importance weights with smaller a variance than their 'unmarginalized' counterpart, which suggests that the produced estimates are also less variable. This is proven by a Rao-Blackwell decomposition, and consequently 'marginalized' particle filters are sometimes referred to as 'Rao-Blackwellized' particle filters. We now extend the argument of these authors by proving that asymptotic variances for every estimator are indeed smaller in the 'marginalized' case. Assume decompositions of π_t and $\widetilde{\pi}_t$ of the form

$$\pi_t(\theta_t) = \pi_t^m(\xi_t)\pi_t^c(\lambda_t|\xi_t), \qquad \widetilde{\pi}_t(\theta_t) = \widetilde{\pi}_t^m(\xi_t)\widetilde{\pi}_t^c(\lambda_t|\xi_t),$$

where (ξ_t, λ_t) identifies to θ_t, and π_t^m, π_t^c, $\widetilde{\pi}_t^m$, $\widetilde{\pi}_t^c$, are, respectively, marginal and conditional densities of ξ_t and λ_t. Consider two particle filters, tracking respectively (π_t) and (π_t^m). It is assumed that both filters resort to the same selection scheme (whether multinomial or residual), and that their mutation steps consist in drawing respectively from kernels k_t and k_t^m, where the latter is the 'marginal' version of the former, that is to say the following probability measures coincide on $\Theta_t = \Xi_t \times \Lambda_t$,

$$\int_{\Lambda_{t-1}} \pi_{t-1}^c(\lambda_{t-1}|\xi_{t-1}) k_t\{(\xi_{t-1}, \lambda_{t-1}), (d\xi_t, d\lambda_t)\}\, d\lambda_{l-1} = k_t^m(\xi_{t-1}, d\xi_t)\widetilde{\pi}_t^c(\lambda_t|\xi_t)\, d\lambda_t, \quad (12)$$

for almost every $\xi_{t-1} \in \Xi_{t-1}$. This equality implies in particular that

$$\int \pi_{t-1}^m(\xi_{t-1}) k_t^m(\xi_{t-1}, \cdot)\, d\xi_{t-1} = \widetilde{\pi}_t^m(\cdot).$$

Asymptotic variances and other quantities are distinguished similarly through the m-suffix for the marginal case, that is $V_t(\varphi)$ and $V_t^m(\varphi)$ and so on.

THEOREM 3. *For any $\varphi : \Xi_t \to \mathbb{R}^d$, we have $V_t^m(\varphi) \leq V_t(\varphi)$ and $V_t^{m,r}(\varphi) \leq V_t^r(\varphi)$. These inequalities are attained for a non constant φ if and only if $\pi_k^c(\cdot|\xi_t) = \widetilde{\pi}_k^c(\cdot|\xi_t)$ for almost every $\xi_t \in \Xi_t$.*

As suggested by the condition for equality above or more clearly exhibited in the proof in the Appendix, marginalizing allows for cancelling weight dispersion due to discrepancy between conditional densities $\widetilde{\pi}_t^c$ and π_t^c, while the part due to discrepancy between marginal densities π_t^m and $\widetilde{\pi}_t^m$ remains identical.

Beyond the small number of cases where this marginalization technique can be effectively carried out, this result has also strong qualitative implications. In the following sections, we will study the behaviour of the time sequence $V_t(\varphi)$ in order to measure whether and to which rate a given particle filter 'diverges'. In this respect, we will be able in some cases to build a marginalized particle filter whose rate of divergence is theoretically known, thus providing a lower bound for the actual rate of divergence of the considered particle filter.

3. Stability of particle filters

3.1. Sequential importance sampling

The study of the sequential importance sampling algorithm is much simpler that any other particle filter. Since particles are not resampled, they remain independent through iterations. It follows through the standard central limit theorem that

$$H^{1/2}\{\frac{\sum_{j=1}^H w_t^{(j)} \varphi(\theta_t^{(j)})}{\sum_{j=1}^H w_t^{(j)}} - \mathbb{E}_{\pi_t}(\varphi)\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, V_t^{sis}(\varphi)\},$$

where the corresponding asymptotic variance is

$$V_t^{sis}(\varphi) = \mathbb{E}_{\widetilde{\pi}_t}\left[\frac{\pi_t}{\widetilde{\pi}_t}\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\}\right]^2,$$

and $\widetilde{\pi}_t$ denotes this time the generating distribution of particles $\theta_t^{(j,H)}$ obtained by recursion of mutation kernels $k_t(\cdot,\cdot)$, that is,

$$\widetilde{\pi}_t(\cdot) = \int \widetilde{\pi}_{t-1}(\theta_{t-1}) k_t(\theta_{t-1}, \cdot)\, d\theta_{t-1},$$

the distribution $\widetilde{\pi}_0$ being arbitrary. Sequential importance sampling is rarely an efficient algorithm, but the value of $V_t^{sis}(\varphi)$ may serve as a benchmark in some occasions, as we will see in the following.

3.2. Sequential importance sampling and resampling in fixed parameter case

We have explained that in the fixed parameter case, that is $\Theta_t = \Theta$ and $\pi_t(\theta) = \pi(\theta|y_{1:t})$, π_t is expected to get more and more informative on θ, and to eventually converge to a Dirac mass at some point θ_0. Sequential importance sampling and resampling algorithms typically diverge in such a situation, since they generate once and for all the set of possible particle values from $\widetilde{\pi}_0$, a majority of which being presumably far from θ_0. The following result quantifies this degeneracy effect.

THEOREM 4. *Let $\varphi : \Theta \to \mathbb{R}^d$, then under suitable regularity conditions, there exists constants c_1, c_2 and c_3 such that*

$$V_t^{sis}(\varphi) \sim c_1 t^{p/2-1}, \quad V_t^r(\varphi) \sim c_2 t^{p/2}, \quad V_t(\varphi) \sim c_3 t^{p/2},$$

where p is the dimension of Θ, and $V_t^r(\varphi)$, $V_t(\varphi)$ refer here to the sequential importance resampling case, that is $k_t(\theta, \cdot) = \delta_\theta$.

The conditions mentioned above amount to assume that π_t is the posterior density of a model regular enough to ensure existence and asymptotic normality of the maximum likelihood estimator. Under such conditions, π_t can be approximated at first order as a Gaussian distribution centred at θ_0 with variance $I(\theta_0)^{-1}/t$, where $I(\theta_0)$ is the Fisher information matrix evaluated at θ_0. Results above are then obtained through the Laplace approximation of integrals, see the Appendix. It may seem paradoxical that $V_t(\varphi)$ converges to zero when $p = 1$, but if we rather study the quantity $V_t(\varphi)/\mathbb{V}_{\pi_t}(\varphi)$, which measures the precision of the algorithm relatively to the variation of the considered function, we see that this ratio diverges even when $p = 1$, since usually $\mathbb{V}_{\pi_t}(\varphi) \sim I(\theta_0)^{-1}/t$ as $t \to +\infty$.

