Working Paper

Recursive Monte Carlo filters algorithms and theoretical analysis

Author(s):
Künsch, Hansruedi

Publication Date:
2003

Permanent Link:

https://doi.org/10.3929/ethz-a-004467791 →

Recursive Monte Carlo Filters: Algorithms and Theoretical Analysis

by

Hans R. Künsch

Research Report No. 112
January 2003

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Recursive Monte Carlo Filters: Algorithms and Theoretical Analysis

Hans R. Künsch

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

January 2003

Abstract

Recursive Monte Carlo filters, also called particle filters, are a powerful tool to perform the computations in general state space models. We discuss and compare the accept-reject version with the more common sampling importance resampling version of the algorithm. In particular, we show how auxiliary variable methods and stratification can be used in the accept-reject version, and we compare different resampling techniques. In a second part, we show laws of large numbers and a central limit theorem for these Monte Carlo filters by simple induction arguments that need only weak conditions. We also show that under stronger conditions the required sample size is independent of the length of the observed series.

AMS 2000 subject classifications. Primary 62M09; secondary 60G35, 60J22, 65C05.

Key words and phrases. State space models, hidden Markov models, filtering and smoothing, particle filters, auxiliary variables, sampling importance resampling, central limit theorem.

1 State Space and Hidden Markov Models

A *general state space model* consists of an unobserved state sequence (X_t) and an observation sequence (Y_t) with the following properties:

State evolution: $X_0, X_1, X_2, \ldots$ is a Markov chain with $X_0 \sim a_0(x)d\mu(x)$ and

$$X_t \mid X_{t-1} = x_{t-1} \;\sim\; a_t(x_{t-1}, x)d\mu(x)$$

Generation of observations: Conditionally on (X_t), the Y_t's are independent and Y_t depends on X_t only with

$$Y_t \mid X_t = x_t \;\sim\; b_t(x_t, y)d\nu(y).$$

If X_t is discrete, it is usually called a *Hidden Markov Model*.

These models occur in a variety of applications. Linear state space models are equivalent to ARMA models, see e.g. Hannan and Deistler (1989), and have become popular under the name of structural models, see e.g. Harvey (1989). Nonlinear state space models occur in finance (stochastic volatility, see e.g. Shephard (1996)), in various fields of engineering (speech, tracking and control problems), in biology (ion channels, DNA and protein sequences) and in geophysics (rainfall at a network of stations, data assimilation). A more detailed survey with many references is given in Künsch (2001), and Doucet et al. (2001) contains many examples, mainly from the engineering field.

In order to apply these models, two kinds of problems have to be solved: Inference about the states based on a stretch of observed values $y_s^t = (y_u, s \le u \le t)$ for a given model, i.e. a_t and b_t known (this is called prediction, filtering and smoothing), and inference about unknown parameters in a_t, b_t. From a statistical point of view, the latter problem is maybe of greater interest, but fast and reliable algorithms for the former are a prerequisite for computing maximum likelihood or Bayesian estimators. The reason for this is briefly mentioned in subsection 2.1. This paper is therefore entirely devoted to algorithms for filtering, prediction and smoothing.

Section 2 recalls the basic recursions for filtering, prediction and smoothing. Section 3 discusses algorithmic aspects of sequential Monte Carlo methods to implement these recursions. Most algorithms in the literature, beginning with the pioneering paper by Gordon, Salmond and Smith (1993), use the sampling importance resampling idea of Rubin (1988). An exception is Hürzeler and Künsch (1998) who use the accept-reject method instead. Here we show how some ideas like stratification and an auxiliary variable method of Pitt and Shephard (1999) can be adapted to rejection sampling, and we give a new result on the performance of a systematic resampling method of Carpenter et al. (1999). In addition, we hope that our view of classifying and comparing approaches is useful.

Section 4 presents results on the convergence of the method as the number of Monte Carlo replicates tends to infinity. We discuss both laws of large numbers and a central limit theorem. Recently, many similar results have been published, see e.g. Del Moral and Miclo (2000), Crisan (2001) and Le Gland and Oudjane (2001). The distinctive features of our presentation here are the weakness of conditions, the use of the total variation distance to measure the difference between the approximate and the true filter density and the simplicity of the techniques used. We basically show that most results follow by induction, in accordance with the recursive nature of the algorithm. The complications that occur are due to a counterintuitive property of Bayes formula, see Lemma 3.6 ii) in Künsch (2001). As a consequence, although one can obtain consistency with very few conditions on the model, the required sample size seems to grow exponentially with the number of time steps. For results that guarantee that the required sample size is independent of the number of time steps (or grows at most logarithmically), one has to use induction over several time steps and this requires very strong conditions on the dynamics of the states.

2 Filtering and smoothing recursions

Let $f_{t|s}(x_t \mid y_1^s)$ be the conditional density of X_t given $Y_1^s = y_1^s$. We distinguish the cases $s < t$ (prediction), $s = t$ (filtering) and $s > t$ (smoothing). The dependence structure of a state space model can be represented by the following graph

$$\begin{array}{ccccccccc} \ldots & \to & X_{t-1} & \to & X_t & \to & X_{t+1} & \to & \ldots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \ldots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \ldots \end{array}$$

From this, various conditional independence properties follow which are used together with the law of total probability and Bayes' theorem to derive recursions for the filter, prediction and smoothing densities. These are well known, see e.g. Künsch (2001), Section 3.3, and we state them without proofs.

The most important result is the following recursion for the filter density:

Propagation: From the filter density we obtain the one step ahead prediction density:

$$f_{t|t-1}(x_t \mid y_1^{t-1}) = \int f_{t-1|t-1}(x \mid y_1^{t-1}) a_t(x, x_t) d\mu(x) \tag{1}$$

Update: From the one step ahead prediction density, we obtain the filter density one time step later

$$f_{t|t}(x_t \mid y_1^t) = \frac{f_{t|t-1}(x_t \mid y_1^{t-1}) b_t(x_t, y_t)}{\int f_{t|t-1}(x \mid y_1^{t-1}) b_t(x, y_t) d\mu(x)} \propto f_{t|t-1}(x_t \mid y_1^{t-1}) b_t(x_t, y_t). \tag{2}$$

In parts of the literature, e.g. in Del Moral and Miclo (2000), Y_t depends on X_{t-1} and not on X_t. Then the filter density is in our setup the prediction density which should be kept in mind when comparing formulae.

2.1 Prediction of observations and likelihood

The denominator in the update step (2) is the conditional density of Y_t given Y_1^{t-1}:

$$p(y_t \mid y_1^{t-1}) = \int f_{t|t-1}(x \mid y_1^{t-1}) b_t(x, y_t) d\mu(x). \tag{3}$$

If $f_{t|t-1}$ is available, we thus can obtain the likelihood from

$$p(y_1^T) = \prod_{t=1}^{T} p(y_t \mid y_1^{t-1}).$$

A different representation of the likelihood is

$$p(y_1^T) = \int a_0(x_0) \prod_{t=1}^{T} a_t(x_{t-1}, x_t) b_t(x_t, y_t) \prod_{t=0}^{T} d\mu(x_t).$$

From this the likelihood ratio can be expressed as an expectation with respect to the smoothing distribution, see e.g. Hürzeler and Künsch (2001).

2.2 Smoothing

The filter densities can also be used for the smoothing problem since conditional on y_1^T, $(X_T, X_{T-1}, \ldots, X_0)$ is an inhomogeneous Markov chain with starting density $f_{T|T}$ and backward transition densities

$$p(x_t \mid x_{t+1}, y_1^T) = p(x_t \mid x_{t+1}, y_1^t) \propto a_{t+1}(x_t, x_{t+1}) f_{t|t}(x_t \mid y_1^t). \tag{4}$$

This is also the basis for the forward-filtering-backward-sampling algorithm, see Frühwirth-Schnatter (1994), equation (20). From (3), we can derive in particular a backward recursion for $f_{t|T}$.

2.3 Recursive filtering in operator notation

A compact notation for the filter recursion which will be useful later on is

$$f_{t|t}(. \mid y_1^t) = B(A_t^* f_{t-1|t-1}(. \mid y_1^{t-1}), b_t(., y_t)). \tag{5}$$

Here

$$A_t^* f(x) = \int f(x') a_t(x', x) d\mu(x')$$

is the Markov transition operator, and

$$B(f, b)(x) = \frac{f(x)b(x)}{\int f(x)b(x)d\mu(x)}$$

is the Bayes operator that assigns the posterior to a prior f and a likelihood b. The operators A_t^* and $B(., b)$ map the space of densities into itself, but they can be extended to the space of probability distributions.

2.4 Implementation of recursions

If X_t is discrete with M possible values, integrals are sums and the recursions need $O(TM^2)$ operations. In a linear Gaussian state space model, all $f_{t|s}$ are Gaussian, and their means and variances are computed with the Kalman filter and smoother.

In practically all other cases, the recursions are difficult to compute. Analytical approximations like the extended Kalman filter are not satisfactory, and numerical integration is problematic in high dimensions. Much current interest focuses on Monte Carlo methods. Standard Markov chain Monte Carlo can be used, but they lack a recursive implementation. There has been a large interest in recursvie Monte Carlo methods in recent years, see e.g. Doucet et al. (2001).

