The Annals of Applied Probability
2004, Vol. 14, No. 1, 144–187

STABILITY AND UNIFORM APPROXIMATION OF NONLINEAR FILTERS USING THE HILBERT METRIC AND APPLICATION TO PARTICLE FILTERS[1]

BY FRANÇOIS LE GLAND AND NADIA OUDJANE

IRISA/INRIA Rennes and EDF R&D Clamart

We study the stability of the optimal filter w.r.t. its initial condition and w.r.t. the model for the hidden state and the observations in a general hidden Markov model, using the Hilbert projective metric. These stability results are then used to prove, under some mixing assumption, the uniform convergence to the optimal filter of several particle filters, such as the interacting particle filter and some other original particle filters.

1. Introduction. The stability of the optimal filter has recently become an active research area. Ocone and Pardoux proved [26] that the filter forgets its initial condition in the L^p sense, without stating any rate of convergence. Recently, a new approach has been proposed using the Hilbert projective metric. This metric allows getting rid of the normalization constant in the Bayes formula and reduces the problem to studying the linear equation satisfied by the unnormalized optimal filter. Using the Hilbert metric, stability results w.r.t. the initial condition have been proved by Atar and Zeitouni [4], and some stability result w.r.t. the model have been proved by Le Gland and Mevel [19, 20], for hidden Markov models (HMM) with finite state space. The results and methods of [4] have been extended to HMM with Polish state space by Atar and Zeitouni [3]; see also [8]. Independently, Del Moral and Guionnet have adopted in [9], for the same class of HMM, another approach based on semigroup techniques and on the Dobrushin ergodic coefficient, to derive stability results w.r.t. the initial condition, which are used to prove uniform convergence of the interacting particle system (IPS) approximation to the optimal predictor. New approaches have been proposed recently, to prove the stability of the optimal filter w.r.t. its initial condition, in the case of a noncompact state space (see, e.g., [1, 2, 6, 7]).

In this article, we use the approach based on the Hilbert metric to study the asymptotic behavior of the optimal filter, and to prove as in [9] the uniform

Received July 2001; revised January 2003.

[1]Supported in part by CNRS projects Méthodes Particulaires en Filtrage Non-Linéaire (97–N23/0019, Modélisation et Simulation Numérique programme), Chaînes de Markov Cachées et Filtrage Particulaire (MathSTIC programme) and Méthodes Particulaires (AS67, DSTIC Action Spécifique programme).

AMS 2000 subject classifications. Primary 93E11, 93E15, 62E25; secondary 60B10, 60J27, 62G07, 62G09, 62L10.

Key words and phrases. Hidden Markov model, nonlinear filter, particle filter, stability, Hilbert metric, total variation norm, mixing, regularizing kernel.

convergence of several particle filters, such as the interacting particle filter (IPF) and some other original particle filters.

A common assumption to prove stability results (see, e.g., [9], Theorem 2.4) is that the Markov transition kernels are mixing, which implies that the hidden state sequence is ergodic. Our results are obtained under the assumption that the nonnegative kernels describing the evolution of the unnormalized optimal filter, and incorporating simultaneously the Markov transition kernels and the likelihood functions, are mixing. This is a weaker assumption (see Proposition 3.9), which allows considering some cases, similar to the case studied in [6], where the hidden state sequence is not ergodic; see Example 3.10. This point of view is further developped by Le Gland and Oudjane [22] and by Oudjane and Rubenthaler [28]. Our main contribution is to study also the stability of the optimal filter w.r.t. the model, when the local error is propagated by mixing kernels and can be estimated in the Hilbert metric, in the total variation norm, or in a weaker distance suitable for random probability distributions.

The uniform convergence of the IPS approximation to the optimal predictor is proved in ([9], Theorem 3.1), under the assumption that the likelihood functions are uniformly bounded away from zero, which is rather strong, and that the predictor is asymptotically stable. The rate $(1/\sqrt{N})^{\alpha}$ for some $\alpha < 1$ is proved under the stronger assumption that the predictor is exponentially asymptotically stable, and the rate $1/\sqrt{N}$ is proved by Del Moral and Miclo ([11], page 36) under an additional assumption which is satisfied, for example, if the Markov kernels are mixing. Our uniform convergence results are obtained under the assumption that the expected values of the likelihood functions integrated against any possible predicted probability distribution are bounded away from zero. This assumption is automatically satisfied under our weaker mixing assumption; see Remark 5.7.

Motivated by practical considerations, we introduce a variant of the IPF, where an adaptive number of particles is used, based on a posteriori estimates. The resulting sequential particle filter (SPF) is shown to converge uniformly to the optimal filter, independently of any lower bound assumption on the likelihood functions. The counterpart is that the computational time is random and that the expected number of particles does depend on the integrated lower bounds of the likelihood functions. Also motivated by practical considerations, that is, to avoid the *degeneracy of particle weights* and the *degeneracy of particle locations*, which are two known causes of divergence of particle filters, we introduce regularized particle filters (RPF), which are shown to converge uniformly to the optimal filter.

The paper is organized as follows: In the next section we define the framework of the nonlinear filtering problem and we introduce some notation. In Section 3, we state some properties of the Hilbert metric, which are used in Section 4 to prove the stability of the optimal filter w.r.t. its initial condition and w.r.t. the model. These stability results are used to prove the uniform convergence of several particle filters to the optimal filter. First, uniform convergence in the weak sense is proved in

Section 5 for interacting particle filters, with a rate $1/\sqrt{N}$, and sequential particle filters with a random number of particles are also considered. Finally, regularized particle filters are defined in Section 6, for which uniform convergence in the weak sense and in the total variation norm are proved.

2. Optimal filter for general HMM. We consider the following model, with a hidden (nonobserved) state sequence $\{X_n, n \geq 0\}$ and an observation sequence $\{Y_n, n \geq 1\}$, taking values in a complete separable metric space E and in $F = \mathbb{R}^d$, respectively (in Section 6, it will be assumed that $E = \mathbb{R}^m$).

- The state sequence $\{X_n, n \geq 0\}$ is defined as an inhomogeneous Markov chain, with transition probability kernel Q_n; that is,

$$\mathbb{P}[X_n \in dx | X_{0:n-1} = x_{0:n-1}] = \mathbb{P}[X_n \in dx | X_{n-1} = x_{n-1}] = Q_n(x_{n-1}, dx),$$

  for all $n \geq 1$, and with initial probability distribution μ_0. For instance, $\{X_n, n \geq 0\}$ could be defined by the following equation:

$$X_n = f_n(X_{n-1}, W_n), \tag{1}$$

  where $\{W_n, n \geq 1\}$ is a sequence of independent random variables, not necessarily Gaussian, independent of the initial state X_0.
- The *memoryless channel* assumption holds; that is, given the state sequence $\{X_n, n \geq 0\}$,

  the observations $\{Y_n, n \geq 1\}$ are independent random variables, and for all $n \geq 1$, the conditional probability distribution of Y_n depends only on X_n.

