On Convergence of Recursive Monte Carlo Filters in Non-Compact State Spaces

Jing Lei and Peter Bickel

Carnegie Mellon University and University of California, Berkeley

Abstract: We consider the particle filter approximation of the optimal filter in non-compact state space models. A time-uniform convergence result is built on top of a filter stability argument developed by Douc, Moulines, and Ritov (2009), under the assumption of a heavy-tailed state process and an informative observation model. We show that an existing set of sufficient conditions for filter stability is also sufficient, with minor modifications, for particle filter convergence. The rate of convergence is also given and depends on both the sample size and the tail behavior of the transition kernel.

Key words and phrases: state space model, particle filter, consistency.

1 Introduction

Consider a state space model consisting of two sequences of random variables: a Markovian state process $(X_i, i \geq 0)$ in probability space $(\mathcal{X}, \mathcal{F}_\mathcal{X}, P)$ with transition density $q(\cdot,\cdot)$ under a base measure μ_1:

$$P(X_{i+1} \in A | X_i = x) = \int_A q(x, x') d\mu_1(x'),\ \forall\ i \geq 0,$$

for all $A \in \mathcal{F}_\mathcal{X}$ and $x \in \mathcal{X}$; and an observation sequence $(Y_i, i \geq 1)$ in $(\mathcal{Y}, \mathcal{F}_\mathcal{Y}, P)$, where Y_i's are conditionally independent given X_i's, with density function $g(y; x)$ under a base measure μ_2:

$$P(Y_i \in B | X_i = x) = \int_B g(y; x) d\mu_2(y),\ \forall\ i \geq 1,$$

for all $B \in \mathcal{F}_\mathcal{Y}$ and $x \in \mathcal{X}$. The joint distribution of $(X_i, Y_i : i \geq 0)$ is determined by q, g, and p_0, the density of X_0. Models of this form are also known as hidden Markov models (Künsch (2001); Cappé, Moulines, and Rydén (2005)). Typical inference tasks in state space models include: 1) parameter estimation for the state process q and the observation model g (Bickel, Ritov, and Rydén (1998); Olsson and Rydén (2008)); and 2) when q and g are known, computing the conditional distribution of state variables X_i given the observations $Y_1^s \equiv (Y_1, \ldots, Y_s)$, whose density function is denoted by $p_{i|s}$ (Liu and Chen (1998)). Calculating $p_{i|s}$ for $s = i$, $s > i$, and $s < i$ are called filtering, smoothing, and predicting, respectively.

State space models have found wide applications in signal processing, robotics, biology, finance, and geophysics (Liu (2001); Künsch (2001); Cappé, Moulines, and Rydén (2005)). In particular, filtering, an important and classical topic in state space models, has been the focus of much research interest and effort. Filtering aims at calculating or approximating the conditional distribution of X_i given Y_1^i, whose density is denoted by $p_{i|i}$. However, exact calculation of the optimal filter $p_{i|i}$ is usually computationally infeasible due to the nonlinearity in q and g. The particle filter, originally introduced by Gordon, Salmon, and Smith (1993), is one of the most important class of filtering methods because of its easy implementation and the modeling flexibility inherited from its nonparametric nature. Various implementation of particle filters have been studied in the statistics literature. For example, Liu and Chen (1998) consider particle filters under the framework of sequential Monte Carlo methods, where sequential importance sampling and related methods are discussed. Künsch (2005) studies another particle filter implementation under the name of recursive Monte Carlo filters, where importance sampling is not carried out explicitly on the particles, but implicitly through the sampling distribution update. For other discussion about practical implementation of particle filters, see Pitt and Shephard (1999); Lin, et al (2005).

In particle filters, the target distribution $p_{i|i}$ is approximated by the sum of weighted point mass distributions:

$$\hat{p}_{i|i} = \sum_{j=1}^{n} w_i^j \zeta(x_i^j),$$

where $\{x_i^j\}_{j=1}^n$ is a set of particles in the state space chosen by the algorithm, $\zeta(x)$ is the point mass distribution at x, and $\{\omega_i^j\}_{j=1}^n$ is a set of weights satisfying $\omega_i^j \geq 0$, $\sum_j \omega_i^j = 1$. Starting from $\hat{p}_{0|0} = p_0$, $\hat{p}_{i|i}$ is used (instead of $p_{i|i}$) together with the next observation y_{i+1} to obtain the next filtering distribution $\hat{p}_{i+1|i+1}$

in a recursive manner. Such a point mass approximation greatly simplifies the computation of the update from $\hat{p}_{i|i}$ to $\hat{p}_{i+1|i+1}$. Obviously, the accuracy of approximation depends on the choice of particles and their weights. Further details are given in Section 2. For a thorough introduction to the basic theory and application of particle filters, see Doucet, de Freitas, and Gordon (2001).

The popularity and successful application of particle filters urge theoretical justification. Specifically, an important topic is time-uniform convergence:

$$\sup_{i\geq 1} E||\hat{p}_{i|i} - p_{i|i}|| \to 0, \text{ as } n \to \infty, \tag{1.1}$$

where n is the number of particles, $\hat{p}_{i|i}$ is the particle filter approximation of $p_{i|i}$, and $||\cdot||$ is a suitable function norm. In applications, it is also desired to know the rate of convergence. Note that the object $\hat{p}_{i|i}$ has two sources of randomness: the observation sequence $(Y_i, i \geq 1)$ and the Monte Carlo sampling in each iteration. In this paper, unless otherwise noted, the expectation is with respect to both sources of randomness.

A common approach (Del Moral and Guionnet (2001); Künsch (2005)) of establishing time-uniform convergence for particle filters is based on the fact that the Monte Carlo sampling error introduced in each iteration is propagated over time by a single Bayes operator and a sequence of Markov operators. The argument generally consists of two components. The first is to develop a uniform upper bound on the single step sampling error after being propagated by the Bayes operator. A particular challenge is to provide a lower bound of the normalizing constant in the Bayes formula. The second is to show that the conditional chain $(X_i|Y_1^s : 1 \leq i \leq s)$ is uniformly contracting so that the single step approximation error vanishes exponentially as a function of time. This usually requires mixing conditions on the state transition kernel of the form

$$c_- a(\cdot) \leq q(x, \cdot) \leq c_+ a(\cdot), \quad \forall x \in \mathcal{X}, \tag{1.2}$$

for some density function $a(\cdot)$ and positive constants c_-, c_+. Condition (1.2) is often too strong to hold when $\mathcal{X}$ is not compact. Meanwhile, the compactness and (1.2) also play an important role in obtaining a lower bound of the normalizing constant in the Bayes formula. Therefore, most convergence results for particle filters are only applicable to compact state spaces.

This paper is part of the effort of proving time-uniform convergence for particle filters in non-compact state spaces. The argument follows the general framework described above. We consider autoregressive models with a heavy-tailed transition kernel and lighter-tailed observation, for which the uniform contracting property has been proven in Douc, Moulines, and Ritov (2009). We show that the sufficient conditions for the uniform contracting property are also sufficient, with minor modifications, for controlling the single step approximation error. There are two key assumptions. First, the tail of the state process is at least exponential or heavier in a sense defined in equation (4.2), which is an important example of the "pseudo-mixing" condition (Le Gland and Oudjane (2003)). Second, the observation likelihood has lighter tails than the transition kernel. This is a relaxation of the "bounded observation" model and enables us to avoid the use of truncation (Le Gland and Oudjane (2003); Heine and Crisan (2008)) for likelihood functions with unbounded support.