3 Algorithms for recursive Monte Carlo filtering

The following is the key observation: $A_t^* f$ is difficult to compute, but easy to sample from if we can sample from f and $a_t(x, .)$. This allows us to generate recursively a sequence of samples ("particles") $(x_{j,t}; j = 1, \ldots N, t = 0, 1, \ldots)$ with approximate distribution $f_{t|t}$ as follows: If $(x_{j,t-1})$ is available, we can replace

$$A_t^* f_{t-1|t-1}(x \mid y_1^{t-1}) = \int f_{t-1|t-1}(x \mid y_1^{t-1}) a_t(x, x_t) d\mu(x)$$

by

$$\frac{1}{N} \sum_{j=1}^{N} a_t(x_{j,t-1}, x).$$

Therefore we sample $(x_{j,t})$ from the distribution with density

$$f_{t|t}^N(. \mid y_1^t) \propto b_t(., y_t) \frac{1}{N} \sum_{j=1}^{N} a_t(x_{j,t-1}, .). \tag{6}$$

In this section we discuss methods to sample from such a density. We simplify the notation somewhat and write the target density as

$$f^N(x) \propto f_u^N(x) = b(x) \sum_{j=1}^{N} a(j, x) \tag{7}$$

(subscript u for unnormalized). We will call b the likelihood and $N^{-1}\sum_j a(j,x)$ the prior. In the filtering context, the prior is the approximate prediction density. For later use, we also introduce

$$\beta_j = \int a(j,x)b(x)d\mu(x)$$

which is in the filtering context equal to the conditional density of Y_t given $X_{t-1} = x_{j,t-1}$). We assume that we have good methods to generate samples from $a(j,.)$ for any j. The methods we discuss fall into two categories: accept-reject and importance sampling with an additional resampling step.

3.1 Accept-reject methods

The accept-reject method for sampling from the density (6) produces values X according to a proposal ρ and if $X = x$ accepts it with probability

$$\pi(x) = \frac{f_u^N(x)}{M\rho(x)}. \tag{8}$$

Here M is an upper bound for the ratio $f_u^N(x)/\rho(x)$

$$M \geq \sup_x \frac{f_u^N(x)}{\rho(x)}.$$

The most obvious proposal $\rho(x)$ is the prior, that is

$$\rho(x) = \frac{1}{N}\sum_{j=1}^{N} a(j,x). \tag{9}$$

Then the evaluation of the acceptance probabilities $\pi(x)$ is easy as long as b is bounded. In order to sample from (**??**), we first choose an index J uniformly from $\{1,\dots N\}$, and given $J = j$, we sample X from $a(j,x)$. Note that in this case, the densities $a(j,x)$ need not be available in analytic form, we only have to be able to sample from them. This is of interest in discretely observed diffusion models.

The average acceptance probability of this algorithm is $\int \rho(x)\pi(x)d\mu(x) = \sum_j \beta_j/M$. In particular, if ρ is the prior and if we use the smallest value of M, it is equal to

$$\frac{\sum_{j=1}^N \beta_j}{N \sup_x b(x)}.$$

This is low if the likelihood is more informative (concentrated) than the prior, or if the likelihood and the prior are in conflict. We discuss here some modifications and tricks that can alleviate this problem in some situations.

3.1.1 The mixture index as auxiliary variable

Other proposal distributions than the prediction density can of course lead to higher acceptance rates, but usually it is difficult to compute a good upper bound M, and the evaluation of the acceptance probability $\pi(x)$ is complicated due to the sum over j. A way to avoid at least the last problem is based on an idea by Pitt and Shephard (1999). Namely we can generate first an index J according to a distribution (τ_j) and given $J = j$

a variable X according to a density $\rho(j,x)$. We then accept the generated pair (j,x) with probability

$$\pi(j,x) = \frac{a(j,x)b(x)}{M\tau_j\rho(j,x)} \tag{10}$$

where now

$$M \geq \sup_{j,x} \frac{a(j,x)b(x)}{\tau_j\rho(j,x)}.$$

If the pair is accepted, we simply discard j and keep x, otherwise we generate a new pair. Because the accepted pairs (J,X) have distribution

$$\frac{a(j,x)b(x)}{\sum_j \beta_j},$$

the marginal distribution of X is the target (6). If we take $\tau_j = 1/N$ and $\rho(j,x) = a(j,x)$ we obtain the usual algorithm discussed before, but one will try to increase the acceptance rate by other choices.

Because j runs over a finite set, we will usually take

$$M = \max_j \frac{M_j}{\tau_j} \quad \text{where} \quad M_j \geq \sup_x \frac{a(j,x)b(x)}{\rho(j,x)}.$$

Lemma 1 *For a given choice of densities $\rho(j,x)$ and bounds M_j, the average acceptance probability is maximal for $\tau_j \propto M_j$.*

Proof: The average acceptance probability is

$$\sum_j \int \pi(j,x)\tau_j\rho(j,x)\mu(dx) = \frac{1}{M}\sum_j \beta_j = \sum_j \beta_j \left(\max_k \frac{M_k}{\tau_k}\right)^{-1}.$$

Clearly

$$\max_k \frac{M_k}{\tau_k} = \sum_j \tau_j \max_k \frac{M_k}{\tau_k} \geq \sum_j M_j,$$

with equality iff M_k/τ_k is constant. □

If $\rho(j,x) = a(j,x)$, the optimal τ_j's are thus constant. This is somewhat surprising since one could conjecture that it is better to give higher probability to those indices j for which the mass of $a(j,x)$ is close to $\arg\sup b(x)$.

The crucial point in implementing this algorithm is the choice of the densities $\rho(j,.)$. We see from the proof of Lemma 1 that for a high acceptance probability all M_j's should be small, i.e. each $\rho(j,x)$ should be a good proposal distribution for the density $a(j,x)b(x)/\beta_j$. Ideally, we would choose that density itself. But then M_j must be close to the normalizing constant β_j which typically is not available in closed form. A more practical approach chooses a parametric family $(\rho(\theta,x))$ where we have available tight upper bounds

$$M(j,\theta) \geq \sup_x \frac{a(j,x)b(x)}{\rho(\theta,x)}.$$

We then optimize over θ, that is

$$\rho(j,x) = \rho(\theta_j,x) \quad \text{where} \quad \theta_j \approx \arg\min_\theta M(j,\theta).$$

Note that it is not necessary to find the optimal θ exactly, but $M(j,\theta)$ should be a true upper bound. By choosing the family $(\rho(\theta,x))$ such that it contains all densities $a(j,x)$, we can make sure that the acceptance probability is at least as high as with the usual algorithm.

As an example, consider the case where $a(j,.)$ is the normal density with mean m_j and variance σ^2 and where b is the likelihood of a $\mathcal{N}(0,\exp(x))$ random variable Y:

$$b(x) = b(x,y) = \exp\left(-\frac{x}{2} - \frac{y^2}{2}\exp(-x)\right).$$

This corresponds to the simplest stochastic volatility model, see e.g. Shephard (1996). If we take as $\rho(\theta,.)$ the normal density with mean θ and variance σ^2, we can compute the supremum of

$$\log\frac{a(j,x)b(x)}{\rho(\theta,x)} = -\frac{x}{2} - \frac{\theta-m_j}{\sigma^2}x - \frac{y^2}{2}\exp(-x) - \frac{m_j^2-\theta^2}{2\sigma^2}$$

over x. It is equal to

$$\frac{\sigma^2}{2}\delta^2 + m_j\delta - (\frac{1}{2}+\delta)(1+\log y^2) + (\frac{1}{2}+\delta)\log(1+2\delta)$$

provided $\delta = (\theta - m_j)/\sigma^2 \geq -1/2$ (otherwise the function is unbounded above). Minimizing this expression with respect to δ subject to $\delta \geq -1/2$ leads to a non-linear equation which has no closed form solution. Using $\log(1+2\delta) \leq 2\delta$, we obtain a quadratic upper bound which is minimized by

$$\theta_j = m_j + \frac{\sigma^2}{2}\max\left(-1, \frac{2}{4+\sigma^2}(\log y^2 - m_j)\right).$$

This choice of θ_j may be slightly suboptimal, but because the bound is sharp for $\theta = m_j$, i.e. $\delta = 0$, we still can guarantee a higher acceptance probability than with the usual method. In practice, the gain can be dramatic if $|y|$ is small.

The above choice of θ_j is somewhat different from the suggestion

$$\theta_j = m_j + \frac{\sigma^2}{2}\left(y^2\exp(-m_j) - 1\right)$$

in Shephard and Pitt (2001), p. 285. In addition, also the choices for τ_j differ.

3.1.2 Stratification

Besides reducing the acceptance rate, we can also try to reduce the variance by using a more systematic sampling. This idea has received much attention in the sampling importance sampling context, see Section 3.2.1 below and the references given there. We have not seen this idea in the accept-reject context. Consider the estimation of

$$m(\psi) = \int f^N(x)\psi(x)d\mu(x) = \frac{\sum_j \beta_j m_j(\psi)}{\sum_j \beta_j}$$

where

$$m_j(\psi) = \frac{\int \psi(x)a(j,x)b(x)d\mu(x)}{\beta_j}$$

and ψ is a bounded "test function". If (X_i) is an i.i.d. sample from f^N, the estimator

$$\hat{m}(\psi) = \frac{\sum_{j=1}^N \psi(X_j)}{N}$$

has variance

$$\frac{1}{N}\sigma^2(\psi) = \frac{1}{N}\frac{\sum_j \int (\psi(x) - m(\psi))^2 a(j,x) b(x) d\mu(x)}{\sum_j \beta_j}.$$

A method to reduce this variance replaces the random selection of an index J by a more systematic procedure. Namely we can propose simultaneously N values, one each from the density $a(j,x)$, and decide whether to accept each of them independently. We repeat the procedure until the total of accepted values is at least N. If we need exactly N values, we can select them at random. We therefore consider the estimator

$$\tilde{m}(\psi) = \frac{\sum_{i=1}^T \sum_{j=1}^N \psi(X_{ij}) 1_{[U_{ij} < b(X_{ij})]}}{\sum_{i=1}^T \sum_{j=1}^N 1_{[U_{ij} < b(X_{ij})]}}$$

where $(X_{ij}, U_{ij}; 1 \le j \le N, i = 1, 2, \ldots)$ are independent random variables with $X_{ij} \sim a(j, .)$, U_{ij} uniform on $(0, \sup b(x))$, and T is the smallest integer such that the denominator is at least N.