  For instance, the observation sequence $\{Y_n, n \geq 1\}$ could be related to the state sequence $\{X_n, n \geq 0\}$ by

$$Y_n = h_n(X_n, V_n)$$

  for all $n \geq 1$, where $\{V_n, n \geq 1\}$ is a sequence of independent random variables, not necessarily Gaussian, independent of the state sequence $\{X_n, n \geq 0\}$. In addition, it is assumed that for all $n \geq 1$, the collection of probability distributions $\mathbb{P}[Y_n \in dy | X_n = x]$ on F, parametrized by $x \in E$, is dominated, that is,

$$\mathbb{P}[Y_n \in dy | X_n = x] = g_n(x, y)\lambda_n^F(dy)$$

  for some nonnegative measure λ_n^F on F. The corresponding likelihood function is defined by $\Psi_n(x) = g_n(x, Y_n)$, and depends implicitly on the observation Y_n.

  The following notation and definitions will be used throughout the paper:

- The set of probability distributions on E and the set of finite nonnegative measures on E are denoted by $\mathcal{P}(E)$ and $\mathcal{M}^+(E)$, respectively.

- The notation $\|\cdot\|$ is used for the total variation norm on the set of signed measures on E and for the supremum norm on the set of bounded measurable functions defined on E, depending on the context.
- With any nonnegative kernel K defined on E is associated a nonnegative linear operator denoted by K and defined by

$$K\mu(dx') = \int_E \mu(dx) K(x, dx')$$

for any nonnegative measure $\mu \in \mathcal{M}^+(E)$.
- With any nonnegative measure $\mu \in \mathcal{M}^+(E)$ is associated the normalized nonnegative measure (i.e., the probability distribution),

$$\bar{\mu} := \begin{cases} \dfrac{\mu}{\mu(E)}, & \text{if } \mu(E) > 0, \text{ that is, if } \mu \text{ is nonzero,} \\ \nu, & \text{otherwise, that is, if } \mu \equiv 0, \end{cases}$$

where $\nu \in \mathcal{P}(E)$ is an arbitrary probability distribution.
- With any nonnegative kernel K defined on E is associated the normalized nonnegative nonlinear operator $\bar{K}$, taking values in $\mathcal{P}(E)$, and defined for any nonzero nonnegative measure $\mu \in \mathcal{M}^+(E)$ by

$$\bar{K}(\mu) := \begin{cases} \dfrac{K\mu}{(K\mu)(E)}, & \text{if } (K\mu)(E) > 0, \text{ that is, if } K\mu \text{ is nonzero,} \\ \nu, & \text{otherwise, that is, if } K\mu \equiv 0, \end{cases}$$

where $\nu \in \mathcal{P}(E)$ is an arbitrary probability distribution. Notice that $\bar{K}(\mu) = \bar{K}(\bar{\mu})$ is nonzero by definition, hence composition of normalized nonnegative nonlinear operators is well defined.

The problem of nonlinear filtering is to compute at each time n, the conditional probability distribution μ_n of the state X_n given the observation sequence $Y_{1:n} = (Y_1, \ldots, Y_n)$ up to time n. The transition from μ_{n-1} to μ_n is described by the following diagram:

$$\mu_{n-1} \underset{\text{prediction}}{\longrightarrow} \mu_{n|n-1} = Q_n \mu_{n-1} \underset{\text{correction}}{\longrightarrow} \mu_n = \Psi_n \cdot \mu_{n|n-1} = \frac{\Psi_n \mu_{n|n-1}}{\langle \mu_{n|n-1}, \Psi_n \rangle},$$

where "$\cdot$" denotes the projective product. In general, no explicit expression is available for the Markov kernel Q_n, or it is so complicated that computing integrals such as

$$\mu_{n|n-1}(dx') = Q_n \mu_{n-1}(dx') = \int_E \mu_{n-1}(dx) Q_n(x, dx')$$

is practically impossible. Instead, throughout this paper we assume that for any $x \in E$, simulating a r.v. with probability distribution $Q_n(x, dx')$ is easy [this is the case, e.g., if (1) holds].

REMARK 2.1. Notice that the normalizing constant $\langle \mu_{n|n-1}, \Psi_n \rangle$ is a.s. positive; hence the projective product $\Psi_n \cdot \mu_{n|n-1}$ is well defined. Indeed,

$$\begin{aligned}\mathbb{P}[Y_n \in dy | Y_{1:n-1}] &= \int_E \mathbb{P}[Y_n \in dy | X_n = x] \mathbb{P}[X_n \in dx | Y_{1:n-1}] \\ &= \left[\int_E g_n(x, y) \mu_{n|n-1}(dx) \right] \lambda_n^F(dy) = \ell_n(y) \lambda_n^F(dy),\end{aligned}$$

hence

$$\langle \mu_{n|n-1}, \Psi_n \rangle = \int_E g_n(x, Y_n)\, \mu_{n|n-1}(dx) = \ell_n(Y_n).$$

Therefore

$$\mathbb{P}[\langle \mu_{n|n-1}, \Psi_n \rangle = 0 | Y_{1:n-1}] = \int_F \mathbf{1}_{\{\ell_n(y)=0\}} \ell_n(y) \lambda_n^F(dy) = 0.$$

REMARK 2.2. Notice also that, for any test function ψ defined on F,

$$\mathbb{E}\left[\frac{\psi(Y_n)}{\langle \mu_{n|n-1}, \Psi_n \rangle} \middle| Y_{1:n-1} \right] = \mathbb{E}\left[\frac{\psi(Y_n)}{\ell_n(Y_n)} \middle| Y_{1:n-1} \right] = \int_F \psi(y) \lambda_n^F(dy).$$

In particular, if $\psi(y) = g_n(x, y)$, then $\psi(Y_n) = \Psi_n(x)$, and

$$\mathbb{E}\left[\frac{\Psi_n(x)}{\langle \mu_{n|n-1}, \Psi_n \rangle} \middle| Y_{1:n-1} \right] = \int_F g_n(x, y) \lambda_n^F(dy) = 1,$$

for any $x \in E$.

For any $n \geq 1$, we introduce the nonnegative kernel

$$R_n(x, dx') = Q_n(x, dx') \Psi_n(x') \tag{2}$$

and the associated nonnegative linear operator $R_n = \Psi_n Q_n$ on $\mathcal{M}^+(E)$, defined by

$$R_n \mu(dx') = \int_E \mu(dx) Q_n(x, dx') \Psi_n(x')$$

for any $\mu \in \mathcal{M}^+(E)$. Notice that R_n depends on the observation Y_n through the likelihood function Ψ_n. With this definition, $(R_n \mu_{n-1})(E) = \langle \mu_{n|n-1}, \Psi_n \rangle$ is a.s. positive, and the evolution of the optimal filter can be written as follows:

$$\mu_n = \Psi_n \cdot (Q_n \mu_{n-1}) = \frac{R_n \mu_{n-1}}{(R_n \mu_{n-1})(E)} = \bar{R}_n(\mu_{n-1}), \tag{3}$$

and iteration yields

$$\mu_n = \bar{R}_n(\mu_{n-1}) = \bar{R}_n \circ \cdots \circ \bar{R}_m(\mu_{m-1}) = \bar{R}_{n:m}(\mu_{m-1}).$$

Equation (3) shows clearly that the evolution of the optimal filter is nonlinear only because of the normalization term coming from the Bayes rule. In the following section a projective metric is introduced precisely to get rid of the normalization and to come down to the analysis of a linear evolution.