Our result is particulary applicable to autoregressive models of the form:

$$\begin{aligned} X_i &= a(X_{i-1}) + U_i, \\ Y_i &= b(X_i) + V_i. \end{aligned} \tag{1.3}$$

Assuming that U_i and V_i have appropriate tail behavior, we show that (Theorem 5)

$$\sup_{i>0} E\big|\big|\hat{p}_{i|i} - p_{i|i}\big|\big|_{\mathrm{tv}} \le c\theta n^{-1/2} + 2P(|U_1| \ge \theta),\ \forall\ \theta > 0, \tag{1.4}$$

where constant $c > 0$ depends on the model only, and $||\cdot||_{\mathrm{tv}}$ denotes the total variation norm. The free parameter $\theta > 0$ can be chosen according to the tail probability of U_1 to optimize the rate of convergence. Comparing with typical compact state space results, this rate has an extra term $P(|U_1| \ge \theta)$, and is slower than the usual $O(n^{-1/2})$ rate. This is actually a price paid for providing a non-trivial lower bound on the normalizing constant in the Bayes formula in non-compact state and observation spaces.

In related work, Le Gland and Oudjane (2003) study time-uniform approximation under the pseudo-mixing condition with application to a truncated particle filter. Heine and Crisan (2008) study uniform convergence of a truncated particle filter under a weaker norm for a different class of autoregressive models, including some special cases that are not pseudo-mixing. The convergence rate is provided in terms of both the truncation parameter and the number of particles. Douc, Moulines, and Ritov (2009) establish the uniform contracting property of the original non-truncated filter for pseudo-mixing state processes with lighter-tailed observation models. The contracting property is then used to prove the filter stability, which says that the filtering distribution $p_{i|i}$ has little dependence on the initial distribution p_0 for large i. van Handel (2009) uses a different approach that employs the ergodicity of Markov process $\{(X_i, p_{i|i}) : i \geq 0\}$, giving a Cesàro-type time-uniform convergence for general state space models, without rates of convergence. The key assumption there is that the approximated filtering distributions are uniformly tight in Cesàro sense. A set of sufficient conditions include geometric ergodicity of the state process $(X_i : i \geq 0)$ and appropriate tail behaviors.

In Section 2 we briefly review the filtering problem with a special focus on particle filters. In Section 3 we present some general arguments to control the error propagation. In Section 4 we apply the arguments to a class of autoregressive models, for which time-uniform convergence is developed. Some further remarks and possible future research topics are given in Section 5. Some lengthy proofs are included in Section 6.

2 Preliminaries on filtering

The objective of this section is to provide necessary prerequisites for the filtering problem, Monte Carlo approximation, and relevant notations. For presentation convenience, we write $g_i(x_i)$ for $g(y_i; x_i)$. The conditional density of X_i given Y_1^s is written as $p_{i|s}(\cdot)$. The base measures μ_1 and μ_2 are not crucial in the argument and results, so we focus on Euclidean spaces for simplicity and assume that both the transition density q and observation density g are positive everywhere under the Lebesgue measure.

The dependence structure of a state space model can be described by the following diagram:

$$\begin{tikzcd}
\dots \arrow[r] & X_{i-1} \arrow[r] \arrow[d] & X_i \arrow[r] \arrow[d] & X_{i+1} \arrow[r] \arrow[d] & \dots \\
\dots & Y_{i-1} & Y_i & Y_{i+1} & \dots
\end{tikzcd}$$

This graph representation leads to some basic recursive formulas.

2.1 The forward propagation and Monte Carlo approximation

Suppose at time $i \geq 1$ we have obtained the ideal (optimal) filtering distribution $p_{i-1|i-1}$, then the conditional distribution of X_i given Y_1^{i-1} is obtained by applying the Markov kernel induced from transition density q on the density function $p_{i-1|i-1}$:

$$p_{i|i-1}(x_i) = \int p_{i-1|i-1}(x_{i-1}) q(x_{i-1}, x_i) dx_{i-1}. \tag{2.1}$$

When the new observation $Y_i = y_i$ is available, the distribution of X_i given $Y_1^i = y_1^i$ is obtained by applying the Bayes formula on the forecast density $p_{i|i-1}$ with likelihood function $g_i(\cdot) \equiv g(y_i; \cdot)$:

$$p_{i|i}(x_i) = \frac{p_{i|i-1}(x_i) g_i(x_i)}{\int p_{i|i-1}(x) g_i(x) dx}. \tag{2.2}$$

The right hand side of (2.2) is well-defined when q and g are positive everywhere.

In practice the prediction (2.1) and Bayes update (2.2) do not allow any closed form solutions. Particle filters tackle this difficulty using Monte Carlo samples to approximate the conditional distributions. We consider the recursive Monte Carlo (RMC, Künsch (2005)) filter as a generic form of particle filter.

In RMC approximations, the integral in (2.1) is substituted by averaging over a random sample:

$$\hat{p}_{i|i-1}(x_i) = \frac{1}{n} \sum_{j=1}^{n} q(x_{i-1}^j, x_i), \tag{2.3}$$

where $\{x_{i-1}^j, j = 1, \ldots, n\}$ is an i.i.d sample from $\hat{p}_{i-1|i-1} \approx p_{i-1|i-1}$. The Bayes update step is similar:

$$\hat{p}_{i|i}(x_i) = \frac{\hat{p}_{i|i-1}(x_i) g_i(x_i)}{\int \hat{p}_{i|i-1}(x) g_i(x) dx}.$$

The recursion starts from $\hat{p}_{0|0} = p_{0|0} = p_0$. See Künsch (2005) for a detailed discussion on implementation of RMC filters.

2.2 The operator notation

For any density function p, transition density q, and likelihood g, define the Markov transition operator Q and Bayes operator B:

$$Qp(x) = \int p(x')q(x',x)dx',$$
$$B(p,g)(x) = \frac{p(x)g(x)}{\int p(x')g(x')dx'}.$$

Then the forward recursion of the optimal filter can be represented by the operator F_i:

$$p_{i|i} = F_i p_{i-1|i-1} := B(Qp_{i-1|i-1}, g_i).$$

For RMC filters, define the random Markov transition kernel $\hat{Q}$:

$$\hat{Q}p(x) = \frac{1}{n}\sum_{j=1}^{n} q(z^j, x),$$

with $\{z^j, j = 1, \ldots, n\}$ an i.i.d sample from $p(\cdot)$. Therefore, the RMC forward recursion becomes

$$\hat{p}_{i|i} = \hat{F}_i \hat{p}_{i-1|i-1} := B(\hat{Q}\hat{p}_{i-1|i-1}, g_i).$$

We wish to control

$$||\hat{p}_{i|i} - p_{i|i}||_{\text{tv}} = ||F_i F_{i-1} \ldots F_1 p_{0|0} - \hat{F}_i \hat{F}_{i-1} \ldots \hat{F}_1 p_{0|0}||_{\text{tv}},$$

where $||\cdot||_{\text{tv}}$ refers to the total variation norm:

$$||f||_{\text{tv}} := \int |f(x)|dx,$$

for any measurable function f.

2.3 The backward recursion and an alternative representation of sequential filtering

The Monte Carlo approximation $\hat{Q}$ of Q introduces a sampling error of order $O_P(n^{-1/2})$ in each iteration. Such an error is subsequently propagated by the Bayes operators $B(\cdot, g_i)$, which may be expanding (Künsch (2001, Lemma 3.6)).