In order to compute the variance of $\tilde{m}(\psi)$ approximately, we use

$$\tilde{m}(\psi) - m(\psi) = \frac{\sum_{i=1}^T \sum_{j=1}^N (\psi(X_{ij}) - m(\psi)) 1_{[U_{ij} < b(X_{ij})]}}{\sum_{i=1}^T \sum_{j=1}^N 1_{[U_{ij} < b(X_{ij})]}}.$$

For simplicity, we assume that $\sup b(x) = 1$. Then, by Wald's identity, the denominator has expected value

$$\mathbf{E}[T] \sum_{j=1}^N \beta_j.$$

In particular, the expected number of random variables that have to be generated is essentially the same as with the basic i.i.d. rejection sampling. Similarly, the numerator has expectation zero and variance

$$\mathbf{E}(T) \sum_{j=1}^N \mathrm{Var}((\psi(X_{1j}) - m(\psi)) 1_{[U_{1j} < b(X_{1j})]}) = \mathbf{E}(T)(\sigma^2(\psi) \sum_j \beta_j - \sum_j \beta_j^2 (m_j(\psi) - m(\psi))^2).$$

Assuming the denominator to be approximately constant and equal to N (which is reasonable if the expected number of accepted values in each round of proposals is small), we obtain the approximation

$$\mathbf{E}[\hat{m}(\psi)] \approx m(\psi), \quad \mathrm{Var}(\hat{m}(\psi)) \approx \frac{1}{N}\left(\sigma(\psi)^2 - \frac{\sum_j \beta_j^2 (m_j(\psi) - m(\psi))^2}{\sum \beta_j}\right).$$

The second term thus quantifies the gain of the method.

3.2 Sampling importance resampling.

This method generates $(z_k; 1 \le k \le R)$ according to some proposal ρ and selects from these a sample of size N with inclusion probabilities

$$\pi(z_k) \propto \frac{b(z_k) \sum_{j=1}^N a(j, z_k)}{\rho(z_k)}. \tag{11}$$

The resampling need not to be made at random. We will discuss below alternative methods with reduced variability.

The standard proposal is again the prior (**??**), leading to the original proposal in Gordon, Salmond and Smith (1993). Situations with a low acceptance rate in rejection sampling typically also have heavily unequal sampling probabilities $\pi(z_k)$, thus leading to many ties in the final sample. Choosing R much bigger than N reduces the number of ties, but at the expense of longer computations. Note that rejection sampling is an automatic way of choosing R such that all ties are avoided. In cases where all $a(j,.)$'s have their main mass in a region where the likelihood is flat and small, sampling importance resampling can be much faster than rejection sampling and still give approximately equal weights to all values. However, this can be misleading since it simply means that no value was proposed in the region where the likelihood is large. It does not guarantee that the target f^N has negligible mass there. A more detailed comparison between rejection and importance sampling in general can be found in Section 3.3.3 of Robert and Casella (1999).

Most of the ideas discussed in connection with rejection sampling can also be used here. The idea of Pitt and Shephard (1999) to include explicitely an index J was originally developed for this case. It proposes a sample (j_k, z_k) of size R with distribution $\tau_j \rho(j,x)$ and then selects a sample of size N with inclusion probabilities

$$\pi(z_k, j_k) \propto \frac{b(z_k)a(j_k, z_k)}{\tau_{j_k}\rho(j_k, z_k)}.$$

In contrast to rejection sampling, combining $\rho(j,.) = a(j,.)$ with unequal τ_j's is a promising idea here. For instance, we can take τ_j to be proportional to $b(m_j)$ where m_j is the mean or the median of $a(j,.)$. If all $a(j,.)$'s have a small spread (relative to the scale at which b varies), then most $\pi(z_k, j_k)$'s will be approximately equal, and therefore $R = N$ is sufficient.

3.2.1 Stratification and the effect of resampling

Reducing the variance by stratification is important and often easy to implement. It can be used both for the proposal and the resampling. If we use the prior (**??**) as the proposal, we can take $R = mN$ and then obtain (z_k) by generating m values from $a(j,.)$ for each j. If we use a proposal distribution $\tau_j \rho(j,x)$, we would similarly like to generate R_j values from $a(j,.)$ where the R_j's are integers close to $R\tau_j$ with $\sum R_j = R$. The same problem occurs in the resampling step where we would like the multiplicities N_k of z_k in the resample to be close to $N\pi(z_k)$. Resampling randomly with replacement implies that (N_k) will be multinomial$(N, (\pi(z_k)))$, but we will discuss here other possibilities and compare them with random sampling.

We require that $\sum N_k \equiv N$ and that resampling is unbiased, that is

$$\mathbf{E}[N_j \mid z_1, \ldots z_r] = N\pi(z_j).$$

Then the estimator

$$\hat{m}(\psi) = \frac{1}{N}\sum_{j=1}^{R} \psi(Z_j)N_j.$$

has the same expected value as the usual importance sampling estimator

$$\tilde{m}(\psi) = \sum_{j=1}^{R} \psi(Z_j)\pi(Z_j).$$

Its variance can be written as

$$\mathrm{Var}[\hat{m}(\psi)] = \mathrm{Var}[\sum_{j=1}^{R} \psi(Z_j)\pi(Z_j)] + \frac{1}{N^2}\,\mathbf{E}[\sum_{i,j} \psi(Z_i)\psi(Z_j)C_R(i,j)].$$

where $C_R(i,j)$ is the conditional covariance of N_i and N_j. The first term is the variance of the usual importance sampling estimator and the second term is the additional variability due to the resampling step. Without resampling, the recursive filter would quickly degenerate, that is practically all the weights would be given to very few values. Resampling is necessary in order to split the particles with large weights into several independent ones and to kill some of the particles with very small weights. Nevertheless, we should try to minimize the additional variability. Because it is not known in advance which functions ψ will be of interest, we presumably should consider the supremum over all (bounded) test functions ψ.

With multinomial N_j's, we have

$$\sum_{i,j} \psi(z_i)\psi(z_j)C_R(i,j) = N(\sum_i \psi(z_i)^2\pi(z_i) - (\sum_i \psi(z_i)\pi(z_i))^2) \le N \sup \psi(x)^2.$$

Hence, resampling randomly with replacement can guarantee that the effect of resampling disappears asymptotically.

Several methods have been proposed which reduce the (conditional) variances $C_R(i,i)$. Residual sampling (Liu and Chen, 1998) takes

$$N_i = [N\pi(z_i)] + N_i', \quad (N_i') \sim \text{multinomial}(N', (\pi'(z_i)))$$

where $[x]$ denotes the integer part of x and

$$N' = N - \sum_i [N\pi(z_i)], \quad \pi'(z_i) = \frac{N\pi(z_i) - [N\pi(z_i)]}{N'}.$$

This reduces $\sum_{i,j} \psi(z_i)\psi(z_j)C_R(i,j)$ by the factor N'/N. Intuitively, we expect the remainder $N\pi(z_i) - [N\pi(z_i)]$ to be uniform on $(0,1)$, leading to an average reduction by a factor of two.

The variance $C_R(k,k)$ is minimal iff N_k is equal to one of the two integers closest to $N\pi(z_k)$. This can be achieved by the following algorithm, see Whitley (1994) and Carpenter et al. (1999):

$$N_{j_k} = \left| [N \sum_{i=1}^{k-1} \pi(z_{j_i}) + U, N \sum_{i=1}^{k} \pi(z_{j_i}) + U) \cap \{1, 2, \ldots, N\} \right|, \tag{12}$$

where $(j_1, j_2, \ldots, j_R)$ is a random permutation of $(1, 2, \ldots R)$, U is uniform on $[0,1)$, and the absolute value of a finite set denotes the number of elements in this set.

But by minimizing $C_R(i,i)$, we usually introduce strong dependence between different N_j's, and the effects of this are hard to control. We know that $|C_R(i,j)| \le 1/4$, but the bound

$$\sum_{i,j} \psi(z_i)\psi(z_j)C_R(i,j) \le \frac{N^2}{4} \sup \psi(x)^2$$

contains no useful information because it does not even allow to conclude that the additional uncertainty due to resampling disappears asymptotically. Because both $\psi(Z_i)$ and $C_R(i,j)$ can be either positive or negative, I do not see how one could obtain a better worst case bound. But the following Lemma supports the conjecture that on average the algorithm (9) will behave well.

Lemma 2 *For arbitrary probabilities (π_i) and arbitrary N, consider the random variables*

$$N_j = \left| [N \sum_{i=1}^{j-1} \pi_i + U, N \sum_{i=1}^{j} \pi_i + U) \cap \{1, 2, \ldots, N\} \right|,$$

where U is uniform on $(0,1)$ (This is the algorithm (9) without the additional permutation). Then for any $j < k$, $\mathrm{Cov}(N_j, N_k)$ depends only on $r_l = N\pi_j \bmod 1$, $r_u = N\pi_k \bmod 1$ and $r_m = N \sum_{i=j+1}^{k-1} \pi_i \bmod 1$, an explicit expression being given in the proof. Moreover, the average of this covariance with respect to the uniform distribution on $(0,1)$ for r_m is zero for all values r_l and r_u.

Proof: Because shifting a uniform random variable modulo 1 does not change the distribution, we may assume that $j = 1$. Moreover, it is clear that only the fractional parts r_l, r_m, r_u matter. If we put $M_j = N_j - [N\pi_j]$ and $M_k = N_j - [N\pi_k]$, we obtain therefore

$$\begin{aligned} \mathbf{E}\, M_j M_k &= \mathbf{P}[U \in (0, r_l) \cap (r_l + r_m - 1, r_l + r_m + r_u - 1)] \\ &+ \mathbf{P}[U \in (0, r_l) \cap (r_l + r_m - 2, r_l + r_m + r_u - 2)]. \end{aligned}$$

It is easy to evaluate the right hand side by distinguishing different cases:

$$\mathbf{E}\, M_j M_k = \begin{cases} (r_l + r_m + r_u - 1)^+ & (r_l + r_m \le 1, r_m + r_u \le 1) \\ r_u & (r_l + r_m > 1, r_m + r_u \le 1) \\ r_l & (r_l + r_m \le 1, r_m + r_u > 1) \\ 1 - r_m & (r_l + r_m > 1, r_m + r_u > 1, r_l + r_m + r_u \le 2) \\ r_l + r_u - 1 & (r_l + r_m + r_u > 2) \end{cases}$$

It is also easy to show that by integrating over $r_m \in (0,1)$, we obtain $r_l r_u$ in all cases. □

We expect that randomizing the order of the values will make the r_m approximately uniform. Therefore it seems wise to use always randomization since it is computationally cheap.