REMARK 2.3. The model considered here is slightly different from the model considered in other works (see [11] and references therein), where it is assumed that an observation Y_0 is already available at time 0, and where the object of study is rather the conditional probability distribution η_n of the state X_n given the observation sequence $Y_{0:n-1} = (Y_0, \ldots, Y_{n-1})$ up to time $(n-1)$. With our notation, the evolution of the optimal predictor in this alternate model can be written as follows:

$$\eta_{n+1} = Q_{n+1}(\Psi_n \cdot \eta_n),$$

and iteration yields

$$\begin{aligned} \eta_{n+1} &= Q_{n+1}\bar{R}_n \circ \cdots \circ \bar{R}_m(\Psi_{m-1} \cdot \eta_{m-1}) \\ &= Q_{n+1}\bar{R}_{n:m}(\Psi_{m-1} \cdot \eta_{m-1}) \\ &= Q_{n+1}\hat{\eta}_n, \end{aligned} \tag{4}$$

with initial condition $\eta_0 = \mu_0$.

3. Hilbert metric on the set of finite nonnegative measures. In this section we recall the definition of the Hilbert metric and its associated contraction coefficient, the Birkhoff contraction coefficient. We introduce also a mixing property for nonnegative kernels, and we state some properties relating the Hilbert metric with other distances on the set of probability distributions, for instance, the total variation norm, or a weaker distance suitable for random probability distributions. In the last part of the section, these definitions and properties are specialized to the optimal filtering context.

DEFINITION 3.1. Two nonnegative measures $\mu, \mu' \in \mathcal{M}^+(E)$ are *comparable*, if they are both nonzero and if there exist positive constants $0 < a \leq b$, such that

$$a\mu'(A) \leq \mu(A) \leq b\mu'(A)$$

for any Borel subset $A \subset E$.

DEFINITION 3.2. The nonnegative kernel K defined on E is *mixing*, if there exist a constant $0 < \varepsilon \leq 1$ and a nonnegative measure $\lambda \in \mathcal{M}^+(E)$, such that

$$\varepsilon\lambda(A) \leq K(x, A) \leq \frac{1}{\varepsilon}\lambda(A)$$

for any $x \in E$, and any Borel subset $A \subset E$.

DEFINITION 3.3. The Hilbert metric on $\mathcal{M}^+(E)$ is defined by

$$h(\mu, \mu') := \begin{cases} \log \dfrac{\sup_{A:\,\mu'(A)>0} \mu(A)/\mu'(A)}{\inf_{A:\,\mu'(A)>0} \mu(A)/\mu'(A)}, & \text{if } \mu \text{ and } \mu' \text{ are comparable,} \\ 0, & \text{if } \mu = \mu' \equiv 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Notice that the two nonnegative measures μ and μ' are comparable if and only if μ and μ' are equivalent, with Radon–Nikodym derivatives $\frac{d\mu}{d\mu'}$ and $\frac{d\mu'}{d\mu}$ bounded and bounded away from zero, and then the following equality holds:

$$h(\mu,\mu') = \log\Big[\sup_{A:\,\mu'(A)>0}\frac{\mu(A)}{\mu'(A)}\sup_{A:\,\mu(A)>0}\frac{\mu'(A)}{\mu(A)}\Big] = \log\Big(\Big\|\frac{d\mu}{d\mu'}\Big\|\Big\|\frac{d\mu'}{d\mu}\Big\|\Big). \tag{5}$$

Moreover, h is a projective distance; that is, it is invariant under multiplication by positive scalars; hence the Hilbert distance between two unnormalized nonnegative measures is the same as the Hilbert distance between the two corresponding normalized measures: $h(\mu,\mu') = h(\bar\mu,\bar\mu')$, for any nonzero $\mu,\mu' \in \mathcal{M}^+(E)$. In the nonlinear filtering context, this property will allow us to consider the linear transformation $\mu \mapsto R_n\mu$ instead of the nonlinear transformation $\mu \mapsto \bar R_n(\mu) = R_n\mu/(R_n\mu)(E)$. This projective property does not hold for other distances. Indeed, the following estimates show how the error between two unnormalized nonnegative measures can be used to bound the error between the two corresponding normalized measures. If $\mu = \mu' \equiv 0$, then $\bar\mu = \bar\mu' = \nu$, hence $\bar\mu - \bar\mu' \equiv 0$. If both μ and μ' are nonzero, then

$$\bar\mu - \bar\mu' = \frac{1}{\mu(E)}\big[\mu - \mu' - \big(\mu(E) - \mu'(E)\big)\bar\mu'\big],$$

hence

$$|\langle \bar\mu - \bar\mu', \phi\rangle| \le \frac{|\langle \mu - \mu', \phi\rangle|}{\mu(E)} + \frac{|\mu(E) - \mu'(E)|}{\mu(E)}\|\phi\| \tag{6}$$

and

$$\|\bar\mu - \bar\mu'\| \le \frac{\|\mu - \mu'\|}{\mu(E)} + \frac{|\mu(E) - \mu'(E)|}{\mu(E)}. \tag{7}$$

Finally, if μ is nonzero and $\mu' \equiv 0$, then

$$\bar\mu - \bar\mu' = \frac{\mu}{\mu(E)} - \nu,$$

hence

$$|\langle \bar\mu - \bar\mu', \phi\rangle| \le \frac{|\langle \mu, \phi\rangle|}{\mu(E)} + \|\phi\| \quad \text{and} \quad \|\bar\mu - \bar\mu'\| \le 2,$$

that is, estimates (6) and (7) still hold [notice that the bounds in estimates (6) and (7) do not depend on the restarting probability distribution ν].

The following two lemmas give several useful relations between the Hilbert metric, the total variation norm and a weaker distance suitable for random probability distributions.

LEMMA 3.4. *For any* $\mu, \mu' \in \mathcal{M}^+(E)$,

$$\|\bar{\mu} - \bar{\mu}'\| \leq \frac{2}{\log 3} h(\mu, \mu'). \tag{8}$$

If the nonnegative kernel K *defined on* E *is mixing, then for any nonzero* $\mu, \mu' \in \mathcal{M}^+(E)$,

$$h(K\mu, K\mu') \leq \frac{1}{\varepsilon^2} \|\bar{\mu} - \bar{\mu}'\|. \tag{9}$$

PROOF. If $\mu = \mu' \equiv 0$, then $\bar{\mu} = \bar{\mu}' = \nu$ hence $\|\bar{\mu} - \bar{\mu}'\| = 0$, while $h(\mu, \mu') = 0$ by definition. If μ is nonzero and $\mu' \equiv 0$, then $h(\mu, \mu') = \infty$ by definition. Finally, if both μ and μ' are nonzero, the proof of the first inequality can be found in [3]. To prove the second inequality, notice first that, for any comparable $\mu, \mu' \in \mathcal{M}^+(E)$,

$$\begin{aligned} h(\mu, \mu') &= \log \sup_{A\,:\,\mu'(A)>0} \frac{\mu(A)}{\mu'(A)} + \log \sup_{A\,:\,\mu(A)>0} \frac{\mu'(A)}{\mu(A)} \\ &\leq \sup_{A\,:\,\mu'(A)>0} \frac{|\mu(A) - \mu'(A)|}{\mu'(A)} + \sup_{A\,:\,\mu(A)>0} \frac{|\mu(A) - \mu'(A)|}{\mu(A)} \end{aligned}$$

since $\log(1+x) \leq |x|$. In order to apply this bound to $h(K\mu, K\mu') = h(K\bar{\mu}, K\bar{\mu}')$, we notice that $K\mu$ and $K\mu'$ are comparable for any nonzero $\mu, \mu' \in \mathcal{M}^+(E)$, since K is mixing, and we introduce