Therefore, propagating through multiple Bayes operators may result in an exponential growth of the sampling error. One can bypass this difficulty by looking at a different way of getting $p_{i|i}$ from $p_{0|0}$. Define the backward function $\beta_{i,s}$ as the conditional probability of observing y_{i+1}^s given x_i:

$$\beta_{i,s}(x_i) = \begin{cases} \int_{\mathcal{X}^{s-i}} \prod_{j=i+1}^{s} q(x_{j-1}, x_j) g_j(x_j) dx_j, & i \leq s-1; \\ 1, & i \geq s, \end{cases}$$

where $\mathcal{X}^{s-i}$ is the $(s-i)$-tuple product space containing the state vector $(x_{i+1}, ..., x_s)$. It is easy to check that for all $i \leq s-1$, $\beta_{i,s}$ follows a backward recursion:

$$\beta_{i,s}(x_i) = \int q(x_i, x_{i+1}) g_{i+1}(x_{i+1}) \beta_{i+1,s}(x_{i+1}) dx_{i+1}. \tag{2.4}$$

The backward function can be used to calculate $p_{i|s}$ for $s > i$:

$$p_{i|s} = B(p_{i|i}, \beta_{i,s}).$$

Based on the backward function, we are ready to introduce an alternative representation of the evolvement from $p_{i|i}$ to $p_{s|s}$, which involves only one Bayes operator $B(\cdot, \beta_{i,s})$. First we state the Markov property of the conditional chain of X_i given Y_1^s.

Lemma 1. *For any $s \geq 1$, the conditional chain $(X_i, 0 \leq i \leq s)$ given $Y_1^s = y_1^s$ is a (possibly non-homogenous) Markov chain, with transition kernel $F_{i+1|s}$: $\mathcal{X} \times \mathcal{F}_{\mathcal{X}} \mapsto [0,1]$,*

$$F_{i+1|s}(x_i, A) = \frac{\int_A q(x_i, x_{i+1}) g_{i+1}(x_{i+1}) \beta_{i+1,s}(x_{i+1}) dx_{i+1}}{\beta_{i,s}(x_i)}.$$

We refer the reader to Cappé, Moulines, and Rydén (2005, Proposition 3.3.2) for a proof of Lemma 1.

Lemma 1 suggests that:

$$p_{s|s} = F_{s|s} \dots F_{1|s} B\left(p_{0|0}, \beta_{0,s}\right), \tag{2.5}$$

or more generally for all $i \leq s-1$ and any density p on $\mathcal{X}$,

$$F_s \dots F_{i+1} p = F_{s|s} \dots F_{i+1|s} B(p, \beta_{i,s}). \tag{2.6}$$

Equations (2.5) and (2.6) show how to obtain $p_{s|s}$ with only a single Bayes operator followed by a sequence of Markov operators. This observation is useful in controlling the propagation of sampling error because Markov operators are contracting under total variation norm: if F is a Markov kernel, its contraction coefficient $\delta(F) \in [0,1]$ is

$$\delta(F) = \sup_{f_1, f_2} \frac{||Ff_1 - Ff_2||_{\text{tv}}}{||f_1 - f_2||_{\text{tv}}}, \tag{2.7}$$

where the supreme is over all pairs of densities f_1 and f_2 on $\mathcal{X}$. Apparently, for any Markov kernels F and F',

$$\delta(FF') \leq \delta(F)\delta(F'). \tag{2.8}$$

3 Convergence of Recursive Monte Carlo Filters

In this section we introduce some general arguments and conditions to develop a uniform upper bound on $E||\hat{p}_{i|i} - p_{i|i}||_{\text{tv}}$. The arguments are applied to a class of autoregressive models with a set of sufficient conditions in Section 4. Consider a decomposition of the total approximation error for $p_{s|s}$:

$$\begin{aligned}
&\left|\left|\hat{p}_{s|s} - p_{s|s}\right|\right|_{\text{tv}} \\
=& \left|\left|\hat{F}_s \cdots \hat{F}_1 p_{0|0} - F_s \cdots F_1 p_{0|0}\right|\right|_{\text{tv}} \\
=& \left|\left|\sum_{i=1}^{s} F_s \cdots F_{i+1}\hat{F}_i \cdots \hat{F}_1 p_{0|0} - F_s \cdots F_i \hat{F}_{i-1} \cdots \hat{F}_1 p_{0|0}\right|\right|_{\text{tv}} \\
\leq& \sum_{i=1}^{s} \left|\left|F_s \cdots F_{i+1}\hat{F}_i \hat{p}_{i-1|i-1} - F_s \cdots F_i \hat{p}_{i-1|i-1}\right|\right|_{\text{tv}} \\
\leq& \sum_{i=1}^{s} \delta(F_{s|s} \cdots F_{i+1|s}) \left|\left|B(\hat{Q}\hat{p}_{i-1|i-1}, g_i\beta_{i,s}) - B(Q\hat{p}_{i-1|i-1}, g_i\beta_{i,s})\right|\right|_{\text{tv}},
\end{aligned} \tag{3.1}$$

where the second step uses a stepwise decomposition of the approximation error; the third step uses the triangle inequality and the definition of $\hat{p}_{i-1|i-1}$; and the last step uses the fact $B(B(p,g),h) = B(p,gh)$ and (2.6). Use the notation

$$\Delta_{i,s} = \left|\left|B(\hat{Q}\hat{p}_{i-1|i-1}, g_i\beta_{i,s}) - B(Q\hat{p}_{i-1|i-1}, g_i\beta_{i,s})\right|\right|_{\text{tv}},$$

then controlling the particle filter approximation error amounts to two tasks: 1) provide an upper bound of the single step error $\Delta_{i,s}$, and 2) show that $\delta(F_{s|s} \cdots F_{i+1|s}) \leq \rho^{s-i}$ uniformly for all y_1^s with $0 < \rho < 1$.

3.1 Single-step approximation error

We first look at the single step sampling error:

$$\begin{aligned}\Delta_{i,s} =& \left|\left| B\left(\hat{Q}\hat{p}_{i-1|i-1}, g_i\beta_{i,s}\right) - B\left(Q\hat{p}_{i-1|i-1}, g_i\beta_{i,s}\right)\right|\right|_{\mathrm{tv}} \\ =& \int \left| \frac{\hat{Q}\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)}{\int \hat{Q}\hat{p}_{i-1|i-1}(x_i')g_i(x_i')\beta_{i,s}(x_i')dx_i'} - \frac{Q\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)}{\int Q\hat{p}_{i-1|i-1}(x_i')g_i(x_i')\beta_{i,s}(x_i')dx_i'}\right| dx_i \\ \leq& \frac{2\int \left|\hat{Q}\hat{p}_{i-1|i-1}(x_i) - Q\hat{p}_{i-1|i-1}(x_i)\right| g_i(x_i)\beta_{i,s}(x_i)dx_i}{\int Q\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)dx_i}. \end{aligned} \tag{3.2}$$

The last inequality stems from the following.