That the sequential importance resampling diverges quicker than the sequential importance sampling is not surprising: when particles are not mutated, the only effect of a selection step is to deplete the particle system. In this respect, we have for any non constant function φ,

$$V_t^{sis}(\varphi) < V_t^r(\varphi) \leq V_t(\varphi).$$

The proof of this inequality is straightforward.

Due to its facility of implementation and results above, it may be recommended to use the sequential importance sampling algorithm for studying short series of observations, provided that the dimension of Θ is low. But in general one should rather implement a more elaborate particle filter which includes mutation steps in order to counter the particle depletion. A further-reaching implication of these results is the following. Consider a dynamical model which involves a fixed parameter θ, and assume that the marginal posterior distributions $\pi(\theta|y_{1:t})$, obtained by marginalizing out latent variables $x_{1:t}$, fulfil the regularity conditions of Theorem 4. Then, following the argument developed in §2.5, we got that the rate of divergence of the sequential importance resampling algorithm for this kind of model is at least of order $O(t^{p/2})$, where p is the dimension of this fixed parameter.

3.3. Sequential importance sampling and resampling for Bayesian filtering and smoothing

For simplicity we assume that $\pi_t(x_{1:t}) = \pi(x_{1:t}|y_{1:t})$ is the posterior density of a state space model with latent Markov process (x_t), $x_t \in \mathcal{X}$ and observed process (y_t), $y_t \in \mathcal{Y}$ which fulfil equations

$$\begin{aligned} y_t|x_t &\sim f(y_t|x_t)\,dy_t, \\ x_t|x_{t-1} &\sim g(x_t|x_{t-1})\,dx_t. \end{aligned}$$

We distinguish two types of functions: those which are defined on common dimensions of the spaces $\Theta_t = \mathcal{X}^t$, say $\varphi : x_{1:t} \to \varphi(x_k)$, for $t \geq k$, and those which are evaluated on the 'last' dimension of Θ_t, that is $\varphi : x_{1:t} \to \varphi(x_t)$. Evaluating these two types of functions amounts to, respectively, 'smoothing' or 'filtering' the states.

The sequential importance sampling algorithm is usually very inefficient in such a setting, whether for smoothing or filtering the states. We illustrate this phenomenon by a simple example. Assume the t-th mutation step consists in drawing x_t from the prior conditional density $g(x_t|x_{t-1})$, which is usually easy to implement. Consider two evolving particles $\theta_t^{(j)} = x_{1:t}^{(j)}$ with weights $w_t^{(j)}$, $j = 1, 2$. We have

$$\log \frac{w_t^{(1)}}{w_t^{(2)}} = \sum_{k=1}^{t} \frac{f(y_k|x_k^{(1)})}{f(y_k|x_k^{(2)})}.$$

Assuming that the joint process $(y_t, x_t^{(1)}, x_t^{(2)})$ is stationary, we should be able to exhibit conditions under which a central limit theorem of the like

$$t^{-1/2} \sum_{k=0}^{t} \log \frac{f(y_k|x_k^{(1)})}{f(y_k|x_k^{(2)})} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \tag{13}$$

holds, where the limiting distribution is centred for symmetry reasons. Note this convergence is with respect to the joint probability space of simulated processes $x_t^{(j)}$, $j = 1, 2$ and the observation process (y_t), while all our previous results were for a given sequence of observations. In this simple example however it is sufficient to assume the equations of the considered state-space model define a stationary joint Markov process(x_t, y_t), (x_t referring here to the 'true' latent process) to ensure that both the observation process (y_t) and the processes $x_t^{(j)}$ (simulated according to the same transitions as (x_t)) are stationary. In this way, (13) yields that the ratio of weights of the two particles either converges or diverges exponentially fast. More generally when H particles are generated initially, very few of them will have a prominent weight after some iterations, thus leading to very unreliable estimates, whether for smoothing or filtering the states. The algorithm suffers from the curse of dimensionality, in that its degeneracy grows exponentially in the dimension of the space of interest Θ_t.

We now turn to the sequential importance resampling algorithm, and remark first that, for $\varphi : x_{1:t} \to \varphi(x_1)$ and $t > 0$

$$V_t(\varphi) \geq V_t^r(\varphi) > V_t^{sis}(\varphi),$$

provided φ is not constant. The proof of this inequality is straightforward. The sequential importance resampling algorithm is even more inefficient than the sequential importance sampling algorithm in smoothing the first state x_1, because the successive selection steps

simply worsen the deterioration of the particle system in the x_1 dimension. The same remarks apply more generally to any state x_k. This is consistent with our claim in §2.4 that a selection step always degrades the inference on current states, but may possibly improve the inference on future states. In this respect, the algorithm is expected to show more capability in filtering the states, as argued in §2.2, and we now turn to the study of the filtering stability.

The functional operator $\mathcal{E}_t$ which appears in the expression of $V_t(\varphi)$, see (9), summarizes two contradictory effects: on one hand, the weight distortion due to the correction step, and on the other hand the rejuvenation of particles due to application of kernel k_t. Stability will be achieved provided that these two effects compensate in some way.

For simplicity, we assume that the state space $\mathcal{X}$ is included in the real line and that the studied filtering function $\varphi : x_{1:t} \to \varphi(x_t)$ is real-valued. Recall that for the sequential importance resampling algorithm, k_t verifies,

$$k_t(x_{1:t-1}, dx_{1:t}) = \delta_{x_{1:t-1}}.q_t(x_t|x_{1:t-1})dx_t,$$

for some arbitrary conditional distribution $q_t(\cdot|\cdot)$. We assume that q_t only depends on the previous value x_{t-1}, and therefore defines a Markov transition. The ability of q_t to 'forget past values' is usually expressed through its contraction coefficient, see Dobrushin (1956),

$$\rho_t = \frac{1}{2} \sup_{x',x'' \in \mathcal{X}} ||q_t(\cdot|x') - q_t(\cdot|x'')||_1,$$

where $||\cdot||_1$ stands for the L_1-norm. Note $\rho_t \leq 1$, and if $\rho_t < 1$, q_t is said to be strictly contractive. Define the variation of a given function φ by

$$\Delta\varphi = \sup_{x,x' \in \mathcal{X}} |\varphi(x) - \varphi(x')|,$$

then the coefficient ρ_t measures the extent to which the application q_t 'contracts' the variation of the considered function, that is for $x', x'' \in \mathcal{X}$,

$$|\int q_t(x|x')\varphi(x)\,dx - \int q_t(x|x'')\varphi(x)\,dx| \leq \rho_t \Delta\varphi. \tag{14}$$

Furthermore, it is known, see for instance Künsch (2001), that if q_t is such that, for all $x, x', x'' \in \mathcal{X}$,

$$\frac{q_t(x|x')}{q_t(x|x'')} \leq C,$$

then its contraction coefficient verifies $\rho_t \leq 1 - C$. We therefore make such assumptions in order to prove the stability of the sequential importance resampling algorithm.