$$\begin{aligned} \Delta(A) &= \frac{(K\bar{\mu})(A) - (K\bar{\mu}')(A)}{(K\bar{\mu})(A)} \\ &= \int_E (\bar{\mu} - \bar{\mu}')(dx)\Phi(x, A) \\ &= \int_E (\bar{\mu} - \bar{\mu}')^+(dx)\Phi(x, A) - \int_E (\bar{\mu} - \bar{\mu}')^-(dx)\Phi(x, A), \end{aligned}$$

where

$$\Phi(x, A) = \frac{K(x, A)}{(K\bar{\mu})(A)} \leq \frac{1}{\varepsilon^2}$$

for any $x \in E$ and any Borel subset $A \subset E$, using the mixing property. By the Scheffé theorem,

$$\int_E (\bar{\mu} - \bar{\mu}')^+(dx) = \int_E (\bar{\mu} - \bar{\mu}')^-(dx) = \tfrac{1}{2}\|\bar{\mu} - \bar{\mu}'\|,$$

hence if $\Delta(A)$ is positive, then

$$|\Delta(A)| \leq \int_E (\bar{\mu} - \bar{\mu}')^+(dx)\Phi(x, A) \leq \frac{1}{2\varepsilon^2}\|\bar{\mu} - \bar{\mu}'\|,$$

and similarly, if $\Delta(A)$ is negative, then

$$|\Delta(A)| \le \int_E (\bar{\mu} - \bar{\mu}')^-(dx)\Phi(x, A) \le \frac{1}{2\varepsilon^2}\|\bar{\mu} - \bar{\mu}'\|. \qquad \square$$

LEMMA 3.5. *If the nonnegative kernel K defined on E is* dominated, *that is, if there exist a constant $c > 0$ and a nonnegative measure $\lambda \in \mathcal{M}^+(E)$, such that*

$$K(x, A) \le c\lambda(A)$$

for any $x \in E$ and any Borel subset $A \subset E$, then

$$\mathbb{E}\|K\mu - K\mu'\| \le c\lambda(E) \sup_{\phi:\, \|\phi\|=1} \mathbb{E}|\langle \mu - \mu', \phi\rangle|$$

for any $\mu, \mu' \in \mathcal{M}^+(E)$, possibly random.

REMARK 3.6. If the nonnegative kernel K is mixing, then it is dominated, with the same nonnegative measure $\lambda \in \mathcal{M}^+(E)$ and with $c = 1/\varepsilon$.

REMARK 3.7. If in addition the nonnegative kernel K is $\mathcal{F}$-measurable, then the same estimate holds for conditional expectations w.r.t. $\mathcal{F}$, that is,

$$\mathbb{E}[\|K\mu - K\mu'\||\mathcal{F}] \le c\lambda(E) \sup_{\phi:\, \|\phi\|=1} \mathbb{E}[|\langle \mu - \mu', \phi\rangle||\mathcal{F}]. \tag{10}$$

PROOF OF LEMMA 3.5. By definition, if K is dominated, then $K(x, \cdot)$ is absolutely continuous w.r.t. λ, with Radon–Nikodym derivative $k(x, \cdot)$ bounded by c, for any $x \in E$. Therefore, the total variation norm $\|K\mu - K\mu'\|$ can be written as an integral as follows:

$$\|K\mu - K\mu'\| = \int_E \left| \int_E (\mu - \mu')(dx)k(x, x') \right| \lambda(dx'),$$

hence, taking expectation yields

$$\begin{aligned} \mathbb{E}\|K\mu - K\mu'\| &= \int_E \mathbb{E}\left| \int_E (\mu - \mu')(dx)k(x, x') \right| \lambda(dx') \\ &\le \sup_{\phi:\, \|\phi\|=1} \mathbb{E}|\langle \mu - \mu', \phi\rangle| \int_E \left[\sup_{x \in E} k(x, x') \right] \lambda(dx'). \end{aligned} \qquad \square$$

LEMMA 3.8. *The nonnegative kernel K defined on E is a contraction under the Hilbert metric, and*

$$\tau(K) := \sup_{0 < h(\mu,\mu') < \infty} \frac{h(K\mu, K\mu')}{h(\mu, \mu')} = \tanh\left[\frac{1}{4} H(K) \right], \tag{11}$$

where the supremum in

$$H(K) := \sup_{\mu, \mu'} h(K\mu, K\mu'),$$

is over nonzero nonnegative measures: $\tau(K)$ *is called the Birkhoff contraction coefficient*.

The proof can be found in [5], Chapter XVI, Theorem 3, or in [17], Theorem 1. Notice that $H(K) < \infty$ implies $\tau(K) < 1$.

Returning to the filtering problem introduced in Section 2, stability results stated in the following sections will in general require that for any $n \geq 1$, the nonnegative kernel R_n is mixing, that is, there exist a constant $0 < \varepsilon_n \leq 1$ and a nonnegative measure $\lambda_n \in \mathcal{M}^+(E)$, such that

$$\varepsilon_n \lambda_n(A) \leq R_n(x, A) \leq \frac{1}{\varepsilon_n} \lambda_n(A)$$

for any $x \in E$ and any Borel subset $A \subset E$. Notice that in full generality ε_n and λ_n depend on the observation Y_n, hence are random variables.

PROPOSITION 3.9. *The nonnegative kernel R_n defined in* (2) *is a contraction under the Hilbert metric, with Birkhoff contraction coefficient $\tau_n = \tau(R_n) \leq 1$. Moreover*:

(i) *If R_n is mixing, with the possibly random constant ε_n, then*

$$\tau_n \leq \frac{1 - \varepsilon_n^2}{1 + \varepsilon_n^2} < 1.$$

(ii) *If the Markov transition kernel Q_n is mixing, with the nonrandom constant ε_n, then R_n is also mixing, with the same constant ε_n, without any condition on the likelihood function Ψ_n, and*

$$\tau_n \leq \tau(Q_n) \leq \frac{1 - \varepsilon_n^2}{1 + \varepsilon_n^2} < 1.$$

Throughout the paper, for any integers $m \leq n$, the contraction coefficient of the product $R_{n:m} = R_n \cdots R_m$ is denoted by $\tau_{n:m} = \tau(R_{n:m}) \leq \tau_n \cdots \tau_m$ and by convention $\tau_{n:n+1} = \tau_{m-1:m} = 1$.

PROOF OF PROPOSITION 3.9. It follows immediately from Lemma 3.8 that R_n is a contraction under the Hilbert metric.

If R_n is mixing, then for any nonzero $\mu, \mu' \in \mathcal{M}^+(E)$ and any Borel subset $A \subset E$,

$$\varepsilon_n^2 \frac{R_n\mu'(A)}{\mu'(E)} \leq \varepsilon_n \lambda_n(A) \leq \frac{R_n\mu(A)}{\mu(E)} \leq \frac{1}{\varepsilon_n} \lambda_n(A) \leq \frac{1}{\varepsilon_n^2} \frac{R_n\mu'(A)}{\mu'(E)},$$

hence $R_n\mu$ and $R_n\mu'$ are comparable. Using (5) yields

$$H(R_n) = \sup_{\mu,\mu'} h(R_n\mu, R_n\mu') = \sup_{\mu,\mu'} \log \left(\left\| \frac{d(R_n\mu)}{d(R_n\mu')} \right\| \left\| \frac{d(R_n\mu')}{d(R_n\mu)} \right\| \right) \leq \log \frac{1}{\varepsilon_n^4},$$

where the supremum is taken over nonzero nonnegative measures. Then using Lemma 3.8 yields

$$\tau_n = \tau(R_n) = \tanh\left[\frac{1}{4}H(R_n)\right] \le \tanh\left(\log\frac{1}{\varepsilon_n}\right) = \frac{1-\varepsilon_n^2}{1+\varepsilon_n^2} < 1,$$

which ends the proof of (i).