Lemma 2 (Künsch (2001, Lemma 3.6)). *Let f and h be two non-negative integrable functions with $\int f(x)dx > 0$ and $\int h(x)dx > 0$, then*

$$\int \left| \frac{f(x)}{\int f(x')dx'} - \frac{h(x)}{\int h(x')dx'}\right| dx \leq \frac{2\int |f(x) - h(x)|dx}{\int f(x)dx}.$$

Proof. With out loss of generality, assume $\int f(x)dx \geq \int h(x)dx$. Then

$$\begin{aligned} &\int \left| \frac{f(x)}{\int f(x')dx'} - \frac{h(x)}{\int h(x')dx'}\right| dx = 2\int \left(\frac{f(x)}{\int f(x')dx'} - \frac{h(x)}{\int h(x')dx'}\right)_+ dx \\ =&2\left(\int f(x')dx'\right)^{-1}\int\left(f(x) - \frac{\int f(x')dx'}{\int h(x')dx'}h(x)\right)_+ dx \\ \leq&2\left(\int f(x')dx'\right)^{-1}\int\left(f(x) - h(x)\right)_+ dx \\ \leq&2\left(\int f(x')dx'\right)^{-1}\int\left|f(x) - h(x)\right| dx, \end{aligned}$$

where $(z)_+ \equiv \max(z, 0)$. The conclusion follows from the assumption $\int f(x)dx \geq \int h(x)dx$. □

Next we give upper and lower bounds on the numerator and denominator in (3.2), respectively. If $||q||_\infty := \sup_{x,x'} q(x, x') \leq M$, then

$$\left|\hat{Q}\hat{p}_{i-1|i-1}(x_i) - Q\hat{p}_{i-1|i-1}(x_i)\right| = O_P\left(\frac{M}{\sqrt{n}}\right). \tag{3.3}$$

Therefore, roughly speaking, the numerator of (3.2) is bounded by $O(n^{-1/2}M\int g_i\beta_{i,s})$.

It remains to control the function $g_i\beta_{i,s}$, which is intractable in general. As $g_i\beta_{i,s}$ appears in both the numerator and denominator in (3.2), one can expect some cancelation if it can be separated out from the integral in the denominator. The following assumption validates such a cancelation, and provides a lower bound on the denominator:

(**A1**) For every y, there exists a compact set $C_y \subseteq \mathcal{X}$, such that

(a) For all x, x', y,

$$\min\left\{\frac{\int_{C_y} q(x,x')g(y;x')dx'}{\int q(x,x')g(y;x')dx'}, \frac{\int_{C_y} g(y;x)q(x,x')dx}{\int g(y;x)q(x,x')dx}, \frac{\int_{C_y} g(y;x)dx}{\int g(y;x)dx}\right\} \geq \kappa > 0,$$

where κ is a positive constant independent of (x, x', y).

(b) $\kappa \leq \int_{C_0} p_0(x_0)dx_0 \leq 1$, where $C_0 \subseteq \mathcal{X}$ is a compact set that depends only on p_0, and κ is the constant of part (a).

(c) $E\sup_{x\in C_{Y_i}, x'\in C_{Y_{i+1}}} (q(x,x'))^{-1} < \infty$ for all $i \geq 0$.

Part (a) essentially requires that, as demonstrated in Section 4, the observation provides more information about the current state than the previous and future states, which is satisfied when the likelihood has lighter tails than the transition kernel. For any y, the set C_y can be thought as the set of state variables that are "likely" to generate observation $Y = y$. In part (b), the choice of C_0 is not very crucial and we can choose it to be a level set defined as $\{x_0 : p(y_1|x_0) \geq \lambda_0\}$, for some $\lambda_0 > 0$.

Part (c) requires C_{Y_i} and $C_{Y_{i+1}}$ to be "close" enough, on average, with respect to the randomness of $(Y_i, i \geq 1)$. It provides a lower bound on the denominator factor $Q\hat{p}_{i-1|i-1}$ on C_i. If

$$\xi_i := \sup_{x_{i-1}\in C_{Y_{i-1}}, x_i \in C_{Y_i}} (q(x_{i-1}, x_i))^{-1}, \tag{3.4}$$

we have the following.

Lemma 3. *Under (A1), we have, for any* $1 \leq i \leq s-1$, *conditioning on* Y_1^s,

$$\inf_{x_i\in C_i} Q\hat{p}_{i-1|i-1}(x_i) \geq \kappa\xi_i^{-1}.$$

The proof of Lemma 3 is postponed to Section 6.1.

Recall that the numerator in (3.2) is roughly $O(\int g_i\beta_i/\sqrt{n})$, and Lemma 3 implies that the denominator is at least $\kappa\int_{C_i} g_i\beta_i/\xi_i$. Then (A1a) and (A1b) finishes the cancelation of $\int g_i\beta_{i,s}$. Formally, let E_i denote the expectation over Monte Carlo samples from $\hat{p}_{i|i}$ conditioning on $(Y_i, i\geq 1)$, then we have the following.

Lemma 4. *Take ξ_i as in (3.4). Assuming (A1), and that $||q||_\infty\leq M$,*

$$E_{i-1}\Delta_{i,s}\leq 2\min\left(1,\frac{M\xi_i}{\kappa^2\sqrt{n}}\right).$$

A detailed proof of Lemma 4 is given in Section 6.1.

In order to verify condition (A1), the construction of C_y requires further investigation on the tail behavior of q and g. We illustrate the idea through an autoregressive model in Section 4. In subsequent arguments we write C_i for C_{y_i}, and $\sup_{C_i\times C_{i+1}}$ for $\sup_{x\in C_i, x'\in C_{i+1}}$.

3.2 A time-uniform result with uniformly contracting kernels

To obtain time-uniform convergence, we also need to show that the sequence of Markov operators $F_{i+1|s},\ldots,F_{s|s}$ are uniformly contracting:

$$\delta\left(\prod_{r=i+1}^{s}F_{r|s}\right)\leq\rho^{s-i},\tag{3.5}$$

for some $\rho\in(0,1)$ independent of the observation sequence $(y_1,\ldots,y_s)$.

If (3.5) holds, by contracting of Markov kernels and the decomposition in (3.1) we have, by Lemma 4,

$$\begin{aligned}E\left|\left|\hat{p}_{s|s}-p_{s|s}\right|\right|_{\mathrm{tv}} &\leq\sum_{i=1}^{s}\delta\left(\prod_{r=i+1}^{s}F_{r|s}\right)E\Delta_{i,s}\\ &\leq 2\sum_{i=1}^{s}\rho^{s-i}E\left(1\bigwedge\frac{M}{\kappa^2\sqrt{n}}\xi_i\right)\\ &\leq\frac{2}{1-\rho}\sup_{i>0}E\left(1\bigwedge\frac{M}{\kappa^2\sqrt{n}}\xi_i\right).\end{aligned}\tag{3.6}$$

Douc, Moulines, and Ritov (2009) introduce a set of sufficient conditions to prove (3.5) for autoregressive models of form (1.3) with a pseudo-mixing kernel and lighter-tail likelihood. In the next section we show that a simple modification of these conditions makes them also sufficient for (A1) and $\sup_{i\geq 0} E(1 \wedge M\kappa^{-2}\xi_i/\sqrt{n}) = o(1)$, and therefore sufficient for time-uniform particle approximation.