THEOREM 5. *Assume that $\Delta\varphi < +\infty$ and there exist constants C, $\underline{f}$ and $\overline{f}$ such that, for any $t \geq 0, x, x', x'' \in \mathcal{X}$, $y \in \mathcal{Y}$,*

$$\frac{g(x|x')}{g(x|x'')} \leq C, \qquad \frac{q_t(x|x')}{q_t(x|x'')} \leq C, \qquad \underline{f} \leq f(y|x) \leq \overline{f} \tag{15}$$

then $V_t(\varphi)$ is bounded from above in t.

This theorem is akin to previous results in the literature, see Del Moral and Guionnet (2001), Legland (2001), and most especially Künsch (2001), except that these authors do not study the stability of the asymptotic variance but rather of some distance (such as the total variation norm of the difference) between the 'true' filtering density $\pi_t(x_t)$ and the empirical density computed from the particle system. Furthermore, the aforementioned references only consider the particular case where the instrumental distribution $q_t(x_t|x_{t-1})$ is set to $g(x_t|x_{t-1})$, while our theorem is more general. Unfortunately all these results, including ours, require strong assumptions such as (15) that are unrealistic when $\mathcal{X}$ is not compact. Further research will hopefully provide weaker assumptions but this may prove an especially arduous problem.

3.4. Resample-move algorithms, variance estimation

We term as 'resample-move algorithm' any particle filter algorithm which includes an MCMC step in order to reduce degeneracy, as described in §2.2. It seems difficult to make general statements about such algorithms and we will rather make some informal comments.

The fixed parameter case is especially well-behaved. Basic particle filters diverge only at a polynomial rate, as seen in §3.2, in contrast with the exponential rate for state-space models. Adding (well calibrated) MCMC mutation steps should consequently lead to stable algorithms in many cases of interest. In fact it is doubtful that a mutation step must be performed at each iteration to achieve stability. Chopin (2002) argues and provides some experimental evidence that it may be sufficient to perform move steps at a logarithmic rate, that is the n-th move step should occur at iteration $t_n \sim \exp(\alpha n)$.

Situations where a latent process intervenes seem less promising. Smoothing the states is especially a difficult problem, and we do not think that there is any solution for circumventing the curse of dimensionality that we have pointed out in previous section. Even if mutation steps are performed at every iteration, the MCMC transition kernels should themselves suffer from the curse of dimensionality, in that their ability to rejuvenate particles of dimension t is likely to decrease in t.

Resample-move algorithms remain an interesting alternative when the considered dynamic model includes a fixed parameter θ. MCMC mutation steps should avoid depletion in simulated values of θ, and make it possible at least to filter the states and estimate the parameter under reasonable periods of time. Unfortunately the corresponding MCMC transition kernels will often depend on the whole past trajectory, so that long term stability remains uncertain.

In such complicated setups it is necessary to monitor at least numerically the degeneracy of the considered particle filter algorithm. We propose the following method. Run k, say $k = 10$, parallel independent particle filters of size H. For any quantity to be estimated, compute the average of the k corresponding estimates. This new estimator is clearly consistent and asymptotically normal. Moreover, the computational cost of this strategy is identical to that of a single particle filter of size kH, while the obtained precision will be also of the same order of magnitude in both cases, that is to say $V_t(\varphi)/(kH)$. This method does not therefore incur unnecessary computational load, and allows for assessing the stability of the algorithm through the evolution of the variability between these k estimates.

Acknowledgments

I would like to thank Professor C.P. Robert for his guidance and helpful comments. This paper is the fourth part of my Ph.D. thesis.

Appendix

A1. Proof of Theorems 1 and 2

The proof works by induction of Lemmas 1, 2 and 3 for Theorem 1, and Lemmas 1, 2 and 4 for Theorem 2. Assume for a given $t > 0$ that, for any $\varphi \in \Phi_{t-1}^{(d)}$,

$$H^{1/2} \left\{ \frac{1}{H} \sum_{j=1}^{H} \varphi(\widehat{\theta}_{t-1}^{(j,H)}) - \mathbb{E}_{\pi_{t-1}}(\varphi) \right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widehat{V}_{t-1}(\varphi)\}. \tag{16}$$

LEMMA 1 (MUTATION). *Let $\psi : \Theta_t \to \mathbb{R}^d$, assume the function $\mu : \theta_{t-1} \mapsto E_{k_t(\theta_{t-1},\cdot)}\{\psi(\cdot) - \mathbb{E}_{\widetilde{\pi}_t}(\psi)\}$ belongs to $\Phi_{t-1}^{(d)}$ and there exists $\delta > 0$ such that $\mathbb{E}_{\widetilde{\pi}_t}\|\psi\|^{2+\delta} < +\infty$, then*

$$H^{1/2} \left\{ \frac{1}{H} \sum_{j=1}^{H} \psi(\theta_t^{(j,H)}) - \mathbb{E}_{\widetilde{\pi}_t}(\psi) \right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widetilde{V}_t(\psi)\}.$$

PROOF. We assume first that is ψ is real-valued ($d = 1$). The generalization to $d > 1$ will follow directly from Cramer-Wold Theorem.

Let $\overline{\psi} = \psi - \mathbb{E}_{\widetilde{\pi}_t}(\psi)$, $\mu(\theta_{t-1}) = \mathbb{E}_{k_t(\theta_{t-1},\cdot)}\{\overline{\psi}(\cdot)\}$, $\sigma^2(\theta_{t-1}) = \mathbb{V}_{k_t(\theta_{t-1},\cdot)}\{\overline{\psi}(\cdot)\}$ and $\sigma_0^2 = \mathbb{E}_{\pi_{t-1}}(\sigma^2)$. We have $\mathbb{E}_{\pi_{t-1}}(\mu) = 0$, and by Jensen inequality

$$\sigma_0^2 = \mathbb{E}_{\pi_{t-1}}\left[\mathbb{V}_{k_t(\theta_{t-1},\cdot)}\{\psi(\cdot)\}\right] \leq \{\mathbb{E}_{\widetilde{\pi}_t}|\psi|^{(2+\delta)}\}^{2/(2+\delta)} < +\infty,$$

which makes it possible to apply the law of large numbers for particle filters toσ^2,

$$H^{-1} \sum_{j=1}^{H} \sigma^2(\theta_{t-1}^{(j,H)}) \to \sigma_0^2 \text{ almost surely.} \tag{17}$$

Defining

$$\begin{aligned} \nu(\theta_{t-1}) &= \mathbb{E}_{k_t(\theta_{t-1},\cdot)}\{|\psi(\cdot) - \mu(\theta_{t-1})|^{2+\delta}\} && (18)\\ &\leq 2^{1+\delta}\{\mathbb{E}_{k_{t-1}(\theta_{t-1},\cdot)}|\psi(\cdot)|^{2+\delta} + |\mathbb{E}_{k_{t-1}(\theta_{t-1},\cdot)}\psi(\cdot)|^{2+\delta}\} && (19)\\ &\leq 2^{2+\delta}\{\mathbb{E}_{k_{t-1}(\theta_{t-1},\cdot)}|\psi(\cdot)|^{2+\delta}\} && (20) \end{aligned}$$

where (19) comes from C_r inequality, and (20) from Jensen inequality, we deduce that

$$\mathbb{E}_{\pi_{t-1}}(\nu) \leq 2^{2+\delta}\mathbb{E}_{\widetilde{\pi}_t}|\psi|^{2+\delta} < +\infty.$$