If Q_n is mixing, then $R_n = \Psi_n Q_n$ is also mixing, since

$$\varepsilon_n \int_A \Psi_n(x')\lambda_n(dx') \le R_n(x, A) \le \frac{1}{\varepsilon_n}\int_A \Psi_n(x')\lambda_n(dx')$$

for any $x \in E$ and any Borel subset $A \subset E$, hence for any nonzero $\mu, \mu' \in \mathcal{M}^+(E)$, $R_n\mu$ and $R_n\mu'$ are comparable, with Radon–Nikodym derivative

$$\frac{d(R_n\mu)}{d(R_n\mu')}(x') = \frac{d(Q_n\mu)}{d(Q_n\mu')}(x')\mathbf{1}_{\{\Psi_n(x')>0\}} \le \frac{d(Q_n\mu)}{d(Q_n\mu')}(x')$$

for any $x' \in E$, and similarly with interchanging the role of μ and μ'. Therefore,

$$H(R_n) \le \sup_{\mu,\mu'} \log\left(\left\|\frac{d(Q_n\mu)}{d(Q_n\mu')}\right\|\left\|\frac{d(Q_n\mu')}{d(Q_n\mu)}\right\|\right) = H(Q_n) \le \log\frac{1}{\varepsilon_n^4},$$

where the supremum is taken over nonzero nonnegative measures. Then using Lemma 3.8 again yields $\tau_n = \tau(R_n) \le \tau(Q_n)$. □

The assumption that the nonnegative kernel R_n is mixing is much weaker than the usual assumption that both the Markov kernel Q_n is mixing and the likelihood function Ψ_n is bounded away from zero. Indeed, if the Markov kernel Q_n is mixing, then it follows from (ii) that the nonnegative kernel R_n is mixing, without any assumption on the likelihood function Ψ_n: in particular, the likelihood function could take the zero value, or even be compactly supported. This is not a necessary condition, however, as illustrated by the example below, where the Markov kernel Q_n is not mixing, but the nonnegative kernel R_n is (equivalent, in a sense to be defined below, to) a mixing kernel.

EXAMPLE 3.10. Assume that μ_0 has compact support $C_0 \subset E$ and that for any $n \ge 1$, the function Ψ_n has compact support $C_n \subset E$, and the transition probability kernel Q_n is defined by

$$Q_n(x, dx') = (2\pi)^{-m/2}\exp\{-\tfrac{1}{2}|x' - f_n(x)|^2\}\,dx' = q_n(x, x')\lambda(dx'),$$

where the function f_n is continuous and where

$$\lambda(dx') = (2\pi)^{-m/2}\exp\{-\tfrac{1}{2}|x'|^2\}\,dx'.$$

Clearly, the Markov kernel Q_n is not mixing, but introducing

$$\Delta_{n-1} = \sup_{x\in C_{n-1}} |f_n(x)| \quad \text{and} \quad \Delta_n' = \sup_{x'\in C_n} |x'|,$$

which are both finite a.s., it holds

$$\exp\{-\Delta_{n-1}\Delta_n' - \Delta_{n-1}^2\} \le q_n(x, x') \le \exp\{\Delta_{n-1}\,\Delta_n'\} \tag{12}$$

for any $x \in C_{n-1}$ and any $x' \in C_n$. Define

$$R_n(x, dx') = Q_n(x, dx')\Psi_n(x'),$$

as usual, and

$$\begin{aligned} R_n^\bullet(x, dx') &= \mathbf{1}_{\{x\in C_{n-1}\}} R_n(x, dx') + \mathbf{1}_{\{x\notin C_{n-1}\}}\Psi_n(x')\lambda(dx') \\ &= \big[\mathbf{1}_{\{x\in C_{n-1}\}} q_n(x, x') + \mathbf{1}_{\{x\notin C_{n-1}\}}\big]\Psi_n(x')\lambda(dx'). \end{aligned}$$

Notice first that the sequence $\{\mu_n,\ n \ge 0\}$ defined by (3) satisfies also

$$\mu_n = \frac{R_n^\bullet \mu_{n-1}}{(R_n^\bullet \mu_{n-1})(E)}. \tag{13}$$

Moreover, it follows from (12) that

$$\begin{aligned} &\exp\{-\Delta_{n-1}\Delta_n' - \Delta_{n-1}^2\} \int_A \Psi_n(x'),\, \lambda(dx') \\ &\qquad \le R_n^\bullet(x, A) \le \exp\{\Delta_{n-1}\,\Delta_n'\} \int_A \Psi_n(x')\lambda(dx') \end{aligned}$$

for any $x \in E$ and any Borel subset $A \subset E$; that is, the nonnegative kernel $R_n^\bullet$ is mixing. Therefore, stability and approximation properties of the sequence $\{\mu_n,\ n \ge 0\}$ defined by (3), can be obtained directly by studying (13) instead, which involves mixing kernels.

4. Stability of nonlinear filters. In practice one rarely has access to the initial distribution of the hidden state process; hence it is important to study the stability of the filter w.r.t. its initial condition. Moreover, the answer to this question will be useful in studing the stability of the filter w.r.t. the model.

Let μ_n denote the filter initialized with the correct μ_0, and let μ_n' denote the filter initialized with a wrong μ_0'; that is, $\mu_n = \bar{R}_{n:1}(\mu_0)$ and $\mu_n' = \bar{R}_{n:1}(\mu_0')$. We are interested in the total variation error at time n induced by the initial error.

THEOREM 4.1. *Without any assumption on the nonnegative kernels, the following inequality holds*:

$$\|\mu_n - \mu_n'\| \le \frac{2}{\log 3}\tau_{n:m} h(\mu_{m-1}, \mu_{m-1}').$$

If in addition the nonnegative kernel R_m is mixing, then

$$\|\mu_n - \mu_n'\| \le \frac{2}{\log 3}\tau_{n:m+1}\frac{1}{\varepsilon_m^2}\|\mu_{m-1} - \mu_{m-1}'\|.$$

COROLLARY 4.2. *If for any $k \geq 1$, the nonnegative kernel R_k is mixing with $\varepsilon_k \geq \varepsilon > 0$, then convergence holds uniformly in time, that is,*

$$\|\mu_n - \mu'_n\| \leq \frac{2}{\varepsilon^2 \log 3} \tau^{n-m} \|\mu_{m-1} - \mu'_{m-1}\| \qquad \text{with } \tau = \frac{1-\varepsilon^2}{1+\varepsilon^2} < 1.$$