4 Functional autoregressive model

Here we look at a nonlinear, non-Gaussian state space model (see also Le Gland and Oudjane (2003); Heine and Crisan (2008); Douc, Moulines, and Ritov (2009)):

$$\begin{aligned} X_i &= a(X_{i-1}) + U_i, \\ Y_i &= b(X_i) + V_i, \end{aligned} \tag{4.1}$$

for $i \geq 1$, with $X_0 \sim p_0$. Here $(U_i : i \geq 1)$ and $(V_i : i \geq 1)$ are two independent sequences of random variables, with continuous densities p_U and p_V, respectively. For presentation simplicity we focus on the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^1$. Extension to $\mathbb{R}^d$ is straightforward.

Consider the following set of conditions.

(C1) Function $a(\cdot)$ is Lipschitz: $|a(x) - a(x')| \leq a_+|x - x'|$; function $b(\cdot)$ is one-to-one differentiable with derivative bounded and bounded away from zero: $0 < b_- \leq |b'(x)| \leq b_+$, $\forall x$.

(C2) p_U is non increasing on $[0, \infty)$, with $p_U(x) = p_U(|x|)$ and $||p_U||_\infty \leq M < \infty$. Moreover, for all $x \geq 0$ and $x' \geq 0$,

$$\frac{p_U(x + x')}{p_U(x)p_U(x')} \geq r > 0. \tag{4.2}$$

(C3) $p_V(y) = p_V(|y|)$, and p_V is non increasing on $[0, \infty)$, satisfying

$$\int [p_U(cx)]^{-2} p_V(x)dx < \infty, \quad \forall c > 0. \tag{4.3}$$

(C4) The initial distribution also has lighter tail than p_U:

$$\int \left[p_U(a_+^{-1}b_-x)\right]^{-1} p_0(x)dx < \infty. \tag{4.4}$$

Similar conditions are considered by Le Gland and Oudjane (2003) in studying uniform particle approximation of truncated filters, and by Douc, Moulines, and Ritov (2009) in proving filter stability for non-truncated filters. Here in (4.1) the state propagation function $a(\cdot)$ and observation function $b(\cdot)$ stay the same for all $i \geq 1$, which is just for presentation simplicity. In fact they can depend on time index i and the results remain valid if we modify condition (C1) to that $|a_i(x) - a_i(x')| \leq a_+|x - x'|$ and $0 < b_- \leq |b_i'(x)| \leq b_+$ for all $i \geq 1$ and (x, x'). Equation (4.2) is an example of the so-called "pseudo-mixing" condition. It indicates a somewhat heavy tail of p_U that is satisfied for exponential, logistic, and Pareto-type tails (not for Gaussian). The conditions in Heine and Crisan (2008) allow p_U to have lighter tails, including Gaussian, but not power-law tails. The condition of p_U and p_V being non-increasing on $[0, \infty)$ can be relaxed to being non-increasing on $[L, \infty)$ and strictly positive on $[0, L]$. The case $L = 0$ is qualitatively not special, but allows concise presentation.

Our main result is the next theorem which is proved in Section 6.2:

Theorem 5. *Under Model (4.1), assuming (C1)- (C4), there exists a constant c depending on the model only, such that*

$$\sup_{s\geq 0} E\left|\left|\hat{p}_{s|s} - p_{s|s}\right|\right|_{\mathrm{tv}} \leq \frac{c\theta}{\sqrt{n}} + 2P(|U_1| > \theta), \quad \forall\ \theta > 0.$$

Theorem 5 indicates that the time-uniform expected approximation error is bounded by the sum of two parts: one determined by the sample size and one by the tail behavior of the state noise. Such a rate is slower than the $O(1/\sqrt{n})$ rate usually seen in compact state spaces (Künsch (2005)). It is a consequence of the need to provide a lower bound away from zero for the normalizing constant in the Bayes formula in non-compact state spaces. In fact, the term $P(|U_1| > \theta)$ comes from ξ_i, which is the supreme of $(q(x_{i-1}, x_i))^{-1}$ on $C_{i-1} \times C_i$, and can be large if C_{i-1} and C_i are far away from each other. Clearly this will never be a problem if the state space is compact.

The free parameter θ can be chosen to optimize the rate of convergence.

Example 6. If the state noise U_1 has Pareto-type (power-law) tail,

$$P(|U_1| > \theta) = O(\theta^{-\alpha})$$

for some $\alpha > 0$. Choosing $\theta = n^{1/(2+2\alpha)}$ in Theorem 5 yields

$$\sup_{s\geq 0} E\left|\hat{p}_{s|s} - p_{s|s}\right| = O(n^{-\frac{\alpha}{2+2\alpha}}).$$

Example 7. If the state noise U_1 has exponential tails,

$$P(|U_1| > \theta) = O\left(e^{-\theta^\alpha}\right)$$

for some $0 < \alpha \leq 1$. With $\theta = (\log n/2)^{1/\alpha}$, then

$$\sup_{s\geq 0} E\left|\hat{p}_{s|s} - p_{s|s}\right| = O\left((\log n)^{\frac{1}{\alpha}} n^{-\frac{1}{2}}\right).$$

That is, when U_1 has exponential tails, the rate of convergence suggested by Theorem 5 is only slightly slower than $n^{-1/2}$.

5 Final remarks

We have shown that the recursive Monte Carlo filter is consistent uniformly over time for noisy autoregressive models when the state process is heavy-tailed and the observation error has lighter tails. Such a tail constraint is used in both establishing the filter stability, as well as controlling the single step error. Although a heavy tail in the state process noise facilitates our argument, it adds an extra term in the convergence rate. A heavier tail indicates a slower convergence. Following this intuition, if the state process has even lighter tails, such as Gaussian, the convergence result still holds. However, some other technique must be used in this case to deal with the backward function and the normalizing constant. We regard this as an interesting topic for future research.

The results in this paper provide upper bounds on the approximation error. In practice, the asymptotic behavior of particle filter based estimators provides useful guidance for further inferences, such as confidence interval and hypothesis testing. In this direction, central limit theorems are given by Del Moral and Jacod (2001) for Gaussian linear models, Chopin (2004) for particle filters with importance sampling and resampling, and Künsch (2005) for recursive Monte Carlo filters. These central limit theorems are not time-uniform, and hence hold under weaker conditions than those typically required by time-uniform results as considered in this paper. Little is known about central limit theorems in a time-uniform sense for particle filters and its variants, and this may be an important topic for future research.

In this paper we assume that the hidden Markov model is completely known. In practice, for example, in finance, biology, and geophysics, it is often desired to estimate unknown parameters and do filtering at the same time. This is indeed the focus of many research endeavors (see Liu and West (2001); Cappé and Moulines (2005); Polson, Stroud, and Müller (2008), for example). It will be also interesting to connect current work in filter convergence with parameter estimation and to provide theoretical understanding for particle based parameter estimation.

6 Proofs

In the proofs we use $c, c_i, (i = 0, 1, 2, \dots)$ to denote positive constants that depend on the model only. Their values may change in different displays.

6.1 Proofs of Section 3

In this section we prove Lemma 4, starting with Lemma 3.