Note this inequality ensures that the expectations defining ν in (18) (and similarly these defining μ and σ^2) are finite for almost every θ_{t-1}. It follows that

$$H^{-1} \sum_{j=1}^{H} \nu(\theta_{t-1}^{(j,H)}) \to \mathbb{E}_{\pi_{t-1}}(\nu) \text{ almost surely,}$$

and combining this result with (17), we get the almost sure convergence of sequence

$$\rho_H = \frac{\sum_{j=1}^H \nu(\theta_{t-1}^{(j,H)})}{\left\{\sum_{j=1}^H \sigma^2(\theta_{t-1}^{(j,H)})\right\}^{(2+\delta)/2}} = H^{-\delta/2} \frac{H^{-1}\sum_{j=1}^H \nu(\theta_{t-1}^{(j,H)})}{\left\{H^{-1}\sum_{j=1}^H \sigma^2(\theta_{t-1}^{(j,H)})\right\}^{(2+\delta)/2}} \to 0. \tag{21}$$

Let $T_H = H^{-1/2}\sum_{j=1}^H \overline{\psi}(\theta_t^{(j,H)})$, S_{t-1} denote the sigma-field generated by the random variables forming the triangular array $(\theta_{t-1}^{(j,H)})_{j\leq H}$, that is the particle system at time $t-1$, and $\mu_H = \mathbb{E}(T_H|S_{t-1})$. Conditionally on S_{t-1}, the $\overline{\psi}(\theta_t^{(j,H)})$'s form a triangular array of independent variables which fulfill Liapounov condition, see (21), and have variances whose mean converges to σ_0^2, see (17). Therefore (Billingsley, 1995, p. 362) the following central limit theorem holds

$$(T_H - \mu_H)|S_{t-1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_0^2). \tag{22}$$

We have also, by applying (16) to function μ,

$$\mu_H = H^{-1/2}\sum_{j=1}^H \mu(\theta_{t-1}^{(j,H)}) \xrightarrow{\mathcal{D}} \mathcal{N}\left\{0, \widehat{V}_{t-1}(\mu)\right\}. \tag{23}$$

The characteristic function of T_H is

$$\begin{aligned}\Phi_{T_H}(u) &= \mathbb{E}\{\exp(iuT_H)\},\\ &= \mathbb{E}[\exp(iu\mu_H)\mathbb{E}\{\exp(iuT_H - iu\mu_H)|S_{t-1}\}],\end{aligned}$$

where $\mathbb{E}\{\exp(iuT_H - iu\mu_H)|S_{t-1}\}$ is the characteristic function of $T_H - \mu_H$ conditionally on S_{t-1}, which according to (22) converges to $\exp(-\sigma_0^2u^2/2)$. It follows from (23) that

$$\exp(iu\mu_H)\mathbb{E}\{\exp(iuT_H - iu\mu_H)|S_{t-1}\} \xrightarrow{\mathcal{D}} \exp(-\sigma_0^2u^2/2)\exp\left[iu\mathcal{N}\left\{0, \widehat{V}_{t-1}(\mu)\right\}\right].$$

The expectation of the left term converges to the expectation of the right term according to the dominated convergence theorem, and this completes the proof.

LEMMA 2 (CORRECTION). *Let $\varphi \in \Phi_t^{(d)}$, assume function $\theta_t \mapsto 1$ belongs to $\Phi_t^{(1)}$, then*

$$H^{1/2}\left\{\frac{\sum_{j=1}^H w_t^{(j,H)}\varphi(\theta_t^{(j,H)})}{\sum_{j=1}^H w_t^{(j,H)}} - \mathbb{E}_{\pi_t}(\varphi)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, V_t(\varphi)\}.$$

PROOF. Let $\overline{\varphi} = \varphi - \mathbb{E}_{\pi_t}(\varphi)$ and apply Lemma 1 to the vector function $\psi = (\upsilon_t.\overline{\varphi}, \upsilon_t)'$:

$$H^{1/2}\left\{\frac{1}{H}\sum_{j=1}^H \begin{pmatrix}\upsilon_t(\theta_t^{(j,H)})\overline{\varphi}(\theta_t^{(j,H)})\\ \upsilon_t(\theta_t^{(j,H)})\end{pmatrix} - \begin{pmatrix}0_{\mathbb{R}^d}\\ 1\end{pmatrix}\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widetilde{V}_t(\psi)\}. \tag{24}$$

Then resorting to the δ-method with function $g(x,y) = x/y$ we obtain

$$H^{1/2}\frac{\sum_{j=1}^H \upsilon_t(\theta_t^{(j,H)})\overline{\varphi}(\theta_t^{(j,H)})}{\sum_{j=1}^H \upsilon_t(\theta_t^{(j,H)})} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{V}) \tag{25}$$

where $\mathcal{V} = \{(\partial g/\partial x, \partial g/\partial y)(0,1)\}\widetilde{V}_t(\psi)\{(\partial g/\partial x, \partial g/\partial y)(0,1)\}' = \widetilde{V}_t\{\upsilon_t.(\varphi - \mathbb{E}_{\pi_t}\varphi)\}$. We can replace the $\upsilon_t(\theta_t^{(j,H)})$'s by weights $w_t^{(j,H)}$ in the above equation since they are proportional.

LEMMA 3 (SELECTION, MULTINOMIAL RESAMPLING). *Let $\widehat{V}_t(\varphi) = V_t(\varphi) + \mathbb{V}_{\pi_t}(\varphi)$ and assume the particle system is resampled according to multinomial scheme, then, under the same conditions as previous Lemma,*

$$H^{1/2}\left\{\frac{1}{H}\sum_{j=1}^{H}\varphi(\widehat{\theta}_t^{(j,H)}) - \mathbb{E}_{\pi_t}(\varphi)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widehat{V}_t(\varphi)\}.$$

PROOF. The proof is most similar to the proof of Lemma 1. Denote S_t the sigma-field generated by the random variables $(\theta_t^{(j,H)}, w_t^{(j,H)})_{j\leq H}$ and let $\overline{\varphi} = \varphi - \mathbb{E}_{\pi_t}(\varphi)$, $T_H = H^{-1/2}\sum_{j=1}^{H}\overline{\varphi}(\widehat{\theta}_t^{(j,H)})$ and $\mu_H = \mathbb{E}(T_H|S_t)$. Conditionally on S_{t-1}, T_H is, up to factor $H^{-1/2}$, a sum of independent draws from the multinomial distribution which produces $\overline{\varphi}(\theta_t^{(j,H)})$ with probability $w_t^{(j,H)}/\sum_{j=1}^{H} w_t^{(j,H)}$. Then, as in Lemma 1, we have

$$(T_H - \mu_H)|S_t \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_0^2)$$

where this time $\sigma_0^2 = \mathbb{V}_{\pi_t}(\varphi)$, which is the limit as $H \to +\infty$ of the variance of the multinomial distribution mentioned above. Proof is completed along the same lines as in Lemma 1.