PROOF OF THEOREM 4.1. Using (8) and the definition (11) of the Birkhoff contraction coefficient yields

$$\|\bar{R}_{n:m}(\mu) - \bar{R}_{n:m}(\mu')\| \leq \frac{2}{\log 3} h(R_{n:m}\mu, R_{n:m}\mu') \leq \frac{2}{\log 3} \tau_{n:m} h(\mu, \mu'), \tag{14}$$

for any $\mu, \mu' \in \mathcal{P}(E)$. If the nonnegative kernel R_m is mixing, then using (9) yields

$$\begin{aligned} \|\bar{R}_{n:m}(\mu) - \bar{R}_{n:m}(\mu')\| &\leq \frac{2}{\log 3} h(R_{n:m+1}R_m\mu, R_{n:m+1}R_m\mu') \\ &\leq \frac{2}{\log 3} \tau_{n:m+1} h(R_m\mu, R_m\mu') \\ &\leq \frac{2}{\log 3} \tau_{n:m+1} \frac{1}{\varepsilon_m^2} \|\mu - \mu'\|. \end{aligned} \tag{15}$$

Taking $\mu = \mu_{m-1}$ and $\mu' = \mu'_{m-1}$ completes the proof. □

To solve the nonlinear filtering problem, one must have a model to describe the state/observation system, $\{X_n, n \geq 0\}$, $\{Y_n, n \geq 1\}$, as presented in Section 2. The general hidden Markov model is based on the initial condition μ_0, on the transition kernels Q_n and on the likelihood functions Ψ_n, which define the evolution operator $\bar{R}_n$ for the optimal filter μ_n. But, as for the initial condition, in practice one has rarely access to the true model. In particular, the prior information on the state sequence is in general unknown and the choice of Q_n is approximative. Similarly, the probabilistic relation between the observation and the state is in general unknown and the choice of Ψ_n is also approximative. As a result, instead of using the true model, it is common to work with a wrong model, based on a wrong transition kernel Q'_n and a wrong likelihood function Ψ'_n, which define the evolution operator $\bar{R}'_n$ for a wrong filter μ'_n.

Another situation is when the evolution operator $\bar{R}_n$ is known, but difficult to compute. For the purpose of practical implementation, one constructs an approximate filter μ'_n such that the evolution $\mu'_{n-1} \mapsto \mu'_n$ is easy to compute and close to the true evolution $\mu'_{n-1} \mapsto \bar{R}_n(\mu'_{n-1})$.

We are interested in bounding the *global error* between μ'_n and μ_n induced by the *local errors* committed at each time step. We suppose here that $\mu_0 = \mu'_0$, since the problem of a wrong initialization has already been studied above. In full generality, we assume that $\{\mu'_n, n \geq 0\}$ is a random sequence with values

in $\mathcal{P}(E)$, satisfying the following property: for any $n \geq k \geq 1$ and for any bounded measurable function F defined on $\mathcal{P}(E)$,

$$\mathbb{E}[F(\mu'_k)|Y_{1:n}] = \mathbb{E}[F(\mu'_k)|Y_{1:k}]. \tag{16}$$

The results stated below are based on the following decomposition of the global error into a sum of local errors transported by a sequence of normalized evolution operators

$$\begin{aligned} \mu'_n - \mu_n &= \sum_{k=1}^{n} \left[\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k}(\mu'_{k-1})\right] \\ &= \sum_{k=1}^{n} \left[\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1} \circ \bar{R}_k(\mu'_{k-1})\right]. \end{aligned} \tag{17}$$

This equation shows the close relation between the stability w.r.t. the initial condition and the stability w.r.t. the model.

Let us consider first the case where we can estimate the local error in the sense of the Hilbert metric.

ASSUMPTION H (Local error bound in the Hilbert metric).

$$\delta_k^{\mathrm{H}} := \mathbb{E}\big[h\big(\mu'_k, \bar{R}_k(\mu'_{k-1})\big)\big|Y_{1:k}\big] < \infty.$$

REMARK 4.3. If the evolution of the wrong filter μ'_k is defined by the nonnegative kernel $R'_k(x, dx') = Q'_k(x, dx')\Psi'_k(x')$, and if

$$Q_k(x, dx') = q_k(x, x')\lambda_k(dx') \quad \text{and} \quad Q'_k(x, dx') = q'_k(x, x')\lambda_k(dx'),$$

then a sufficient condition for Assumption H to hold is that there exist constants $\delta_k \geq 0$ and $a_k > 0$, such that

$$a_k \leq \frac{\Psi_k(x')q_k(x, x')}{\Psi'_k(x')q'_k(x, x')} \leq a_k \exp(\delta_k)$$

for all $x, x' \in E$, in which case $\delta_k^{\mathrm{H}} \leq \delta_k$.

THEOREM 4.4. *If for any $k \geq 1$, Assumption* H *holds, then*

$$\mathbb{E}[\|\mu'_n - \mu_n\||Y_{1:n}] \leq \frac{2}{\log 3} \sum_{k=1}^{n} \tau_{n:k+1}\delta_k^{\mathrm{H}}. \tag{18}$$

COROLLARY 4.5. *If for any $k \geq 1$, the nonnegative kernel R_k is mixing with $\varepsilon_k \geq \varepsilon > 0$, and Assumption* H *holds with $\delta_k^{\mathrm{H}} \leq \delta$, then convergence holds uniformly in time, that is,*

$$\mathbb{E}[\|\mu_n - \mu'_n\||Y_{1:n}] \leq \frac{2}{\varepsilon^2 \log 3}\delta. \tag{19}$$

Indeed, (19) follows from

$$\sum_{k=1}^{n} \tau^{n-k} = \frac{1-\tau^n}{1-\tau} \leq \frac{1}{1-\tau} = \frac{1+\varepsilon^2}{2\varepsilon^2} \leq \frac{1}{\varepsilon^2}.$$

PROOF OF THEOREM 4.4. Using the decomposition (17), the triangle inequality and estimate (14), yields

$$\begin{aligned}\|\mu'_n - \mu_n\| &\leq \sum_{k=1}^{n} \|\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1} \circ \bar{R}_k(\mu'_{k-1})\| \\ &\leq \frac{2}{\log 3} \sum_{k=1}^{n} \tau_{n:k+1} h(\mu'_k, \bar{R}_k(\mu'_{k-1})).\end{aligned}$$

Taking conditional expectation w.r.t. the observations and using (16), yields (18). □

Let us consider next the case where we can estimate the local error in the sense of the total variation norm

$$\delta_k^{\mathrm{TV}} := \mathbb{E}[\|\mu'_k - \bar{R}_k(\mu'_{k-1})\| | Y_{1:k}] \leq 2.$$

THEOREM 4.6. *If for any $k \geq 1$, the nonnegative kernel R_k is mixing, then*

$$\mathbb{E}[\|\mu_n - \mu'_n\| | Y_{1:n}] \leq \delta_n^{\mathrm{TV}} + \frac{2}{\log 3} \sum_{k=1}^{n-1} \tau_{n:k+2} \frac{\delta_k^{\mathrm{TV}}}{\varepsilon_{k+1}^2}. \tag{20}$$

COROLLARY 4.7. *If for any $k \geq 1$, the nonnegative kernel R_k is mixing, with $\varepsilon_k \geq \varepsilon > 0$, and $\delta_k^{\mathrm{TV}} \leq \delta$, then convergence holds uniformly in time, that is,*

$$\mathbb{E}[\|\mu_n - \mu'_n\| | Y_{1:n}] \leq \left(1 + \frac{2}{\varepsilon^4 \log 3}\right)\delta.$$