Proof of Lemma 3. For $i \geq 2$, let $\{x_{i-2}^j\}_{j=1}^n$ be an i.i.d sample from $\hat{p}_{i-2|i-2}$. Then by (A1), for all $x_i \in C_i$,

$$\begin{aligned}
Q\hat{p}_{i-1|i-1}(x_i) =& \int \hat{p}_{i-1|i-1}(x_{i-1})q(x_{i-1}, x_i)dx_{i-1} \\
\geq& \int_{C_{i-1}} \hat{p}_{i-1|i-1}(x_{i-1})q(x_{i-1}, x_i)dx_{i-1} \\
\geq& \xi_i^{-1} \int_{C_{i-1}} \hat{p}_{i-1|i-1}(x_{i-1})dx_{i-1} \\
=& \xi_i^{-1} \frac{\int_{C_{i-1}} n^{-1} \sum_{j=1}^n q(x_{i-2}^j, x_{i-1})g_{i-1}(x_{i-1})dx_{i-1}}{\int n^{-1} \sum_{j=1}^n q(x_{i-2}^j, x_{i-1})g_{i-1}(x_{i-1})dx_{i-1}} \\
=& \xi_i^{-1} \frac{\sum_{j=1}^n \int_{C_{i-1}} q(x_{i-2}^j, x_{i-1})g_{i-1}(x_{i-1})dx_{i-1}}{\sum_{j=1}^n \int q(x_{i-2}^j, x_{i-1})g_{i-1}(x_{i-1})dx_{i-1}} \\
\geq& \xi_i^{-1}\kappa.
\end{aligned}$$

For $i = 1$ we have, according to (A1), for all $x_1 \in C_1$,

$$\begin{aligned} p_{1|0}(x_1) =& \int p_0(x_0) q(x_0, x_1) dx_0 \\ \geq& \int_{C_0} p_0(x_0) q(x_0, x_1) dx_0 \\ \geq& \xi_1^{-1} \int_{C_0} p_0(x_0) dx_0 \\ \geq& \xi_1^{-1} \kappa. \end{aligned}$$

□

The next lemma enables us to restrict the integral of $g_i \beta_{i,s}$ on C_i.

Lemma 8. *Under (A1), we have for all* $1 \leq i \leq s$*, and all* y_1^s,

$$\frac{\int_{C_i} g_i(x_i) \beta_{i,s}(x_i) dx_i}{\int g_i(x_i) \beta_{i,s}(x_i) dx_i} \geq \kappa. \tag{6.1}$$

Proof of Lemma 8. When $i = s$, we have $\beta_{i,s} \equiv 1$, and the result follows easily from (A1).

When $i \leq s - 1$, by (A1),

$$\begin{aligned} \int_{C_i} g_i(x_i) \beta_{i,s}(x_i) dx_i =& \int_{C_i} g_i(x_i) \int q(x_i, x_{i+1}) g_{i+1}(x_{i+1}) \beta_{i+1,s}(x_{i+1}) dx_{i+1} dx_i \\ =& \int \int_{C_i} g_i(x_i) q(x_i, x_{i+1}) dx_i g_{i+1}(x_{i+1}) \beta_{i+1,s}(x_{i+1}) dx_{i+1} \\ \geq& \kappa \int \int g_i(x_i) q(x_i, x_{i+1}) dx_i g_{i+1}(x_{i+1}) \beta_{i+1,s}(x_{i+1}) dx_{i+1} \\ =& \kappa \int g_i(x_i) \beta_{i,s}(x_i) dx_i. \end{aligned}$$

□

With Lemma 3 and Lemma 8, we can cancel out $g_i \beta_{i,s}$ in Equation (3.2).

Proof of Lemma 4. First consider the Monte Carlo approximation error for a random sample $\{x_{i-1}^j\}_{j=1}^n$ from $\hat{p}_{i-1|i-1}$:

$$\begin{aligned} & E_{i-1} \left| \hat{Q} \hat{p}_{i-1|i-1}(x_i) - Q \hat{p}_{i-1|i-1}(x_i) \right| \\ \leq & \left(E_{i-1} \left| \hat{Q} \hat{p}_{i-1|i-1}(x_i) - Q \hat{p}_{i-1|i-1}(x_i) \right|^2 \right)^{1/2} \end{aligned}$$

$$=\left(E_{i-1}\left|n^{-1}\sum_{j=1}^{n}q(x_{i-1}^{j},x_i)-\int\hat{p}_{i-1|i-1}(x_{i-1})q(x_{i-1},x_i)dx_{i-1}\right|^2\right)^{1/2}$$
$$\leq n^{-1}M^2.$$

Combined with (3.2), we have

$$\begin{aligned}E_{i-1}\Delta_{i,s}\leq&\frac{2\int E_{i-1}\left|\hat{Q}\hat{p}_{i-1|i-1}(x_i)-Q\hat{p}_{i-1|i-1}(x_i)\right|g_i(x_i)\beta_{i,s}(x_i)dx_i}{\int Q\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)dx_i}\\ \leq&\frac{2M\int g_i(x_i)\beta_{i,s}(x_i)dx_i}{\sqrt{n}\int Q\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)dx_i}\\ \leq&\frac{2M\int g_i(x_i)\beta_{i,s}(x_i)dx_i}{\sqrt{n}\int_{C_i}Q\hat{p}_{i-1|i-1}(x_i)g_i(x_i)\beta_{i,s}(x_i)dx_i}\\ \leq&\frac{2M\xi_i\int g_i(x_i)\beta_{i,s}(x_i)dx_i}{\sqrt{n}\kappa\int_{C_i}g_i(x_i)\beta_{i,s}(x_i)dx_i}\\ \leq&\frac{2M\xi_i}{\sqrt{n}\kappa^2}.\end{aligned}$$

By definition $||\Delta_{i,s}||_{\text{tv}}\leq 2$. As a result, $E_{i-1}\Delta_{i,s}\leq 2(1\wedge M\kappa^{-2}\xi_i/\sqrt{n})$. □

6.2 Proof of Theorem 5

The proof of Theorem 5 consists of three parts:

1. Verify condition (A1) to enable application of Lemma 4. This is to be done in Lemma 9.
2. Establish uniform contracting property for the conditional Markov operators $F_{i|s}$. This is an existing result from Douc, Moulines, and Ritov (2009).
3. Control $E(1\wedge M\kappa^{-2}\xi_i/\sqrt{n})$.

With these three components, Theorem 5 follows immediately from (3.6). In the following we give detailed arguments for each of these components.

Part I: verify condition (A1). To verify condition (A1), we first specify C_y:

$$C_y:=\{x:|x-b^{-1}(y)|\leq D\},$$

with a constant D to be chosen later with

$$\inf_{[0,D]} p_V > 0. \tag{6.2}$$

Under these conditions, one can show the following, and hence verify (A1).

Lemma 9. *Assuming (C1)-(C4), then for each y the C_y's defined above satisfy*

$$\min\left\{\frac{\int_{C_y} q(x,x')g(y;x')dx'}{\int q(x,x')g(y;x')dx'},\ \frac{\int_{C_y} g(y;x)q(x,x')dx}{\int g(y;x)q(x,x')dx},\ \frac{\int_{C_y} g(y;x)dx}{\int g(y;x)dx}\right\}$$
$$\geq \kappa > 0,$$

for some $\kappa \in (0,1)$, independent of (x,x',y).

The proof is just a minor modification of Douc, Moulines, and Ritov (2009, Lemmas 11,12). We postpone it to the end of this section.

Now we have verified all the conditions necessary to apply Lemma 4. Next we develop a bound for $E\left(1 \wedge M\kappa^{-2}\xi_i/\sqrt{n}\right)$.