Crisan et al. (1999) have proposed a different algorithm that also minimizes the $C_R(i,i)$, see also Crisan (2001). It builds first a binary tree such that the leaves correspond to the values z_j. To each node one attaches the value N times the sum of the probabilities of those leaves that originate from this node. Then one lets N particles propagate down the tree from the root such that the number of particles at each node differs at most by one from the value of the node and such that the expected number of particles at each node is equal to its value. For this algorithm, it can be shown that $C_R(i,j) \le 0$ for all $i \ne j$. Together with $C_R(i,i) \le 1/4$ this implies

$$\sum_{i,j} \psi(z_i)\psi(z_j) C_R(i,j) \le \frac{N}{2} \sup \psi(x)^2,$$

see Crisan (2001), p. 31. It is therefore guaranteed that this algorithm reduces the additional variance due to resampling by a factor of at least two compared to multinomial sampling.

3.2.2 Sampling importance resampling as recursive prediction

Sampling importance resampling can also be considered as a natural way for generating a recursive approximation of the prediction densities $f_{t|t-1}$ by particles $(z_{j,t}; j = 1, \ldots, N, t = 1, 2, \ldots)$. By the same rationale as for the filter, this means to sample $(z_{j,t})$ from the density

$$f^N_{t|t-1}(. \mid y_1^{t-1}) = \frac{\sum_{j=1}^N b_{t-1}(z_{j,t-1}) a_t(z_{j,t-1}, .)}{\sum_{j=1}^N b_{t-1}(z_{j,t-1})}.$$

The right hand side is a mixture density with weights proportional to $b_{t-1}(z_{j,t-1})$. Sampling from this mixture is thus the same as resampling the particles at time $t-1$ followed by a propagation according to the state transition density.

3.3 Computation of the likelihood

Combining (**??**) and (5), we see that

$$p(y_t \mid y_1^{t-1}) \approx \sum_{j=1}^N \int \frac{1}{N} a_t(x_{j,t-1}, x) b_t(x, y_t) d\mu(x)$$

which is in the short notation of this section equal to $\sum \beta_j / N$. If we use $\tau_j \rho(j, x)$ as our proposal, then the usual importance sampling estimator of $p(y_t \mid y_1^{t-1})$ is

$$\widehat{p}(y_t \mid y_1^{t-1}) = \frac{1}{NR} \sum_{k=1}^R \frac{b(z_k) a(j_k, z_k)}{\tau_{j_k} \rho(j_k, z_k)}.$$

3.4 Monte Carlo backward smoothing

There is a similar recursive simulation method that generates samples from the conditional distribution of X_0^T given $Y_1^T = y_1^T$. At time T, we use the recursive filter sample: $x^{sm}_{j,T} = x_{j,T}$. We then proceed backward in time, using (3) together with an approximation of $f_{t|t}$. In order to avoid problems with discreteness, we recommend to use (5) as in Hürzeler and Künsch (1998) instead of replacing $f_{t|t}$ by the empirical distribution of the particles at time t as in Godsill, Doucet and West (2000). This means that we generate $x^{sm}_{j,t}$ from $x^{sm}_{j,t+1}$ and $(x_{i,t-1})$ by simulating from the density proportional to

$$a_{t+1}(x, x^{sm}_{j,t+1}) b_t(x, y_t) \frac{1}{N} \sum_{i=1}^N a_t(x_{i,t-1}, x). \tag{13}$$

(At time $t = 0$ we will use the density proportional to $a_1(x, x^{sm}_{j,1}) a_0(x)$.) Clearly, this has the same structure as (6) and so the same methods as discussed before apply in principle. However, we need one value from the density (**??**) for each j and thus sampling importance resampling does not seem to be useful here. For the same reason, care is needed when using the mixture index as auxiliary variable. Since sampling from (τ_i) typically involves computing the partial sums of the τ_i's, one should use the same distribution (τ_i) for all j. Then the computational cost of the approach is $O(TN)$ and thus at least comparable to a standard MCMC method. The main disadvantage of this approach is that we have to store all the filter samples.

4 Theoretical properties

In this section, we analyze the convergence of the approximation $f_{t|t}^N$ to the true filtering density $f_{t|t}$. We will hold the observations y_1^t fixed and drop them from the notation. In particular, we do not make any assumption about how the observations were obtained. The true filtering densities $f_{t|t}$ are then deterministic, but the approximations $f_{t|t}^N$ are still random since their computation involves random sampling. All expectations and probabilities in this section concern the randomness of the Monte Carlo methods, and not the randomness of the state space model. We assume throughout that X_t takes its values in a complete, separable metric space equiped with the Borel σ-field, and we denote the metric on this state space by $d(.,.)$.

The operator notation for recursive Monte Carlo filters introduced in Section 2.3 will be used extensively. In addition we denote by $E_N(f)$ the empirical distribution of a sample of size N from f. Then the approximate filter density is

$$f_{t|t}^N = B(A_t^* E_N(f_{t-1|t-1}^N), b_t(., y_t))$$

and it has to be compared with

$$f_{t|t} = B(A_t^* f_{t-1|t-1}, b_t(., y_t)),$$

see (5) and (4). We present two approaches for showing convergence of $f_{t|t}^N$ to $f_{t|t}$ as $N \to \infty$. We measure the error by the the L_1-distance between densities, see e.g. Devroye (1987), Chapter 1, which can be written in several equivalent forms

$$\begin{aligned} \|f-g\|_1 &= \int |f(x)-g(x)| d\mu(x) = 2\int (f(x)-g(x))^+ d\mu(x) \\ &= 2\sup_B |P_f[B] - P_g[B]| = 2\int (f(x) - \min(f(x), g(x))) d\mu(x). \end{aligned} \tag{14}$$

(x^+ denotes the positive part of x). Clearly, if $\left\| f_{t|t}^N - f_{t|t} \right\|_1$ converges to zero in probability or almost surely, then for any bounded function ψ on the state space the law of large number holds:

$$\frac{1}{N}\sum_{j=1}^N \psi(x_{j,t}) \longrightarrow \int \psi(x) f_{t|t}(x) d\mu(x),$$

in probability or almost surely. In the last section, we show the corresponding central limit theorem.

4.1 Stepwise error propagation

The obvious first attempt to show convergence uses the following decomposition

$$\begin{aligned} f_{t|t}^N - f_{t|t} &= B(A_t^* E_N(f_{t-1|t-1}^N), b_t) - B(A_t^* f_{t-1|t-1}^N, b_t) \\ &+ B(A_t^* f_{t-1|t-1}^N, b_t) - B(A_t^* f_{t-1|t-1}, b_t). \end{aligned} \tag{15}$$

The first term is the error due to sampling at time $t-1$ (propagated once) and the second term is the propagation of the error at time $t-1$. For a recursive inequality for $\left\| f_{t|t}^N - f_{t|t} \right\|_1$, we have to study the Lipschitz-continuity of Bayes and Markov operators with respect to the L_1-distance and to control the sampling error.

The continuity of Markov operators is well known, see Dobrushin (1956), Section 3.

Lemma 3 *We have*

$$\|A^*f - A^*g\|_1 \leq \rho(A^*)\ \|f - g\|_1$$

where

$$\rho(A^*) = \frac{1}{2}\sup_{x,x'} \|a(x,\cdot) - a(x',\cdot)\|_1 \leq 1.$$

Note that for compact state space the Markov operator is typically contracting.

The continuity of Bayes' formula with respect to the prior is more problematic. We have, see Künsch (2001), Lemma 3.6 i),

Lemma 4

$$\|B(f,b) - B(g,b)\|_1 \leq \beta(f,b)\,\|f - g\|_1$$

where

$$\beta(f,b) = \frac{\sup_x b(x)}{\int b(x)f(x)d\mu(x)} \in \left(1, \frac{\sup_x b(x)}{\inf_x b(x)}\right].$$

The difficulty is that this bound cannot be improved in general. Lemma 3.6 ii) from Künsch (2001) shows that the Bayes operator is not contracting for any f at least for some "directions" g.

Finally, we have the following bound on sampling errors:

Lemma 5 *If $x \to a(x,.)$ is continuous with respect to the L_1-norm, then under i.i.d. sampling from g*

$$\mathbf{P}[\|A^*E_N(g) - A^*g\|_1 > \varepsilon] \overset{N\to\infty}{\longrightarrow} 0$$

exponentially fast in N for any $\varepsilon > 0$. The convergence is uniform for all g such that $\int_K g d\mu \geq 1 - \varepsilon/6$ for some fixed compact set K.

Proof: The proof follows closely the arguments in Devroye (1987), Chapter 3. We denote by μ_N the empirical distribution $E_N(g)$ and by μ_g the distribution $g(x)d\mu(x)$.