LEMMA 4 (SELECTION, RESIDUAL RESAMPLING). *Let $\widehat{V}_t(\varphi)$ takes the value given by (7) and assume the particle system is resampled according to residual resampling scheme, then, under the same conditions as Lemma 2,*

$$H^{1/2}\left\{\frac{1}{H}\sum_{j=1}^{H}\varphi(\widehat{\theta}_t^{(j,H)}) - \mathbb{E}_{\pi_t}(\varphi)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \widehat{V}_t(\varphi)\}.$$

PROOF. The proof is identical to the proof of previous Lemma, except that conditionally on S_t, T_H is $H^{1/2}$ times a constant plus a sum of independent draws from multinomial distribution which produces $\overline{\varphi}(\theta_t^{(j,H)})$ with probability $r(Hw_t^{(j,H)}/\sum_{j=1}^{H} w_t^{(j,H)})$. This yields a different value for σ_0^2,

$$\sigma_0^2 = E_{\widetilde{\pi}_t}\{r(\upsilon_t).\varphi^2\} - \frac{1}{\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\}}\left[\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t).\varphi\}\right]^2.$$

A2. Proof of Theorem 3

Let $\varphi : \Xi_t \to \mathbb{R}^d$ and $\overline{\varphi} = \varphi - \mathbb{E}_{\pi_t}(\varphi) = \varphi - \mathbb{E}_{\pi_t^m}(\varphi)$ for a given $t \geq 0$. To simplify notations it is assumed that $d = 1$, but the adaptation to the general case is straightforward. All quantities related to the 'marginalized' particle filter are distinguished by the$m$-suffix. For instance, $\mathcal{E}_t^m(\varphi)$ stands for function $\xi_t \mapsto \mathbb{E}_{k_t^m(\xi_t,\cdot)}\{\upsilon_t^m(\cdot)\varphi(\cdot)\}$, in agreement with the definition of $\mathcal{E}_t(\varphi)$ in (10). In this respect, the marginal weight function $\upsilon_t^m(\cdot)$ is $\widetilde{\pi}_t^m(\cdot)/\pi_t^m(\cdot)$, and if we define the 'conditional' weight function $\upsilon_t^c(\lambda_t|\xi_t) = \pi_t^c(\lambda_t|\xi_t)/\widetilde{\pi}_t^c(\lambda_t|\xi_t)$, we have the identity

$$\upsilon_t(\theta_t) = \upsilon_t^m(\xi_t)\upsilon_t^c(\lambda_t|\xi_t).$$

It follows from (12) that

$$\mathbb{E}_{\pi_{t-1}^c}\{\mathcal{E}_t(\varphi)\} = \mathbb{E}_{k_t^m}\{\upsilon_t^m.\varphi.\mathbb{E}_{\widetilde{\pi}_t^c}(\upsilon_t^c)\} = \mathcal{E}_t^m(\varphi),$$

since $\mathbb{E}_{\widetilde{\pi}_t^c}(\upsilon_t^c) = 1$, and by induction we show similarly for $k \leq t$ that

$$\mathbb{E}_{\pi_k^c}\{\mathcal{E}_{k+1:t}(\varphi)\} = \mathcal{E}_{k+1:t}^m(\varphi).$$

Hence, for $k \leq t$,

$$\begin{aligned}
\mathbb{E}_{\widetilde{\pi}_k}\left[\upsilon_k.\mathcal{E}_{k+1:t}\{\overline{\varphi}\}\right]^2 &= \mathbb{E}_{\widetilde{\pi}_k^m}\left[\upsilon_k^m.\mathbb{E}_{\widetilde{\pi}_k^c}\{\upsilon_k^c\mathcal{E}_{k+1:t}\overline{\varphi}\}^2\right], \\
&\geq \mathbb{E}_{\widetilde{\pi}_k^m}\left[\upsilon_k^m.\{\mathbb{E}_{\widetilde{\pi}_k^c}\upsilon_k^c\mathcal{E}_{k+1:t}\overline{\varphi}\}^2\right], \\
&\geq \mathbb{E}_{\widetilde{\pi}_k^m}\left[\upsilon_k^m.\mathcal{E}_{k+1:t}^m\{\overline{\varphi}\}\right]^2,
\end{aligned}$$

by Jensen inequality. From the closed form (9) of $V_t(\varphi)$ we deduce the inequality $V_t^m(\varphi) \leq V_t(\varphi)$ for the case when the selection step follows the multinomial scheme. Alternatively, if the selection step consists in residual resampling, let $\underline{\varphi} = \varphi - \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\varphi\}/\mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\}$, then

$$\begin{aligned}
R_t(\varphi) - R_t^m(\varphi) &= \mathbb{E}_{\widetilde{\pi}_t}\{r(\upsilon_t)\underline{\varphi}^2\} - \mathbb{E}_{\widetilde{\pi}_t^m}\{r(\upsilon_t^m)\underline{\varphi}^2\} + \frac{\{\mathbb{E}_{\widetilde{\pi}_t^m} r(\upsilon_t^m)\underline{\varphi}\}^2}{\mathbb{E}_{\widetilde{\pi}_t^m} r(\upsilon_t^m)} \\
&\geq \mathbb{E}_{\widetilde{\pi}_t^m}\left[\{\mathbb{E}_{\widetilde{\pi}_t^c} r(\upsilon_t) - r(\upsilon_t^m)\}\underline{\varphi}^2\right],
\end{aligned}$$

and since $\mathbb{E}_{\widetilde{\pi}_t^c}(\upsilon_t) = \upsilon_t^m$, we have $\mathbb{E}_{\widetilde{\pi}_t^c}\lfloor\upsilon_t\rfloor \leq \lfloor\upsilon_t^m\rfloor$, hence $\mathbb{E}_{\widetilde{\pi}_t^c} r(\upsilon_t) \geq r(\upsilon_t^m)$, and consequently $R_t(\varphi) \geq R_t^m(\varphi)$ for any φ. It follows from (11) that the desired inequality is also verified in the residual case.

A3. Regularity conditions and proof of Theorem 4

Let $\pi_0(\theta)$ denote the prior density and $p(y_{1:t}|\theta)$ the likelihood of t first observations, so that through Bayes formula,

$$\pi_t(\theta) = \pi(\theta|y_{1:t}) \propto \pi_0(\theta)p(y_{1:t}|\theta).$$

Let $l_t(\theta) = \log p(y_{1:t}|\theta)$. The following assumptions holds almost surely.