Part II: Control $E(1\wedge M\kappa^{-2}\xi_i/\sqrt{n})$. Under the autoregressive model (4.1), using (4.2) repeatedly, we have for $i \geq 2$

$$\begin{aligned}
\xi_i &= \sup_{C_{i-1}\times C_i} p_U^{-1}\left(x_i - a(x_{i-1})\right) \\
&= \sup_{C_{i-1}\times C_i} p_U^{-1}\big(x_i - b^{-1}(Y_i) + b^{-1}(Y_i) \\
&\qquad\qquad -a(b^{-1}(Y_{i-1})) + a(b^{-1}(Y_{i-1})) - a(x_{i-1})\big) \\
&\leq c p_U^{-1}(b^{-1}(Y_i) - a(b^{-1}(Y_{i-1}))) \\
&\leq c p_U^{-1}(b^{-1}(Y_i) - X_i + X_i - a(X_{i-1}) + a(X_{i-1}) - a(b^{-1}(Y_{i-1}))) \\
&\leq c p_U^{-1}(b_-^{-1}V_i) p_U^{-1}(U_i) p_U^{-1}(a_+ b_-^{-1}(V_{i-1})),
\end{aligned} \tag{6.3}$$

noting again that the constant c may take different values in different displays.

Therefore, for any $\theta > 0$,

$$\begin{aligned}
&E\left(1 \bigwedge \frac{M}{\kappa^2\sqrt{n}}\xi_i \,\middle|\, V_i, V_{i-1}\right) \\
\leq& \frac{c}{\sqrt{n}} p_U^{-1}(b_-^{-1}V_i) p_U^{-1}(a_+ b_-^{-1}V_{i-1}) \int_{-\theta}^{\theta} p_U^{-1}(u) p_U(u) du + \int_{[-\theta,\theta]^c} p_U(u) du
\end{aligned}$$

$$=\frac{c\theta}{\sqrt{n}}p_U^{-1}(b_-^{-1}V_i)p_U^{-1}(a_+b_-^{-1}V_{i-1})+P(|U_1|>\theta). \tag{6.4}$$

Equation (4.3) in condition (C3) ensures that $E[p_U^{-1}(b_-^{-1}V_i)p_U^{-1}(a_+b_-^{-1}V_{i-1})]$ is a finite constant.

The case $i=1$ is similar. Actually (6.4) still holds by realizing that

$$\begin{aligned}\xi_1 =& \sup_{(x_0,x_1)\in C_0\times C_1} p_U^{-1}(x_1-a(x_0)) \\ \leq& \sup_{C_0\times C_1} p_U^{-1}(b_-^{-1}V_1)p_U^{-1}(U_1)p_U^{-1}(a_+b_-^{-1}|X_0|).\end{aligned} \tag{6.5}$$

As a result, we obtain a bound on the expected one step propagated sampling error:

$$E\left(1\bigwedge\frac{M}{\kappa^2\sqrt{n}}\xi_i\right)\leq\frac{c\theta}{\sqrt{n}}+P(|U_1|>\theta),$$

for some constant c depending only on the model.

Part III: Uniform contracting property of $F_{i|s}$. The uniform contracting property of $F_{i|s}$ under this setting has been established by Douc, Moulines, and Ritov (2009). We state it without proof:

Lemma 10. *Under Model (4.1), assuming (C1)-(C4), then*

$$\delta(F_{i+1|s}F_{i|s})\leq\rho<1,\quad\forall\ 1\leq i\leq s-1,$$

for some constant ρ depending only on the model.

In the end we give the proof of Lemma 9, which follows largely from Douc, Moulines, and Ritov (2009, Lemmas 10, 11 and 12), with small modifications.

Lemma 11 (Lemma 10 of Douc, Moulines, and Ritov (2009)). *Assume diam$(C)<\infty$. Then for all $x\in C$ and $x'\in\mathcal{X}$,*

$$\rho(C)h_C(x')\leq q(x,x')\leq\rho^{-1}(C)h_C(x'), \tag{6.6}$$

with

$$\begin{aligned}\rho(C)&=rp_U(\mathrm{diam}(C))\wedge\inf_{|u|\leq\mathrm{diam}(C)}p_U\wedge\left(\sup_{|u|\leq\mathrm{diam}(C)}p_U\right)^{-1},\\ h_C(x')&=\mathbb{1}\left(x'\in a(C)\right)+\mathbb{1}\left(x'\notin a(C)\right)p_U(|x'-a(z_0)|),\end{aligned}$$

where r is defined in (4.2), and z_0 is an arbitrary element of C. In addition, for all $x \in \mathcal{X}$ and $x' \in C$,

$$\nu(C)k_C(x) \leq q(x, x'), \tag{6.7}$$

with

$$\begin{aligned} \nu(C) &= \inf_{|u|\leq \operatorname{diam}(C)} p_U, \\ k_C(x) &= \mathbb{1}\left(a(x) \in C\right) + r\mathbb{1}\left(a(x) \notin C\right) p_U(|z' - a(x)|), \end{aligned}$$

where z' is an arbitrary element in C.

Proof of Lemma 9. Recall that $C_y = \{x : |x - b^{-1}(y)| \leq D\}$, for some $D > 0$.

We first show

$$\inf_y \frac{\int_{C_y} g(y; x)dx}{\int g(y; x)dx} > 0.$$

In fact, we have $g(y; x) = p_V(y - b(x))$, and then

$$\int_{C_y^c} p_V(y - b(x))dx \leq \int_{C_y^c} p_V\left(b_-|b^{-1}(y) - x|\right)dx \leq \int_{|x|\geq D} p_V(b_-|x|)dx,$$

which is independent of y. Also note that $\int p_V(y - b(x))dx$ is bounded from below uniformly in y by change of variables and the assumption that $|b'|$ is bounded and bounded away from zero. Now we can choose D large enough so that $\int_{|x|\geq D} p_V(b_-|x|)dx < \inf_y \int p_V(y - b(x))dx$, and hence

$$\inf_y \int_{C_y} g(y; x)dx > 0.$$

Then we are going to show

$$\inf_{y,x} \frac{\int_{C_y} q(x, x')g(y; x')dx'}{\int q(x, x')g(y; x')dx'} > 0, \tag{6.8}$$

which is equivalent to

$$\inf_{y,x} \frac{\int_{C_y} q(x, x')g(y; x')dx'}{\int_{C_y^c} q(x, x')g(y; x')dx'} > 0.$$

Note that in Lemma 11 the constants $\rho(C_y)$ and $\nu(C_y)$ depend on C_y only through its diameter and hence are independent of y. In the following argument we drop the dependence on y when using these notations.

Consider two cases.

1. $a(x) \in C_y$.

   In this case $k_{C_y}(x) \equiv 1$ as defined in Lemma 11. We have

$$\frac{\int_{C_y} q(x,x')g(y;x')dx'}{\int q(x,x')g(y;x')dx'} \geq \frac{\nu}{M}\frac{\int_{C_y} g(y;x')dx'}{\int g(y;x')dx'} \geq \frac{\nu}{M}\inf_y \frac{\int_{C_y} g(y;x')dx'}{\int g(y;x')dx'} > 0, \tag{6.9}$$

   where we used the fact $q(x,x') \leq ||p_U||_\infty \leq M$.

2. $a(x) \notin C_y$.