Let $\varepsilon > 0$ be given. Choose a compact K such that $\mu_g(K) \geq 1 - \varepsilon/6$. Next, choose δ such that $\|a(x,.) - a(x',.)\|_1 \leq \varepsilon/6$ for all $x, x' \in K$ with $d(x,x') \leq \delta$. Then choose a partition $\{B_1, \ldots B_J\}$ of K such that each B_j has diameter at most δ and choose a point z_j in B_j for each j. Finally put $B_0 = K^c$. Then

$$\begin{aligned}
&\left\| \frac{1}{N}\sum_{i=1}^N a(x_i,.) - \sum_{j=1}^J \mu_N(B_j)a(z_j,.) \right\|_1 \\
&= \int \left| \frac{1}{N}\sum_{i=1}^N 1_{B_0}(x_i)a(x_i,x) + \sum_{j=1}^J \frac{1}{N}\sum_{i=1}^N 1_{B_j}(x_i)(a(x_i,x) - a(z_j,x)) \right| d\mu(x) \\
&\leq \mu_N(B_0) + \sum_{j=1}^J \frac{1}{N}\sum_{i=1}^N 1_{B_j}(x_i) \int |a(x_i,x) - a(z_j,x)|\, d\mu(x) \\
&\leq |\mu_N(B_0) - \mu_g(B_0)| + \frac{\varepsilon}{3}.
\end{aligned}$$

Similarly we obtain

$$\left\| \int a(x,.)g(x)d\mu(x) - \sum_{j=1}^J \mu_g(B_j)a(z_j,.) \right\|_1 \leq \frac{\varepsilon}{3}.$$

Finally

$$\left\| \sum_{j=1}^{J} \mu_g(B_j) a(z_j, .) - \sum_{j=1}^{J} \mu_N(B_j) a(z_j, .) \right\|_1 \leq \sum_{j=1}^{J} |\mu_N(B_j) - \mu_g(B_j)| .$$

Taking these three inequalities together, we obtain

$$\|A^* E_N(g) - A^* g\|_1 \leq \frac{2\varepsilon}{3} + \sum_{j=0}^{J} |\mu_N(B_j) - \mu_g(B_j)| .$$

Hence, the large deviation estimate for the multinomial distribution

$$\mathbf{P}[\sum_{j=0}^{J} |\mu_N(B_j) - \mu_g(B_j)| > \frac{\varepsilon}{3}] \leq 2^{J+2} \exp(-N\varepsilon^2/18)$$

(Devroye (1987), Theorem 3.2) implies

$$\mathbf{P}[\|A^* E_N(g) - A^* g\|_1 > \varepsilon] \leq 2^{J+2} \exp(-N\varepsilon^2/18).$$

From this, the Lemma follows (note that once K is fixed, J depends only on the transition kernel a and not on g). □

Theorem 1 *If $x \to a_t(x, .)$ is continuous and if for all t, all x and all y*

$$0 < b_t(x, y) \leq C(t, y) < \infty,$$

then for all t and all y_1^t

$$\left\| f_{t|t}^N - f_{t|t} \right\|_1 \longrightarrow 0$$

in probability as $N \to \infty$.

Proof: The proof proceeds by induction on t. For $t = 0$ there is nothing to prove because $f_{0|0}^N = f_{0|0} = a_0$. From Lemmas 3 and 4 it follows that

$$\left\| B(A_t^* f_{t-1|t-1}^N, b_t) - B(A_t^* f_{t-1|t-1}, b_t) \right\|_1 \leq \frac{C(t, y_t)}{p(y_t \mid y_1^{t-1})} \left\| f_{t-1|t-1}^N - f_{t-1|t-1} \right\|_1 \leq \varepsilon$$

if

$$\left\| f_{t-1|t-1}^N - f_{t-1|t-1} \right\|_1 \leq \varepsilon \frac{p(y_t \mid y_1^{t-1})}{C(t, y_t)} =: \delta. \tag{16}$$

By the induction assumption, there is an N_1 such that for $N > N_1$ (**??**) holds with probability at least $1 - \eta$.

In order to bound the first term in (10), some care is needed when applying the bounds provided by Lemmas 4 and 5 with $f_{t-1|t-1}^N$ which is random. We have to show that when (**??**) holds, we can obtain bounds which depend only on $f_{t-1|t-1}$. Note first that

$$\int b_t(x, y_t)(A_t^* f_{t-1|t-1}^N(x) - A_t^* f_{t-1|t-1}(x)) d\mu(x) \geq -\frac{1}{2} C(t, y_t) \left\| f_{t-1|t-1}^N - f_{t-1|t-1} \right\|_1 .$$

Hence if (**??**) is satisfied,

$$\int b_t(x, y_t) A_t^* f_{t-1|t-1}^N(x) d\mu(x) \geq (1 - \varepsilon/2) p(y_t \mid y_1^{t-1}) \geq \frac{1}{2} p(y_t \mid y_1^{t-1})$$

and therefore by Lemma 4 also

$$\left\| B(A_t^* E_N(f_{t-1|t-1}^N), b_t) - B(A_t^* f_{t-1|t-1}^N, b_t) \right\|_1 \leq \frac{2C(t, y_t)}{p(y_t \mid y_1^{t-1})} \left\| A_t^* E_N(f_{t-1|t-1}^N) - A_t^* f_{t-1|t-1}^N \right\|_1 .$$

Next we observe that if K is compact such that $\int_K f_{t|t} d\mu \geq 1 - \delta/2$ and if (**??**) holds, then $\int_K f_{t|t}^N d\mu \geq 1 - \delta$. Therefore by Lemma 5 we can find N_2 such that for $N > N_2$

$$\left\| A_t^* E_N'(f_{t-1|t-1}^N) - A_t^* f_{t-1|t-1}^N \right\|_1 \leq 6\delta \tag{17}$$

holds with probability at least $1 - \eta$. Collecting all the bounds shows that for $N > \max(N_1, N_2)$

$$\left\| f_{t|t}^N - f_{t|t} \right\|_1 \leq 13\,\varepsilon$$

with probability at least $1 - 2\eta$. □

The conditions of this theorem are weak. However, the arguments in the proof require $\left\| f_{t-1|t-1}^N - f_{t-1|t-1} \right\|_1$ to be smaller than $\left\| f_{t|t}^N - f_{t|t} \right\|_1$. This means that the required sample size N grows with t. It is easy to see that in general N has to grow exponentially with t, and thus from a practical point of view, the theorem is not of great use. Strengthening the assumptions like for instance assuming a compact state space does not help because by Lemma 3.6 ii) from Künsch (2001), the Bayes operator is expanding. Hence for a more useful result, we need a different approach which is provided in the next section.

4.1.1 Error propagation with sampling importance resampling

The results so far assumed that the Monte Carlo filter uses i.i.d. samples of $f_{t|t}^N$ which means using the accept-reject method (with or without auxiliary variables). It does not cover sampling importance resampling. In order to extend the results above, we need to adapt Lemma 5 to the different sampling method when g has the form $g = B(h, b)$. We only sketch the main idea. Let (x_j) be the sample from h, let $\pi(x_j) = b(x_j)/\sum b(x_k)$ be the resampling weights and let (N_j) the multiplicities of the resample. Then

$$\begin{aligned} \left\| \frac{1}{N} \sum_{i=1}^N N_i a(x_i, .) - A^* B(h, b)(.) \right\|_1 \quad &\leq \quad \left\| \frac{1}{N} \sum_{i=1}^N (N_i - N\pi(x_i)) a(x_i, .) \right\|_1 \\ &+ \quad \left\| \frac{\sum_{j=1}^N b(x_j) a(x_j, .)}{\sum_{j=1}^N b(x_j)} - \frac{\int h(x) b(x) a(x, .) d\mu(x)}{\int h(x) b(x) d\mu(x)} \right\|_1 . \end{aligned}$$

The first term can be handled similarly as in Lemma 5 if we assume b to be continuous. For the second term we use the same decomposition as in formula (21) of Del Moral and Miclo (2000). It can then be bounded by the sum of

$$\frac{1}{\int h(x) b(x) d\mu(x)} \left\| \frac{1}{N} \sum_{j=1}^N b(x_j) a(x_j, .) - \int h(x) b(x) a(x, .) d\mu(x) \right\|_1$$

which again can be handled similarly as in Lemma 5, and

$$\frac{1}{\int h(x) b(x) d\mu(x)} \left| \frac{1}{N} \sum_{j=1}^N b(x_j) - \int h(x) b(x) d\mu(x) \right| \left\| \frac{\sum_{j=1}^N b(x_j) a(x_j, .)}{\sum_{j=1}^N b(x_j)} \right\|_1 .$$

This can be handled by Hoeffding's inequality because the L_1-norm is equal to one.

4.2 Analysis based on considering several steps

Clearly, we can look at error propagation over more than one time step. If we define

$$K_{s,t}(f) = K_{s+1,t}(A^*_{s+1}B(f, b_s)) \ (s < t), \quad K_{t,t}(f) = B(f, b_t),$$

then for any $s < t$ $f_{t|t} = K_{s+1,t}(A^*_{s+1}f_{s|s})$ and hence

$$f^N_{t|t} - f_{t|t} = \sum_{r=s+1}^{t} (K_{r,t}(A^*_r E_N(f^N_{r-1|r-1})) - K_{r,t}(A^*_r f^N_{r-1|r-1})) + K_{s+1,t}(A^*_{s+1}f^N_{s|s}) - K_{s+1,t}(A^*_{s+1}f_{s|s}).$$

Here, the last difference is the error at time s propagated over $t - s$ steps. The other differences are the errors due to sampling at time $r - 1$, propagated over $t - r$ steps.

This is only useful if we can give a bound on the error propagated over k steps which is better than the sum over k single steps. It is possible because an alternative way to get from $f_{s|s}$ to $f_{t|t}$ is to apply first the Bayes operator once with likelihood equal to the conditional density of y^t_{s+1} given x_s, followed by $t-s$ Markov operators for the conditional transitions from x_r to x_{r+1} given y^t_{r+1}. The contractivity of the Markov operators can then beat the expansion of the Bayes operator. It requires however a uniform nontrivial upper bound for the contraction coefficient of the conditional chain given y^t_{r+1}, and for this we need the following condition

$$C_a := \sup_{t,x,x',x''} \frac{a_t(x, x'')}{a_t(x', x'')} < \infty. \tag{18}$$

It implies that there are densities h_t such that for all x and x'

$$C_a^{-1} \, h_t(x) \le a_t(x', x) \le C_a \, h_t(x) \tag{19}$$

(choose as h_t any $a_t(x'', .)$). Conversely, the inequalities (12) imply (11) with C_a replaced by C_a^2.