(a) The maximum $\widehat{\theta}_t$ of $l_t(\theta)$ exists and converges as $t \to +\infty$ to θ_0 such that $\pi_0(\theta_0) > 0$.

(b) The matrix

$$\Sigma_t = -\left\{\frac{1}{t}\frac{\partial^2 l_t(\theta)}{\partial\theta\partial\theta'}\right\}^{-1}$$

is positive definite and converges to $I(\theta_0)$, the Fisher information matrix at θ_0.

(c) There exists $\Delta > 0$ such that

$$0 < \delta < \Delta \Rightarrow \limsup_{t\to+\infty}\left[\frac{1}{t}\sup_{\|\theta-\widehat{\theta}_t\|>\delta}\{l_t(\theta) - l_t(\widehat{\theta}_t)\}\right] < 0.$$

(d) Functions $\pi_0(\theta)$ and $l_t(\theta)$ are six times continuously differentiable, the partial derivatives of order six of $l_t(\theta)/t$ are bounded over any compact set $\Theta' \subset \Theta$, and the bound does not depend on t and the observations.

(e) $\varphi : \Theta \to \mathbb{R}^d$ is six times continuously differentiable, $\varphi'(\theta_0) \neq 0$.

For convenience, we start with the one-dimensional case ($p = 1$). The Laplace approximation of an integral (see for instance Tierney et al., 1989) is

$$\int \psi(\theta) \exp\{-th(\theta)\}\, d\theta =$$
$$(2\pi/t)^{1/2}\sigma \exp\{-t\widehat{h}\} \left[\widehat{\psi} + \frac{1}{2}\{\sigma^2\widehat{\psi}'' - \sigma^4\widehat{\psi}'\widehat{h}''' + \frac{5}{12}\sigma^6\widehat{\psi}\widehat{h}''' - \frac{1}{4}\sigma^4\widehat{b}\widehat{h}^{iv}\}t^{-1} + O(t^{-2})\right]$$

where hats on ψ, h and their derivatives indicate evaluation at the point which minimizes h, and $\sigma = -1/\widehat{h}''$. This approximation remains valid for a function h_t depending on t, provided that the fluctuations of h_t or its derivatives can be controlled in some way. Conditions (c) and (d) above allow for instance for applying this approximation to functions $h_t^1(\theta) = -l_t(\theta)/t$ and $h_t^2(\theta) = -2l_t(\theta)/t$, see Schervish (1995, p. 446) for technical details. It is necessary however to assume that $\psi(\theta_0) \neq 0$, so that ψ is either strictly positive or strictly negative at least in a neighbourhood of θ_0. Since $V_t^{sis}(\varphi) = V_t^{sis}(\varphi + \lambda)$ for any $\lambda \in \mathbb{R}$ we assume without loss of generality that $\varphi(\theta_0) \neq 0$. $V_t^{sis}(\varphi)$ equals

$$\frac{\int \psi_1(\theta)p(y_{1:t}|\theta)^2 d\theta - 2E_{\pi_t}(\varphi)\int \psi_2(\theta)p(y_{1:t}|\theta)^2 d\theta + \{E_{\pi_t}(\varphi)\}^2 \int \psi_3(\theta)p(y_{1:t}|\theta)^2 d\theta}{\left\{\int \pi(\theta)p(y_{1:t}|\theta)d\theta\right\}^2}, \quad (26)$$

where $\psi_1 = \pi_0(\theta)^2\varphi(\theta)^2/\widetilde{\pi}_0(\theta)$, $\psi_2 = \pi_0(\theta)^2\varphi(\theta)/\widetilde{\pi}_0(\theta)$ and $\psi_3 = \pi_0(\theta)^2/\widetilde{\pi}_0(\theta)$. Combining the appropriate Laplace approximations, we get that

$$\begin{aligned} V_t^{sis}(\varphi) &= \frac{1}{2}(\pi\Sigma_t)^{-1/2}t^{1/2}\frac{\{\psi_1(\widehat{\theta}_t) - 2\mathbb{E}_{\pi_t}(\varphi)\psi_2(\widehat{\theta}_t) + \mathbb{E}_{\pi_t}(\varphi)^2\psi_3(\widehat{\theta}_t) + At^{-1} + O(t^{-2})\}}{\{\pi_0(\widehat{\theta}_t) + Bt^{-1} + O(t^{-2})\}^2} \\ &= \frac{1}{2}(\pi\Sigma_t)^{-1/2}t^{1/2}\left[\{\varphi(\widehat{\theta}_t) - \mathbb{E}_{\pi_t}(\varphi)\}^2\{1 - 2B/\pi_0(\widehat{\theta}_t)t^{-1}\} + A/\pi_0(\widehat{\theta}_t)^2t^{-1} + O(t^{-2})\right] \end{aligned}$$

where A is the sum of $O(t^{-1})$ terms corresponding to the three Laplace expansions of the numerator, and B is the $O(t^{-1})$ term of the denominator. Since $\varphi(\widehat{\theta}_t) - E_{\pi_t}(\varphi) = O(t^{-1})$, $\Sigma_t = I(\theta_0) + O(t^{-1})$ and $\psi(\widehat{\theta}_t) = \psi(\theta_0) + O(t^{-1})$ for any continuous function ψ, we get through appropriate derivation that

$$V_t^{sis}(\varphi) = \frac{I(\theta_0)\varphi'(\theta_0)^2}{4\pi_0(\theta_0)^2}t^{-1/2} + O(t^{-3/2}).$$

Derivations in multi-dimensional cases are much the same, except that notations are more cumbersome. When $p > 1$, the factor $t^{-1/2}$ in Laplace expansion is replaced by $t^{-p/2}$, so that in the ratio (26) we get a factor $t^{p/2}$, and since the $t^{p/2}$ order cancels as in one-dimensional case, the actual rate of divergence is $t^{p/2-1}$, and this completes the first part of the proof.

In the specific case of the sequential importance sampling, $q_t(\theta, \cdot) = \delta_\theta$ and $\widetilde{\pi}_t = \pi_{t-1}$, and according to (9),

$$V_t(\varphi) = V_t^{sis}(\varphi) + \sum_{k=1}^{t} \mathbb{E}_{\pi_{k-1}}\left[\frac{\pi_t}{\pi_{k-1}}\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\}\right]^2, \quad (27)$$

then through a direct adaptation of expansions above we obtain a divergence rate for $V_t(\varphi)$ of order $(\sum_{k=0}^{t}(t-k)^{p/2-1}) = O(t^{p/2})$. For the residual case, it follows from (11) and (27)

that

$$V_t^r(\varphi) = V_t^{sis}(\varphi) + \sum_{k=0}^{t-1} R_k[\frac{\pi_t}{\pi_k}\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\}].$$

The difficulty in this case is that the non continuous function$r(\cdot)$ intervenes in the expression of $R_k(\cdot)$, see (8). It is clear however that the Laplace expansion generalizes to cases where regularity conditions for the likelihood and other functions are fulfilled only locally around θ_0. The additional assumption that $\pi_t(\theta_0)/\pi_{t-1}(\theta_0)$ is not an integer for any $t > 0$ allows $r(\upsilon_t)$ for being six times continuously differentiable in a neighbourhood around θ_0, and therefore makes it possible to expand the terms of sum above, which leads to a rate of divergence of order $O(t^{p/2})$ in the same way as in multinomial case.