   In this case $k_{C_y}(x) = rp_U(|b^{-1}(y) - a(x)|)$ as defined in Lemma 11, where z' is chosen as $b^{-1}(y)$. In (4.2) let $w = x' - a(x)$, $w' = b^{-1}(y) - x'$, so by monotonicity of p_U, (4.2), we have

$$p_U(|w + w'|) \geq p_U(|w| + |w'|) \geq rp_U(|w|)p_U(|w'|),$$

   which implies

$$\frac{p_U(|x' - a(x)|)}{p_U(|b^{-1}(y) - a(x)|)} \leq r^{-1}p_U^{-1}(|b^{-1}(y) - x'|).$$

   Therefore, using (4.3),

$$\begin{aligned}
\frac{\int_{C_y} q(x,x')g(y;x')dx'}{\int_{C_y^c} q(x,x')g(y;x')dx'} &\geq \frac{r\nu p_U(|b^{-1}(y) - a(x)|)\int_{C_y} g(y;x')dx'}{\int_{C_y^c} p_U(|x' - a(x)|)p_V(y - b(x'))dx'} \\
&\geq \frac{r^2\nu \int_{C_y} g(y;x')dx'}{\int_{C_y^c} p_U^{-1}(|b^{-1}(y) - x'|)p_V(y - b(x'))dx'} \\
&\geq \frac{r^2\nu \int_{C_y} g(y;x')dx'}{\int_{C_y^c} p_U^{-1}(|b^{-1}(y) - x'|)p_V(b_-|b^{-1}(y) - x'|)dx'} \\
&\geq \frac{r^2\nu \int_{C_y} g(y;x')dx'}{\int_{|z|\geq D} p_U^{-1}(|z|)p_V(b_-|z|)dz} \\
&\geq \frac{r^2\nu \inf_y \int_{C_y} g(y;x')dx'}{\int_{|z|\geq D} p_U^{-1}(|z|)p_V(b_-|z|)dz} > 0,
\end{aligned} \tag{6.10}$$

   where the last inequality is based on (4.3) of condition (C3). Note that the bounds of both (6.9) and (6.10) are independent of y and x. As a result (6.8) is true.

It remains to show

$$\inf_{y,x'} \frac{\int_{C_y} g(y;x)q(x,x')dx}{\int g(y;x)q(x,x')dx} > 0. \tag{6.11}$$

The argument is very similar to that of (6.8). Again, consider two cases.

1. $x' \in a(C_y)$.

   In this case $h_{C_y}(x') \equiv 1$, and we have

$$\frac{\int_{C_y} g(y;x)q(x,x')dx}{\int g(y;x)q(x,x')dx} \geq \frac{\rho \int_{C_y} g(y;x)dx}{M \int g(y;x)dx} \geq \frac{\rho}{M} \inf_y \frac{\int_{C_y} g(y;x)dx}{\int g(y;x)dx} > 0. \tag{6.12}$$

2. $x' \notin a(C_y)$.

   In this case $h_{C_y}(x') = p_U(|x' - a(b^{-1}(y))|)$, choosing $z_0 = b^{-1}(y)$ in Lemma 11, and

$$\frac{\int_{C_y} g(y;x)q(x,x')dx}{\int_{C_y^c} g(y;x)q(x,x')dx} \geq \frac{\rho p_U(|x' - a(b^{-1}(y))|) \int_{C_y} g(y;x)dx}{\int_{C_y^c} p_V(y - b(x))p_U(x' - a(x))dx}. \tag{6.13}$$

   It suffices to show

$$\int_{C_y^c} p_U^{-1}(|x' - a(b^{-1}(y))|)p_U(x' - a(x))p_V(y - b(x))dx$$

   is bounded uniformly for all y and x'.

   Again, let $w = x' - a(x)$, $w' = a(x) - a(b^{-1}(y))$. Then $|w + w'| = |x' - a(b^{-1}(y))| > L$. Therefore,

$$p_U(|w + w'|) \geq p_U(|w| + |w'|) \geq r p_U(|w|)p_U(|w'|),$$

   which implies

$$p_U^{-1}(|x' - a(b^{-1}(y))|)p_U(x' - a(x)) \leq r^{-1} p_U^{-1}(|a(x) - a(b^{-1}(y))|).$$

   Also note that for all $z, z' \in \mathcal{X}$,

$$p_U^{-1}(|a(z) - a(z')|) \leq p_U^{-1}(a_+|z - z'|).$$

   As a result,

$$\int_{C_y^c} p_U^{-1}(|x' - a(b^{-1}(y))|)p_U(x' - a(x))p_V(y - b(x))dx \tag{6.14}$$

$$\begin{aligned}
&\leq \int_{C_y^c} r^{-1} p_U^{-1}(|a(x) - a(b^{-1}(y))|) p_V(y - b(x)) dx \\
&\leq r^{-1} \int_{C_y^c} p_U^{-1}(a_+|x - b^{-1}(y)|) p_V(y - b(x)) dx \\
&\leq r^{-1} \int_{|x|>D} p_U^{-1}(a_+|x|) p_V(b_- x) dx \\
&< \infty,
\end{aligned}$$

where the last inequality uses (4.3). Therefore, (6.11) is true because the bounds in (6.12) and (6.14) do not depend on y or x'.

□

Acknowledgment Lei and Bickel were supported by NSF DMS-0906808. The authors would like to thank an anonymous reviewer for comments and suggestions on an early version of this paper.

References

Bickel, P., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **26**, 1614–1635.

Cappé, O. and Moulines, E. (2005). On the use of particle filtering for maximum likelihood parameter estimation. In proceedings of *European Signal Processing Conference.*

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

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

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

Del Moral, P. and Jacod, J. (2001). Interacting particle filtering with discrete-time observations: asymptotic behaviour in the Gaussian case. *Trends in Mathematics* (T. Hida, R. Krandikar, H. Kunita, B. Rajput, S. Watanabe, and J. Xiong, Eds.) 101–123. Birkhäuser.

Douc, R., Moulines, E., and Ritov, Y. (2009). Forgetting of the initial condition for the filter in general state-space hidden Markov chain: a coupling approach. *Electron. J. Probab.* **14**, 27–49.

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

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

Heine, K. and Crisan, D. (2008). Uniform approximations of discrete-time filters. *Adv. in Appl. Probab.* **40**, 979–1001.

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

Künsch, H. R. (2005). Recursive Monte Carlo filters: algorithms and theoretical analysis. *Ann. Statist.* **33**, 1983–2021.

Le Gland, F. and Oudjane, N. (2003). A robustification approach to stability and to uniform particle approximation of nonlinear filters: the example of pseudo-mixing signals. *Stochastic Process. Appl.* **106**, 279–316.

Le Gland, F. and Oudjane, N. (2004). Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann. Appl. Probab.* **14**, 144–187.

Lin, M. T., Zhang, J. L., Cheng, Q., and Chen, R. (2005). Independent particle filters. *J. Amer. Statist. Assoc.* **100**, 1412–1421.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

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

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas, and N. Gordon, Eds.) 197–217. Springer-Verlag.

Olsson, J. and Rydén, T. (2008). Asymptotic properties of particle filter-based maximum likelihood estimators for state space models. *Stochastic Process. Appl.* **118**, 649–680.

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

Polson, N. G., Stroud, J. R., and Müller, P. (2008). Practical filtering with sequential parameter learning. *J. R. Statist. Soc.* B **70**, 413–428.

van Handel, R. (2009). Uniform time average consistency of Monte Carlo particle filters. *Stochastic Process. Appl.* **119**, 3835–3861.

132 Baker Hall, Department of Statistics
Carnegie Mellon University
Pittsburgh, PA, 15213
E-mail: jinglei@andrew.cmu.edu

367 Evans Hall, Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720
E-mail: bickel@stat.berkeley.edu