The conditions (11) or (12) on a_t are reasonable when the state space is compact although they are slightly stronger than uniform ergodicity. Using (**??**), we see that the lower bound $a_t(x', x) \ge C_a^{-1}h(x)$ of (12) alone implies $\rho(A^*_t) \le 1 - C_a^{-1}$ and thus also uniform ergodicity. Conditions (11) or (12) include even some examples with unbounded state space. For instance, (11) holds for the model

$$X_t = g(X_{t-1}) + V_t$$

if g is bounded and V_t has a density whose logarithm is uniformly Lipschitz continuous. This is satisfied for most heavy-tailed distributions, but not for the Gaussian. For Gaussian V_t, (11) is false, but the lower bound in (12) holds. We thus have an example of an uniformly ergodic chain that we cannot treat with our arguments.

Concerning b_t, there is an almost minimal condition, namely

$$0 < \int a_{t-1}(x_{t-1}, x) b_t(x, y_t) d\mu(x) < \infty \tag{20}$$

for all t, y_t and some x_{t-1}. Under condition (11), this holds then automatically for any x_{t-1}. Some arguments become much simpler however, if we replace (13) by

$$C_b := \sup_{t,x,x',y} \frac{b_t(x, y)}{b_t(x', y)} < \infty. \tag{21}$$

The following Lemma shows that under condition (11) the error propagated over several steps decreases exponentially. Many versions of this exponential forgetting of the initial conditions of the filter have appeared in the literature, see e.g. Del Moral and Miclo (2000), Del Moral and Guionnet (2001), Le Gland and Oudjane (2001) and the references given there. We use the version of Künsch (2001), Theorem 3.9.

Lemma 6 *Assume condition (11) and condition (13) for all t, y_t and some x_{t-1}. Then $\rho(A_t) \leq 1 - 1/C_a < 1$ and for any two densities f and g and any $s < t$ we have*

$$\left\| K_{s+1,t}(A^*_{s+1} f) - K_{s+1,t}(A^*_{s+1} g) \right\|_1 \leq C_a (1 - C_a^{-2})^{t-s} \left\| f - g \right\|_1 .$$

Theorem 2 *Assume that the transition densities a_t are the same for all t, that they are continuous in the L_1-norm and satisfy (11), and that (**??**) holds. Then to any $\varepsilon > 0$ there are constants c_1 and c_2 such that for all t and all N*

$$\mathbf{P}[\left\| f^N_{t|t} - f_{t|t} \right\|_1 > \varepsilon] \leq c_1 \exp(-c_2 N).$$

Proof: Because a_t and thus also A^*_t are the same for all t, we drop the time index during this proof. Let $\varepsilon > 0$ be given. Choose k such that

$$2C_a(1 - C_a^{-2})^k \leq \varepsilon.$$

Assume first that $k < t$. Because the L_1-distance between densities is at most 2, we obtain in this case from the decomposition (4.2) with $s = t - k$ and Lemmas **??** and 4

$$\begin{aligned}
&\left\| f^N_{t|t} - f_{t|t} \right\|_1 \\
&\quad \leq \quad C_a \sum_{r=t-k}^{t-1} (1 - C_a^{-2})^{t-r-1} \left\| B(A^* E_N(f^N_{r|r}), b_{r+1}) - B(A^* f^N_{r|r}), b_{r+1}) \right\|_1 + \varepsilon \\
&\quad \leq \quad C_b C_a \sum_{r=t-k}^{t-1} (1 - C_a^{-2})^{t-r-1} \left\| A^* E_N(f^N_{r|r}) - A^* f^N_{r|r} \right\|_1 + \varepsilon.
\end{aligned}$$

If $k > t$ we obtain a similar result by considering the decomposition (4.2) with $s = 0$. (Because $f^N_{0|0} = f_{0|0} = a_0$ the ε at the end is then absent). Hence if

$$\sup_{t-k \leq r < t} \left\| A^* E_N(f^N_{r|r}) - A^* f^N_{r|r} \right\|_1 \leq \varepsilon \tag{22}$$

holds, then by the formula for a geometric series

$$\left\| f^N_{t|t} - f_{t|t} \right\|_1 \leq (C_a^3 C_b + 1)\varepsilon.$$

We are now going to bound the probability that (**??**) occurs. Note that ε and thus also k are fixed. Because of Lemma 5, all we need to show is that the set of distributions $(f^N_{r|r} d\mu)$ is tight. By the definition of $f^N_{r|r}$ and by the conditions (11) and (**??**), we have

$$f^N_{r|r}(x) = \frac{\sum_{j=1}^N a(x_{j,r-1}, x) b_r(x, y_r)}{\sum_{j=1}^N \int a(x_{j,r-1}, x) b_r(x, y_r) d\mu(x)} \leq C_b C_a a(x', x)$$

for an arbitrary fixed x'. Clearly this implies the desired tightness. □

The important feature of the above theorem is that the same N works for all times t. By Bonferroni's inequality we obtain

$$\mathbf{P}[\sup_{t\leq T}\left\|f^N_{t|t}-f_{t|t}\right\|_1>\varepsilon]\leq Tc_1\exp(-c_2N).$$

Hence it is sufficient to let N increase logarithmically with the length of the series to guarantee uniform convergence of the filter approximation at all time points. It is not difficult to extend the above theorem to cases where the state transitions depend on t as long as the continuity is uniform in t.

The condition (**??**) is used in the proof for bounding

$$\left\|B(A^*E_N(f^N_{r|r}),b_{r+1})-B(A^*f^N_{r|r}),b_{r+1})\right\|_1$$

by applying Lemmas 4 and 5. The following Lemma provides a direct way to bound the above distance by imposing only conditions on a, but assuming a compact state space.

Lemma 7 *Let a be a transition density on a compact state space that satisfies (12) and*

$$\Delta(x',x):=\sup_{x''}\frac{|a(x,x'')-a(x',x'')|}{h(x'')}\to 0\quad(d(x,x')\to 0)\tag{23}$$

with the same density h as in (12). Then under i.i.d. sampling from g

$$\mathbf{P}[\|B(A^*E_N(g),b)-B(A^*g,b)\|_1>\varepsilon]\overset{N\to\infty}{\longrightarrow}0$$

exponentially fast in N for any $\varepsilon>0$, uniformly over all densities g and all likelihoods b with $0<\int h(x')b(x')d\mu(x')<\infty$.

Proof: As in Lemma 5, we choose a partition $\{B_1,\ldots,B_J\}$ such that each B_j has diameter at most δ, and we choose for each j a point $z_j\in B_j$. We are going to show that for a suitable choice of δ

$$\|B(A^*E_N(g),b)-B(A^*g,b)\|_1\leq\frac{4\varepsilon}{5}+C_a^2\sum_{j=1}^J|\mu_N(B_j)-\mu_g(B_j)|.$$

Then the proof is completed as with Lemma 5.

To make the notation more compact, we introduce

$$q(x',x)=\frac{a(x',x)b(x)}{\beta(x')},\quad\beta(x)=\int a(x,x')b(x')d\mu(x').$$

Then $q(x',x)$ is again a transition density and we can write

$$B(A^*E_N(g),b)(x)=\sum_{i=1}^N\frac{\beta(x_i)}{\sum_{k=1}^N\beta(x_k)}q(x_i,x).$$

and

$$B(A^*g,b)(x)=\int\frac{g(x')\beta(x')}{\int g(x'')\beta(x'')d\mu(x'')}q(x',x)d\mu(x')=\sum_{j=1}^J\frac{\int_{B_j}g(x')\beta(x')q(x',x)d\mu(x')}{\sum_{k=1}^J\int_{B_k}g(x'')\beta(x'')d\mu(x'')}.$$

In order to estimate the L_1-distance between these two densities, we will build a chain of four intermediate densities. Putting $x'_i = z_j$ if $x_i \in B_j$, these four densities are

$$\begin{aligned}
g_1(x) &:= \sum_{i=1}^{N} \frac{\beta(x'_i)}{\sum_{k=1}^{N} \beta(x'_k)} q(x_i, x), \\
g_2(x) &:= \sum_{i=1}^{N} \frac{\beta(x'_i)}{\sum_{k=1}^{N} \beta(x'_k)} q(x'_i, x) = \sum_{j=1}^{J} \frac{\mu_N(B_j)\beta(z_j)}{\sum_{k=1}^{J} \mu_N(B_k)\beta(z_k)} q(z_j, x), \\
g_3(x) &:= \sum_{j=1}^{J} \frac{\mu_g(B_j)\beta(z_j)}{\sum_{k=1}^{J} \mu_g(B_k)\beta(z_k)} q(z_j, x), \\
g_4(x) &:= \sum_{j=1}^{J} \frac{\int_{B_j} g(x')\beta(x')d\mu(x')}{\sum_{k=1}^{J} \int_{B_k} g(x'')\beta(x'')} q(z_j, x).
\end{aligned}$$

Hence we have to bound five L_1-distances between successive intermediate densities. The crucial estimate is

$$\|g_3 - g_2\|_1 \leq \sum_{j=1}^{J} \left| \frac{\mu_N(B_j)\beta(z_j)}{\sum_{k=1}^{J} \mu_N(B_k)\beta(z_k)} - \frac{\mu_g(B_j)\beta(z_j)}{\sum_{k=1}^{J} \mu_g(B_k)} \right|.$$

By Lemma 4 this is bounded by

$$\frac{\max \beta(z_j)}{\min \beta(z_j)} \sum_{j=1}^{J} |\mu_N(B_j) - \mu_g(B_j)| \leq C_a^2 \sum_{j=1}^{J} |\mu_N(B_j) - \mu_g(B_j)|$$

since by condition (12) for any x

$$C_a^{-1} \int h(x')b(x')d\mu(x') \leq \beta(x) \leq C_a \int h(x')b(x')d\mu(x').$$