PROOF OF THEOREM 4.6. The decomposition (17) is written as

$$\mu'_n - \mu_n = [\mu'_n - \bar{R}_n(\mu'_{n-1})] + \sum_{k=1}^{n-1} [\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1} \circ \bar{R}_k(\mu'_{k-1})], \tag{21}$$

hence using the triangle inequality and estimate (15) yields

$$\begin{aligned}\|\mu'_n - \mu_n\| &\leq \|\mu'_n - \bar{R}_n(\mu'_{n-1})\| + \sum_{k=1}^{n-1} \|\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1} \circ \bar{R}_k(\mu'_{k-1})\| \\ &\leq \|\mu'_n - \bar{R}_n(\mu'_{n-1})\| + \frac{2}{\log 3} \sum_{k=1}^{n-1} \tau_{n:k+2} \frac{1}{\varepsilon_{k+1}^2} \|\mu'_k - \bar{R}_k(\mu'_{k-1})\|.\end{aligned}$$

Taking conditional expectation w.r.t. the observations and using (16), yields (20). □

Let us consider finally the case where we can only estimate the local error in the weak sense

$$\delta_k^{\mathrm{W}} := \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \mu'_k - \bar{R}_k(\mu'_{k-1}), \phi\rangle||Y_{1:k}\big] \le 2.$$

This typically happens if the approximate filter μ'_k is an empirical probability distribution associated with $\bar{R}_k(\mu'_{k-1})$: in this case, bounding the local error requires using the law of large numbers, which can only provide estimates in the weak sense. However, if the nonnegative kernel R_{k+1} is dominated, then using Lemma 3.5, the local error transported by R_{k+1} can be bounded in total variation with the same precision δ_k^{W} as in the weak sense.

THEOREM 4.8. *If for any $k \ge 1$, the nonnegative kernel R_k is mixing, then*

$$\begin{aligned}\sup_{\phi\,:\,\|\phi\|=1} &\mathbb{E}[|\langle \mu_n - \mu'_n, \phi\rangle||Y_{1:n}] \\ &\le \delta_n^{\mathrm{W}} + 2\frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2} + \frac{4}{\log 3}\sum_{k=1}^{n-2} \tau_{n:k+3}\frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2\varepsilon_{k+1}^2}.\end{aligned} \tag{22}$$

COROLLARY 4.9. *If for any $k \ge 1$, the nonnegative kernel R_k is mixing with $\varepsilon_k \ge \varepsilon > 0$, and $\delta_k^{\mathrm{W}} \le \delta$, then convergence holds uniformly in time, that is,*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}[|\langle \mu_n - \mu'_n, \phi\rangle||Y_{1:n}] \le \left(1 + \frac{2}{\varepsilon^2} + \frac{4}{\varepsilon^6 \log 3}\right)\delta.$$

PROOF OF THEOREM 4.8. Using the decomposition (21) and the triangle inequality yields

$$\begin{aligned}&|\langle \mu'_n - \mu_n, \phi\rangle| \\ &\quad\le \big|\langle \mu'_n - \bar{R}_n(\mu'_{n-1}), \phi\rangle\big| + \sum_{k=1}^{n-1} \big\|\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1}\circ\bar{R}_k(\mu'_{k-1})\big\|\,\|\phi\|.\end{aligned} \tag{23}$$

For any $1 \le k \le n-2$, using estimate (15) yields

$$\begin{aligned}&\big\|\bar{R}_{n:k+1}(\mu'_k) - \bar{R}_{n:k+1}\circ\bar{R}_k(\mu'_{k-1})\big\| \\ &\quad= \big\|\bar{R}_{n:k+2}\circ\bar{R}_{k+1}(\mu'_k) - \bar{R}_{n:k+2}\circ\bar{R}_{k+1}\circ\bar{R}_k(\mu'_{k-1})\big\| \\ &\quad\le \frac{2}{\log 3}\tau_{n:k+3}\frac{1}{\varepsilon_{k+2}^2}\big\|\bar{R}_{k+1}(\mu'_k) - \bar{R}_{k+1}\circ\bar{R}_k(\mu'_{k-1})\big\|.\end{aligned}$$

For any $1 \leq k \leq n-1$, using estimate (7) yields

$$\|\bar{R}_{k+1}(\mu'_k) - \bar{R}_{k+1} \circ \bar{R}_k(\mu'_{k-1})\| \leq 2\frac{\|R_{k+1}(\mu'_k - \bar{R}_k(\mu'_{k-1}))\|}{(R_{k+1}\mu'_k)(E)},$$

and the mixing property yields

$$(R_{k+1}\mu'_k)(E) \geq \varepsilon_{k+1}\lambda_{k+1}(E).$$

Taking conditional expectation w.r.t. the observations, using estimate (10) with $K = R_{k+1}$, $\mu = \bar{R}_k(\mu'_{k-1})$, $\mu' = \mu'_k$ and $\mathcal{F} = Y_{1:n}$ and using (16), yields

$$\begin{aligned}
&\mathbb{E}\big[\|R_{k+1}(\mu'_k - \bar{R}_k(\mu'_{k-1}))\| \big| Y_{1:n}\big] \\
&\quad \leq \frac{\lambda_{k+1}(E)}{\varepsilon_{k+1}} \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \mu'_k - \bar{R}_k(\mu'_{k-1}), \phi\rangle| \big| Y_{1:n}\big] \\
&\quad \leq \frac{\lambda_{k+1}(E)}{\varepsilon_{k+1}} \delta_k^{\mathrm{W}}.
\end{aligned}$$

Combining these estimates yields

$$\mathbb{E}\big[\|\bar{R}_{k+1}(\mu'_k) - \bar{R}_{k+1} \circ \bar{R}_k(\mu'_{k-1})\| \big| Y_{1:n}\big] \leq 2\frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+1}^2}.$$

Finally, taking conditional expectation w.r.t. the observations in (23) yields (22). □

5. Uniform convergence of interacting particle filters. In this section and the next we consider again the framework introduced in Section 4, but now the wrong model is chosen deliberately, such that the wrong filter can easily be computed, and remains close to the optimal filter. More specifically, we are interested in particle methods to approximate the optimal filter numerically and we provide estimates of the approximation error. The idea common to all particle filters is to generate an N-sample $(\xi^1_{n|n-1}, \ldots, \xi^N_{n|n-1})$ of i.i.d. random variables, called a *particle system*, with common probability distribution $Q_n\mu^N_{n-1}$, where μ^N_{n-1} is an approximation of μ_{n-1}, and to use the corresponding empirical probability distribution

$$\mu^N_{n|n-1} = \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i_{n|n-1}}$$

as an approximation of $\mu_{n|n-1} = Q_n\mu_{n-1}$. The method is very easy to implement, even in high-dimensional problems, since it is sufficient in principle to simulate independent samples of the hidden state sequence. A major and the earliest

contribution in this field was made by Gordon, Salmond and Smith [15], who proposed to use sampling/importance resampling (SIR) techniques in the correction step: the positive effect of the resampling step is to automatically select particles with larger values of the likelihood function, that is, to concentrate particles in regions of interest of the state space. A very complete account of the currently available mathematical results can be found in [11]. Theoretical and practical aspects can be found in [14].