A4. Proof of Theorem 5

As a preliminary, we state without proof the following inequality. Let $\varphi, \psi : \mathbb{R} \to \mathbb{R}$ such that $\varphi \geq 0$, $\sup \psi \geq 0$ and $\inf \psi \leq 0$, then

$$\Delta(\varphi\psi) \leq \sup \varphi . \Delta\psi. \tag{28}$$

Due to particular cancellations, the weight function $\upsilon_t(x_{1:t})$ only depends on x_{t-1} and x_t in the state space case,

$$\upsilon_t(x_{1:t}) = \upsilon_t(x_{t-1}, x_t) \propto \frac{f(y_t|x_t)g(x_t|x_{t-1})}{q_t(x_t|x_{t-1})}. \tag{29}$$

Straightforward consequences of this expression are the identities,

$$\pi_t(x_t|x_{t-1}) = \frac{q_t(x_t|x_{t-1})\upsilon_t(x_{t-1}, x_t)}{\int q_t(x|x_{t-1})\upsilon_t(x_{t-1}, x)\, dx}, \tag{30}$$

$$\pi_{t+1}(x_{t+1}|x_k) = \frac{\int \pi_t(x_t|x_k)q_{t+1}(x_{t+1}|x_t)\upsilon_{t+1}(x_t, x_{t+1})\, dx_t}{\int \pi_t(x_t|x_k)q_{t+1}(x|x_t)\upsilon_{t+1}(x_t, x)\, dx_t dx}, \tag{31}$$

for $k < t+1$, where $\pi_t(x_t|x_k)$ denotes the conditional posterior density of x_t given x_k and the t first observations, that is $\pi_t(x_t|x_k) = \pi(x_t|x_k, y_{1:t}) = \pi(x_t|x_k, y_{k+1:t})$. We start by proving some useful lemmas.

LEMMA 5. *The conditional posterior density $\pi_t(x_t|x_k)$, $k < t$, defines a Markov transition from x_k to x_t whose contraction coefficient is less than or equal to $(1 - C^{-2})^{t-k}$.*

PROOF. For $x_k, x_k', x_{k+1} \in \mathcal{X}$, $k < t$,

$$\frac{\pi_t(x_{k+1}|x_k)}{\pi_t(x_{k+1}|x_k')} = \frac{g(x_{k+1}|x_k)p(y_{k+1:t}|x_k')}{g(x_{k+1}|x_k')p(y_{k+1:t}|x_k)} \leq C^2,$$

since $g(x_{k+1}|x_k) \leq Cg(x_k|x_k')$ and

$$\begin{aligned} p(y_{k+1:t}|x_k') &= \int g(x_{k+1}|x_k')p(y_{k+1:t}|x_{k+1})\, dx_{k+1}, \\ &\leq C \int g(x_{k+1}|x_k)p(y_{k+1:t}|x_{k+1})\, dx_{k+1}. \end{aligned}$$

Therefore the contraction coefficients of Markov transitions $\pi_t(x_{k+1}|x_k)$ and $\pi_t(x_t|x_k)$ are less than or equal to, respectively, $(1 - C^{-2})$ and $(1 - C^{-2})^{t-k}$.

LEMMA 6. *Let λ a density on $\mathcal{X}$ and $h(x|x')$ a conditional density defining a Markov transition on $\mathcal{X}$, then for any $x' \in \mathcal{X}$, $y \in \mathcal{Y}$,*

$$\frac{\int f(y|x)h(x|x')\,dx}{\mathbb{E}_{\lambda(x'')}\{\int f(y|x)h(x|x'')\,dx\}} \leq 1 + \rho_h C_f,$$

where ρ_h is the contraction coefficient of $h(\cdot|\cdot)$, and $C_f = \overline{f}/\underline{f} - 1$.

PROOF. It follows from the definition of ρ_h, see (14), that, for $x', x'' \in \mathcal{X}$,

$$|\int f(y|x)h(x|x')\,dx - \int f(y|x)h(x|x'')\,dx| \leq \rho_h(\overline{f} - \underline{f}),$$

and therefore

$$\sup_{x' \in \mathcal{X}}\{\int f(y|x)h(x|x')\,dx\} \leq \mathbb{E}_{\lambda(x'')}\{\int f(y|x)h(x|x'')\,dx\} + \rho_h(\overline{f} - \underline{f}),$$

so that

$$\begin{aligned}\frac{\sup_{x' \in \mathcal{X}}\{\int f(y|x)h(x|x')\,dx\}}{\mathbb{E}_{\lambda(x'')}\{\int f(y|x)h(x|x'')\,dx\}} &\leq 1 + \rho_h \frac{(\overline{f} - \underline{f})}{\mathbb{E}_{\lambda(x'')}\{\int f(y|x)h(x|x'')\,dx\}}, \\ &\leq 1 + \rho_h \left(\frac{\overline{f}}{\underline{f}} - 1\right).\end{aligned}$$

LEMMA 7. *Let $\rho = 1 - C^{-1}$, $\rho_2 = 1 - C^{-2}$, then for $k < t$,*

$$\Delta\mathcal{E}_{k+1:t}\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\} \leq \prod_{i=1}^{t-k}(1 + \rho\rho_2^{i-1}C_f)\rho_2^{t-k}\Delta\varphi.$$

PROOF. Let $\overline{\varphi} = \varphi - \mathbb{E}_{\pi_t}(\varphi)$. Note the arguments of $\mathcal{E}_{k+1:t}(\overline{\varphi})$ are $x_{1:k}$ in general, but in the case considered in §3.3 it only depends on x_k and is therefore treated as a function $\mathcal{X} \to \mathcal{X}$. For the sake of clarity we treat the case $k = t - 2$ but the reasoning easily generalizes. The following decomposition is deduced from the identity (30).

$$\begin{aligned}\mathcal{E}_{t-1:t}(\overline{\varphi})(x_{t-2}) &= \mathbb{E}_{q_{t-1}(x_{t-1}|x_{t-2})}\{\upsilon_{t-1}(x_{t-2}, x_{t-1})\mathcal{E}_t(\overline{\varphi})(x_{t-1})\}, \\ &= \mathbb{E}_{q_{t-1}(x_{t-1}|x_{t-2})}\{\upsilon_{t-1}(x_{t-2}, x_{t-1})\}\mathbb{E}_{\pi_{t-1}(x_{t-1}|x_{t-2})}\{\mathcal{E}_t(\overline{\varphi})(x_{t-1})\}.\end{aligned}$$