The other four distances can be all made less than $\varepsilon/5$ by a suitable choice of δ. Since all cases are similar, we give the one for

$$\|g_4 - B(A^* g, b)\|_1 \leq \sup_j \sup_{x' \in B_j} \|q(x', .) - q(z_j, .)\|_1 .$$

Because of (**??**), it is sufficient to show that $q(x', x)/q(z_j, x)$ converges to one uniformly over x, $x' \in B_j$ and j as δ goes to zero. This is true since by condition (**??**)

$$\frac{q(x', x)}{q(z_j, x)} = \left(1 + \frac{a(x', x) - a(z_j, x)}{a(z_j, x)}\right) \left(1 + \frac{\beta(z_j) - \beta(x')}{\beta(x')}\right) \leq (1 + C_a \Delta(x', z_j))^2.$$

□

By looking at the proof of Theorem 2, this Lemma implies immediately

Theorem 3 *The claim of Theorem 2 is valid if the state space is compact, the transition densities do not depend on t and (12), (13) and (**??**) hold.*

4.3 Central limit theorems

The goal of this section is to show by a simple induction argument that

$$\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^{N}\psi_s(x_{j,s}) - \int \psi_s(x) f_{s|s}(x) d\mu(x)\right)_{0\leq s\leq t}$$

is asymptotically centered normal for any fixed t, any y_1^t and functions ψ_s, $0 \leq s \leq t$, which are square integrable w.r. to $f_{s|s}$. Del Moral and Miclo (2000), Corollary 20, have obtained a similar result, but we do not assume the ψ_s's to be bounded nor the likelihood $b_t(., y_t)$ to be bounded way from zero.

Our argument proceeds by induction on the number t of time steps. For $t = 0$, the result is obvious because $(x_{j,0})$ is an i.i.d sample from $f_{0|0} = a_0$. The key idea for the induction step is to condition on $(x_{j,t-1})$. We first explain the argument heuristically Introducing the notation

$$\begin{aligned} M_{N,t}(\psi) &= \frac{1}{N}\sum_{j=1}^{N}\psi(x_{j,t}), \\ m_{N,t}(\psi) &= \int \psi(x) f_{t|t}^N(x) d\mu(x), \\ m_t(\psi) &= \int \psi(x) f_{t|t}(x) d\mu(x), \end{aligned}$$

we can split

$$\sqrt{N}(M_{N,t}(\psi) - m_t(\psi)) = \sqrt{N}(M_{N,t}(\psi) - m_{N,t}(\psi)) + \sqrt{N}(m_{N,t}(\psi) - m_t(\psi)). \quad (24)$$

Conditionally on all samples up to time $t-1$, $(x_{j,t})$ is an i.i.d sample from $f_{t|t}^N$. Thus the first term in (14) has the conditional limit distribution $\mathcal{N}\left(0, \sigma_{N,t}^2(\psi)\right)$ where

$$\sigma_{N,t}^2(\psi) = \int (\psi(x) - m_{N,t}(\psi))^2 f_{t|t}^N(x) d\mu(x) \approx \sigma_t^2(\psi) = \int (\psi(x) - m_t(\psi))^2 f_{t|t}(x) d\mu(x)$$

if $f_{t|t}^N$ converges to $f_{t|t}$. By the recursions for $f_{t|t}$ and $f_{t|t}^N$, ((1) – (2) and (5) respectively,

$$\sqrt{N}(m_{N,t}(\psi) - m_t(\psi)) = \sqrt{N}\left(\frac{\sum_j L_t\psi(x_{j,t-1})}{\sum_j L_t 1(x_{j,t-1})} - \frac{m_{t-1}(L_t\psi)}{m_{t-1}(L_t 1)}\right) \quad (25)$$

where

$$L_t\psi(x_{t-1}) = \int a_t(x_{t-1}, x_t) b_t(x_t, y_t) \psi(x_t) d\mu(x_t).$$

Asymptotic normality of the second term of (14) follows therefore from the induction assumption and the delta method.

We now state and prove a rigorous result.

Theorem 4 *If $x \to a_t(x, .)$ is continuous and if for all t, all x and all y*

$$0 < b_t(x, y) \leq C(t, y) < \infty,$$

then for all t, all y_1^t and all functions ψ with

$$\sigma_t^2(\psi) = \int (\psi(x) - m_t(\psi))^2 f_{t|t}(x) d\mu(x) < \infty,$$

the recursively defined asymptotic variance

$$V_t(\psi) = \sigma_t^2(\psi) + \frac{1}{p(y_t \mid y_1^{t-1})^2} V_{t-1}\left(L_t(\psi - m_t(\psi))\right).$$

is finite. Moreover, if $\sigma_s^2(\psi_s) < \infty$ *for* $s = 0, 1, \ldots t$, *then the vector* $\sqrt{N}(M_{N,s}(\psi_s) - m_s(\psi_s))_{s=0,\ldots t}$ *converges in distribution to a* $\mathcal{N}(0, (V_{r,s}(\psi_r, \psi_s)))$ *random vector where*

$$V_{r,t}(\psi_r, \psi_t) = V_{r,t-1}(\psi_r, L_t(\psi_t - m_t(\psi_t))).$$

for $r < t$ *and* $V_{t,t}(\psi_t, \phi_t) = (V_t(\psi_t + \phi_t) - V_t(\psi_t) - V_t(\phi_t))/2$.

Proof: Using the Cramér-Wold device, it is sufficient to show that

$$Z_N = \sqrt{N} \sum_{s=0}^{t} (M_{N,s}(\psi_s) - m_s(\psi_s))$$

is asymptotically centered normal with variance

$$\tau^2 = \sum_{r,s=0}^{t} V_{r,s}(\psi_r, \psi_s).$$

For $t = 0$, the theorem is trivially satisfied, and for the induction argument, we decompose $Z_N = Z_N^{(1)} + Z_N^{(2)}$ where

$$Z_N^{(1)} = Z_N = \sqrt{N}(M_{N,t}(\psi_t) - m_{N,t}(\psi_t))$$

and

$$Z_N^{(2)} = \sqrt{N}(m_{N,t}(\psi_t) - m_t(\psi_t)) + \sqrt{N} \sum_{s=0}^{t-1} (M_{N,s}(\psi_s) - m_s(\psi_s)).$$

We first assume that ψ_t is bounded. Denoting by $\mathcal{F}_t$ the σ-field generated by the $(x_{j,s}; 1 \leq j \leq N, 0 \leq s \leq t)$, we can write

$$\mathbf{E}\left[\exp(i\lambda Z_N)\right] = \mathbf{E}\left[\mathbf{E}\left[\exp(i\lambda Z_N^{(1)}) \mid \mathcal{F}_{t-1}\right] \exp(i\lambda Z_N^{(2)})\right].$$

Since conditionally on $\mathcal{F}_{t-1}$ the $x_{j,t}$'s are i.i.d, we have

$$\mathbf{E}\left[\exp(i\lambda Z_N^{(1)}) \mid \mathcal{F}_{t-1}\right] = \left(\mathbf{E}\left[\exp\left(i\frac{\lambda}{\sqrt{N}}(\psi_t(x_{1,t}) - m_{N,t}(\psi_t))\right) \mid \mathcal{F}_{t-1}\right]\right)^N.$$

Furthermore, by a Taylor expansion of $\exp(iu)$

$$\left|\mathbf{E}\left[\exp\left(i\frac{\lambda}{\sqrt{N}}(\psi_t(x_{1,t}) - m_{N,t}(\psi_t))\right) \mid \mathcal{F}_{t-1}\right] - 1 + \frac{\lambda^2 \sigma_{N,t}^2(\psi_t)}{2N}\right| \leq \frac{|\lambda|^3 \sup|\psi_t(x)|^3}{6N^{3/2}}.$$

Similarly, because $1 - u \leq \exp(-u) \leq 1 - u + u^2$ for all $u \geq 0$,

$$\left|1 - \frac{\lambda^2 \sigma_{N,t}^2(\psi_t)}{2N} - \exp(-\lambda^2 \sigma_{N,t}^2(\psi_t)/(2N))\right| \leq \frac{\lambda^4 \sup|\psi_t(x)|^4}{4N^2}.$$

Because $|u^N - v^N| \leq N|u - v|$ for $|u| \leq 1, |v| \leq 1$, we therefore obtain that for any λ

$$\mathbf{E}\left[\exp(i\lambda Z_N^{(1)}) \mid \mathcal{F}_{t-1}\right] - \exp(-\lambda^2 \sigma_{N,t}^2(\psi_t)/2)$$

converges to zero as $N \to \infty$ uniformly. By Theorem 1, $\left\| f_{t|t}^N - f_{t|t} \right\|_1$ converges to zero for $N \to \infty$. Because ψ_t is bounded, this implies that $\sigma_{N,t}^2(\psi_t)$ converges to $\sigma_t^2(\psi_t)$. Therefore

$$\mathbf{E}\left[\left|\exp(-\lambda^2\sigma_{N,t}^2(\psi)/2) - \exp(-\lambda^2\sigma_t^2(\psi)/2)\right|\right] \stackrel{N\to\infty}{\longrightarrow} 0.$$

We now turn to the second term, $Z_N^{(2)}$. The conditions of the theorem guarantee that

$$m_{t-1}(L_t 1) = \int\int f_{t-1|t-1}(x_{t-1}) a_t(x_{t-1}, x_t) b_t(x_t, y_t) d\mu(x_{t-1}) d\mu(x_t) = p(y_t \mid y_1^{t-1})$$

is strictly positive, and $L_t\psi_t$ and $L_t 1$ are easily seen to be bounded if ψ_t is bounded. Hence the conditions for the delta method are satisfied, and so $\sqrt{N}(m_{N,t}(\psi_t) - m_t(\psi_t))$ is asymptotically equivalent to

$$\frac{1}{\sqrt{N} p(y_t \mid y_1^{t-1})} \left(\sum_j (L_t\psi_t(x_{j,t-1}) - m_{t-1}(L_t\psi_t) - m_t(\psi_t) \sum_j (L_t 1(x_{j,t-1}) - m_{t-1}(L_t 1)) \right).$$

This equal to $\sqrt{N}(M_{N,t-1}(\phi_{t-1}) - m_{t-1}(\phi_{t-1}))$ where $\phi_{t-1} = L_t(\psi_t - m_t(\psi_t))/p(y_t \mid y_1^{t-1})$. Hence, by the induction assumption, $\mathbf{E}\left[\exp(i\lambda Z_N^{(2)})\right]$ converges to

$$\exp\left(-\frac{\lambda^2}{2} \left(\sum_{r,s=0}^{t-2} V_{r,s}(\psi_r, \psi_s) + 2 \sum_{s=0}^{t-2} V_{s,t-1}(\psi_s, \psi_{t-1} + \phi_{t-1}) + V_{t-1}(\psi_{t-1} + \phi_{t-1}) \right) \right)$$

which is equal to $\exp(-\lambda^2(\tau^2 - \sigma_t^2(\psi_t))/2)$ because $V_{r,t}(.,.)$ is bilinear.