It comes from (29) that the first term verifies

$$\mathbb{E}_{q_{t-1}(x_{t-1}|x_{t-2})}\{\upsilon_{t-1}(x_{t-2}, x_{t-1})\} \propto \int f(y_{t-1}|x_{t-1})g(x_{t-1}|x_{t-2})\,dx_{t-1},$$

where the proportionality constant can be retrieved by remarking that the expectation of this term over π_{t-2} equals one, and therefore,

$$\begin{aligned}\mathbb{E}_{q_{t-1}(x_{t-1}|x_{t-2})}\{\upsilon_{t-1}(x_{t-2}, x_{t-1})\} &= \frac{\int f(y_{t-1}|x_{t-1})g(x_{t-1}|x_{t-2})\,dx_{t-1}}{\mathbb{E}_{\pi_{t-2}(x_{t-2})}\{\int f(y_{t-1}|x_{t-1})g(x_{t-1}|x_{t-2})\,dx_{t-1}\}} \\ &\leq 1 + \rho C_f\end{aligned}$$

according to Lemma 6. Note $\pi_{t-2}(x_{t-2})$ denotes the π_{t-2}-marginal density of x_{t-2}. It follows from the decomposition above and the inequality in (28) that

$$\Delta\mathcal{E}_{t-1:t}(\overline{\varphi}) \leq (1+\rho C_f)\Delta\psi$$

where ψ is the function

$$\begin{aligned}\psi(x_{t-2}) &= \mathbb{E}_{\pi_{t-1}(x_{t-1}|x_{t-2})}\{\mathcal{E}_t(\overline{\varphi})(x_{t-1})\} \\ &= \mathbb{E}_{\pi_{t-1}(x_{t-1}|x_{t-2})}\left[\mathbb{E}_{q_t(x_t|x_{t-1})}\{\upsilon_t(x_{t-1},x_t)\overline{\varphi}(x_t)\}\right].\end{aligned}$$

Note that ψ does take positive and negative values, since the expectation of $\mathcal{E}_{t-1:t}(\overline{\varphi})$ over π_{t-2} is null. We now decompose ψ in the same way as for $\mathcal{E}_{t-1:t}(\overline{\varphi})$, that is

$$\psi(x_{t-2}) = \mathbb{E}_{\pi_{t-1}(x_{t-1}|x_{t-2})}\left[\mathbb{E}_{q_t(x_t|x_{t-1})}\{\upsilon_t(x_{t-1},x_t)\}\right] E_{\pi_t(x_t|x_{t-2})}\{\overline{\varphi}(x_t)\},$$

by consequence of identity (31). The expectation of the first term over $\pi_{t-1}(x_{t-2})$ equals one, so that

$$\begin{aligned}&\mathbb{E}_{\pi_{t-1}(x_{t-1}|x_{t-2})}\left[\mathbb{E}_{q_t(x_t|x_{t-1})}\{\upsilon_t(x_{t-1},x_t)\}\right] \\ &= \frac{\int \pi_{t-1}(x_{t-1}|x_{t-2})f(y_t|x_t)g(x_t|x_{t-1})\,dx_{t-1}\,dx_t}{\mathbb{E}_{\pi_{t-1}(x_{t-2})}\left\{\int \pi_{t-1}(x_{t-1}|x_{t-2})f(y_t|x_t)g(x_t|x_{t-1})\,dx_{t-1}\,dx_t\right\}}, \\ &\leq 1+\rho\rho_2 C_f,\end{aligned}$$

according to Lemma 6. Resorting again to inequality (28), we get

$$\Delta\psi \leq (1+\rho\rho_2 C_f)\rho_2^2\Delta\varphi,$$

which leads to the desired inequality, and this completes the proof of Lemma 7.

To conclude the proof of Theorem 5, remark that $\mathbb{E}_{\widetilde{\pi}_k}(\upsilon_k) = 1$, therefore

$$\begin{aligned}\upsilon_k(x_{k-1},x_k) &= \frac{f(y_k|x_k)g(x_k|x_{k-1})/q_k(x_k|x_{k-1})}{\mathbb{E}_{\widetilde{\pi}_k(x_{1:k})}\left\{f(y_k|x_k)g(x_k|x_{k-1})/q_k(x_k|x_{k-1})\right\}}, \\ &\leq C^2,\end{aligned}$$

and since the expectation of the function $\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}$ over π_k is null, the function $\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}$ is ensured to take positive and negative values, so that

$$\sup_{x_k\in\mathcal{X}} \left|\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}(x_k)\right| \leq \Delta\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}$$

and finally,

$$\begin{aligned}\mathbb{E}_{\widetilde{\pi}_k}\left[\upsilon_k^2\mathcal{E}_{k+1:t}\left\{\varphi - \mathbb{E}_{\pi_t}(\varphi)\right\}^2\right] &\leq C^4\prod_{i=1}^{t-k}(1+\rho\rho_2^{i-1}C_f)^2\rho_2^{2(t-k)}(\Delta\varphi)^2, \\ &\leq C^4\exp(2\rho C_f\sum_{i=1}^{t-k}\rho_2^{i-1})\rho_2^{2(t-k)}(\Delta\varphi)^2, \\ &\leq C^4\exp\{2\rho C_f/(1-\rho_2)\}\rho_2^{2(t-k)}(\Delta\varphi)^2.\end{aligned}$$

It follows from (9) that $V_t(\varphi)$ is bounded from above by a convergent series.

References

Andrieu, C. and Doucet, A. (2002). Particle filtering for partially observed Gaussian state space models. *J. R. Statist. Soc. B*, (to appear).

Billingsley, P. (1995). *Probability and measure*. 3rd ed., Wiley.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar, Sonar Navigation*, 146(1):2–7.

Chen, R. and Liu, J. (2000). Mixture Kalman filters. *J. R. Statist. Soc. B*, 62:493–508.

Chopin, N. (2001). Sequential inference and state number determination for discrete state-space models through particle filtering. *CREST Working Paper*, 2001-34.

Chopin, N. (2002). A sequential particle filter for static models. *Biometrika*, 89:539–552.

Del Moral, P. and Guionnet, A. (1999). Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Prob.*, 9:275–297.

Del Moral, P. and Guionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré*, 37(2):155–194.

Dobrushin (1956). Central limit theorem for non-stationary Markov chains I, II. *Theory Prob. Appl.*, 1:65–80, 329–383.

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

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, 10(3):197–208.

Gilks, W. R. and Berzuini, C. (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. B*, 63:127–146.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Comm., Radar, Signal Proc.*, 140(2):107–113.

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

Legland, F. Oudjane, N. (2001). Stability and approximation of nonlinear filters using the Hilbert metric, and application to particle filters. *Technical report, INRIA*.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Am. Statist. Ass.*, 93:1032–1044.

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

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

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, M., DeGroot, K., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3*. Oxford University Press.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximmations to expectations and variances of non-positive functions. *J. Am. Statist. Ass.*, 84:710–716.

Whitley (1994). A genetic algorithm tutorial. *Stat. Comp.*, 4:65–85.