Taking all this together, we obtain that for bounded ψ_t

$$\begin{aligned} \left| \mathbf{E}\left[\exp(i\lambda Z_N)\right] - \exp(-\lambda^2\tau^2/2) \right| &\leq \mathbf{E}\left[\left| \mathbf{E}\left[\exp(i\lambda Z_N^{(1)}) \mid \mathcal{F}_{t-1}\right] - \exp(-\lambda^2\sigma_t^2(\psi_t)/2) \right| \right] \\ &+ \left| \mathbf{E}\left[\exp(i\lambda Z_N^{(2)})\right] - \exp(-\lambda^2(\tau^2 - \sigma_t^2(\psi_t))/2) \right| \end{aligned}$$

converges to zero.

The last part of the proof deals with the case when ψ_t is unbounded. We show first that $\sigma_t(\psi_t) < \infty$ implies $V_t(\psi_t) < \infty$. Again, we use induction. For $t = 0$, this is clear because $\sigma_0^2(\psi) = V_0(\psi)$. For the induction step, it is sufficient to show that $\sigma_{t-1}(L_t(\psi_t - m_t(\psi_t))) < \infty$ because by our assumptions $p(y_t \mid y_1^{t-1}) > 0$. By Schwarz' inequality $(L_t\psi)^2 \leq L_t(\psi^2) L_t 1$, and by our assumption $L_t 1 \leq C(t, y_t)$ is finite. Hence by the definition of L_t and the recursions (1) – (2)

$$\begin{aligned} \sigma_{t-1}^2(L_t(\psi_t - m_t(\psi))) &\leq m_{t-1}((L_t(\psi_t - m_t(\psi)))^2) \leq C(t, y_t) m_{t-1}(L_t((\psi_t - m_t(\psi))^2)) \\ &= C(t, y_t) p(y_t \mid y_1^{t-1}) \sigma_t^2(\psi_t) < \infty. \end{aligned}$$

For the asymptotic normality, we use a truncation argument. We set

$$\psi_{t,c}(x) = \psi_t(x) 1_{\{|\psi_t(x)| \leq c\}}, \quad \overline{\psi}_{t,c}(x) = \psi_t(x) - \psi_{t,c}(x).$$

Because $V_t(\psi_t) < \infty$, it follows by dominated convergence that

$$V_{r,t}(\psi_r, \psi_{t,c}) \stackrel{c\to\infty}{\longrightarrow} V_{r,t}(\psi_r, \psi_t). \tag{26}$$

Next, we are going to show that

$$\lim_{c\to\infty} \limsup_N \mathbf{P}[\sqrt{N} \left| M_{N,t}(\overline{\psi}_{t,c}) - m_t(\overline{\psi}_{t,c}) \right| \geq \epsilon] = 0. \tag{27}$$

We first condition on $\mathcal{F}_{t-1}$. By Chebyshev's inequality

$$\begin{aligned}\mathbf{P}[\sqrt{N}\left|M_{N,t}(\overline{\psi}_{t,c}) - m_t(\overline{\psi}_{t,c})\right| \geq \epsilon \mid \mathcal{F}_{t-1}] &\leq 1_{\{\sqrt{N}|m_{N,t}(\overline{\psi}_{t,c})-m_t(\overline{\psi}_{t,c})|\geq\epsilon/2\}} \\ &+ \min\left(1, \frac{4}{\epsilon^2} m_{N,t}(\overline{\psi}^2_{t,c})\right).\end{aligned}$$

We therefore have to study the expectations of the two terms on the right. By (**??**),

$$\sqrt{N}(m_{N,t}(\overline{\psi}_{t,c}) - m_t(\overline{\psi}_{t,c})) = \sqrt{N}\left(\frac{\sum_j L_t\overline{\psi}_{t,c}(x_{j,t-1})}{\sum_j L_t 1(x_{j,t-1})} - \frac{m_{t-1}(L_t\overline{\psi}_{t,c})}{m_{t-1}(L_t 1)}\right),$$

which by the induction assumption is asymptotically $\mathcal{N}(0, V_{t-1}(L_t(\overline{\psi}_{t,c} - m_t(\overline{\psi}_{t,c}))))$-distributed. For $c \to \infty$, this variance goes to zero, implying the desired behavior of the first term. By the recursion for $f^N_{t|t}$,

$$m_{N,t}(\overline{\psi}^2_{t,c}) = \frac{\sum_j L_t\overline{\psi}^2_{t,c}(x_{j,t-1})}{\sum_j L_t 1(x_{j,t-1})}.$$

which by the induction assumption converges in probability to $\int \overline{\psi}^2_{t,c}(x) f_{t|t}(x) d\mu(x)$. Hence by dominated convergence, the second term also has the desired behavior, and thus (16) follows.

Now we have all the ingredients to complete the proof. We write

$$Z_{N,c} = \sqrt{N}\sum_{s=0}^{t-1}(M_{N,s}(\psi_s) - m_s(\psi_s)) + (M_{N,t}(\psi_{t,c}) - m_t(\psi_{t,c}))$$

and τ_c^2 for the asymptotic variance of $Z_{N,c}$. Then

$$\begin{aligned}&\left|\mathbf{E}\left[\exp(i\lambda Z_N)\right] - \exp(-\lambda^2\tau^2/2)\right| \leq \left|\mathbf{E}\left[\exp(i\lambda Z_{N,c})\right] - \exp(-\lambda^2\tau_c^2/2)\right| + \\ &\left|\exp(-\lambda^2\tau_c^2/2) - \exp(-\lambda^2\tau^2/2)\right| + \\ &\mathbf{E}\left[\left|\exp(i\lambda\sqrt{N}(M_{N,t}(\psi_{t,c}) - m_t(\psi_{t,c}) - M_{N,t}(\psi) + m_t(\psi))) - 1\right|\right].\end{aligned}$$

By (15) the second term is arbitrarily small if c is large enough. Using $|\exp(iu) - 1| \leq \min(2, |u|)$ and (16), the same thing holds also for the last term, uniformly in N. Finally, the first term goes to zero for any fixed c as $N \to \infty$. □

Similarly as in the case of convergence of $f^N_{t|t}$, one would like to know whether the variances $V_t(\psi)$ remain bounded as t increases. Chopin (2002) has shown that this is the case under the assumptions (11) and (**??**).

Acknowledgment: I am grateful to Neil Shephard and Eric Moulines for helpful comments on an earlier version of this paper. In particular, I thank Eric Moulines for showing me the proof of Theorem 3.

References

Carpenter, J., Clifford, P. and Fearnhead, P. (1999). An improved particle filter for nonlinear problems. *IEE Proceedings Part F, Radar, Sonar and Navigation*, **146**, 2-7.

Chopin, N. (2002). Central limit theorem for sequential Monte Carlo methods and its applications to Bayesian inference. Preprint, INSEE, Paris.

Crisan, D. (2001). Particle filters – A theoretical perspective. In *Sequential Monte Carlo Methods in Practice*, Doucet, A., de Freitas, N., and Gordon, N., eds., 17–41. Springer, Berlin.

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

Del Moral, P. and Guionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. Henri Poincaré, Probab. Statist.*, **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. In *Séminaire de Probabilités XXXIV*, Azéma, J., Émery, M., Ledoux, M., and Yor, M., eds., Lecture Notes in Mathematics, 1729, 1–145. Springer, Berlin.

Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Basel.

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

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

Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear modeling. *J. Time Ser. Anal.*, **15**, 183–202.

Godsill, S. J., Doucet, A., and West, M. (2000). Monte Carlo smoothing for nonlinear time series. In *Proc. International Symposium on Frontiers of Time Series Modelling*. Institute of Statistical Mathematics, Tokyo.

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

Hannan, E. J. and Deistler, M. (1988) . *The Statistical Theory of Linear Systems*. Wiley, New York.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Hürzeler, M. and Künsch, H. R. (1998). Monte Carlo Approximations for general state-space models. *J. Comp. and Graph. Statist.*, **7**, 175–193.

Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximizing the likelihood for a general state space model. In *Sequential Monte Carlo Methods in Practice*, Doucet, A., de Freitas, N., and Gordon, N., eds., 159–175. Springer, New York.

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

Le Gland, F., and Oudjane, N. (2001). Stability and uniform approximation of nonlinear filters using the Hilbert metric, and application to particle filters. Preprint Nr. 1404, IRISA, Rennes.

Liu, J. S., and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, **93**, 1032–1044.

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

Pitt, M. K. and Shephard, N. (2001). Auxiliary variable based particle filters. In *Sequential Monte Carlo Methods in Practice*, Doucet, A., de Freitas, N., and Gordon, N., eds., 273–293. Springer, Berlin.

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

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., eds., 395–402. Oxford University Press, Oxford.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models with Econometric, Finance and Other Applications*, Cox, D. R., Hinkley, D. V. and Barndorff-Nielsen, O. E., eds., 1-67. Chapman and Hall, London.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, **4**, 65-85.