5.1. *Notation and preliminary results.* Throughout the paper, $S^N(\mu)$ is a shorthand notation for the empirical probability distribution of an N-sample with probability distribution μ, that is,

$$S^N(\mu) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi^i} \qquad \text{with } (\xi^1, \dots, \xi^N) \text{ i.i.d. } \sim \mu.$$

LEMMA 5.1. *For any* $\mu \in \mathcal{P}(E)$,

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\langle S^N(\mu) - \mu, \phi\rangle| \le \frac{1}{\sqrt{N}}.$$

PROOF. It holds

$$\langle S^N(\mu) - \mu, \phi\rangle = \frac{1}{N}\sum_{i=1}^{N} [\phi(\xi^i) - \langle \mu, \phi\rangle],$$

hence

$$\mathbb{E}|\langle S^N(\mu) - \mu, \phi\rangle|^2 = \frac{1}{N}\big[\langle \mu, \phi^2\rangle - \langle \mu, \phi\rangle^2\big] \le \frac{1}{N}\|\phi\|^2. \qquad \square$$

REMARK 5.2. If in addition ϕ and μ are $\mathcal{F}$-measurable r.v.s, and if conditionally w.r.t. $\mathcal{F}$ the r.v.s $(\xi^1, \dots, \xi^N)$ are i.i.d. with (conditional) probability distribution μ, then the same estimate holds for conditional expectation w.r.t. $\mathcal{F}$, that is,

$$\mathbb{E}\big[|\langle S^N(\mu) - \mu, \phi\rangle| \big| \mathcal{F}\big] \le \frac{1}{\sqrt{N}}\|\phi\|. \tag{24}$$

For any nonnegative and bounded measurable function Λ defined on E and any probability distribution μ defined on E, the projective product $\Lambda \cdot \mu$ is defined by

$$\Lambda \cdot \mu := \begin{cases} \dfrac{\Lambda\mu}{\langle \mu, \Lambda\rangle}, & \text{if } \langle \mu, \Lambda\rangle > 0, \\ \nu, & \text{otherwise,} \end{cases}$$

where ν is an arbitrary probability distribution defined on E. If $\langle \mu, \Lambda \rangle > 0$, it follows immediately from Lemma 5.1 and using estimate (6), that

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\langle \Lambda \cdot S^N(\mu) - \Lambda \cdot \mu, \phi \rangle| \leq 2 \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\left| \frac{\langle S^N(\mu) - \mu, \Lambda\phi \rangle}{\langle \mu, \Lambda \rangle} \right| \leq \frac{2}{\sqrt{N}} \frac{\sup_{x\in E} \Lambda(x)}{\langle \mu, \Lambda \rangle}.$$

REMARK 5.3. If in addition ϕ, Λ and μ are $\mathcal{F}$-measurable r.v.s, and if conditionally w.r.t. $\mathcal{F}$ the r.v.s $(\xi^1, \ldots, \xi^N)$ are i.i.d. with (conditional) probability distribution μ, then the same estimate holds for conditional expectation w.r.t. $\mathcal{F}$, that is,

$$\mathbb{E}\big[|\langle \Lambda \cdot S^N(\mu) - \Lambda \cdot \mu, \phi \rangle| \big| \mathcal{F}\big] \leq \frac{2}{\sqrt{N}} \frac{\sup_{x\in E} \Lambda(x)}{\langle \mu, \Lambda \rangle} \|\phi\|. \tag{25}$$

The following procedure, classical in sequential analysis, can be used alternatively.

LEMMA 5.4. *Let $\mu \in \mathcal{P}(E)$, and let Λ be a nonnegative bounded measurable function defined on E, such that $\langle \mu, \Lambda \rangle > 0$. For any $\delta > 0$, define the stopping time*

$$T = \inf\left\{ N : \delta^2 \sum_{i=1}^{N} \Lambda(\xi^i) \geq \sup_{x\in E} \Lambda(x) \right\} \qquad \textit{with } (\xi^1, \ldots, \xi^N, \ldots) \textit{ i.i.d. } \sim \mu.$$

Then

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\langle \Lambda \cdot S^T(\mu) - \Lambda \cdot \mu, \phi \rangle| \leq 2\delta\sqrt{1+\delta^2}.$$

To obtain an error estimate $O(\delta)$, the expected sample size should be $O(1/\delta^2)$, that is,

$$\frac{\rho}{\delta^2} \leq \mathbb{E}[T] \leq \frac{\rho}{\delta^2}(1+\delta^2),$$

where $\rho = (\sup_{x\in E} \Lambda(x))/\langle \mu, \Lambda \rangle$.

The method proposed here to approximate the posterior probability distribution $\Lambda \cdot \mu$ is somehow intermediate, between the classical importance sampling method, which uses a fixed number of random variables, and the acceptance/rejection method, which requires a random number of random variables. In Lemma 5.4, the number of random variables generated is random as in the acceptance/rejection method, but there is no rejection, since all the random variables generated are explicitly used in the approximation, as in the importance sampling method.

PROOF OF LEMMA 5.4. Notice first that a.s.

$$\langle S^N(\mu), \Lambda\rangle = \frac{1}{N}\sum_{i=1}^{N} \Lambda(\xi^i) \longrightarrow \langle \mu, \Lambda\rangle > 0,$$

as $N \uparrow \infty$, hence the stopping time T is a.s. finite. By definition, $\langle S^T(\mu), \Lambda\rangle > 0$, hence using estimate (6) yields

$$\sup_{\phi:\, \|\phi\|=1} \mathbb{E}|\langle \Lambda \cdot S^T(\mu) - \Lambda\cdot\mu, \phi\rangle| \le 2 \sup_{\phi:\, \|\phi\|=1} \mathbb{E}\frac{|\langle S^T(\mu) - \mu, \Lambda\phi\rangle|}{\langle S^T(\mu), \Lambda\rangle},$$

and we define

$$M_N = \sum_{i=1}^{N} [\Lambda(\xi^i)\phi(\xi^i) - \langle \mu, \Lambda\phi\rangle] \quad \text{and} \quad D_N = \sum_{i=1}^{N} \Lambda(\xi^i).$$

By definition of the stopping time T, it holds

$$\frac{\lambda}{\delta^2} \le D_T = D_{T-1} + \Lambda(\xi^T) \le \frac{\lambda}{\delta^2} + \lambda = \frac{\lambda}{\delta^2}(1+\delta^2),$$

where $\lambda = \sup_{x\in E} \Lambda(x)$, and the Cauchy–Schwarz inequality yields

$$\mathbb{E}\frac{|M_T|}{D_T} \le \frac{\delta^2}{\lambda}(\mathbb{E}[M_T^2])^{1/2}.$$

In addition, for any $a > 0$,

$$\mathbb{P}(T > N) \le \exp\left\{a\frac{\lambda}{\delta^2}\right\} r^N \qquad \text{where } r = \int_E \exp\{-a\Lambda(x)\}\mu(dx) < 1,$$

hence the stopping time T is integrable.

It follows from the Wald identity (see, e.g., [25], Proposition IV–4–21), that

$$\frac{\lambda}{\delta^2} \le \mathbb{E}[D_T] = \mathbb{E}[T]\langle \mu, \Lambda\rangle \le \frac{\lambda}{\delta^2}(1+\delta^2)$$

and

$$\begin{aligned}\mathbb{E}[M_T^2] &= \mathbb{E}[T]\big[\langle \mu, \Lambda^2\phi^2\rangle - \langle \mu, \Lambda\phi\rangle^2\big] \\ &\le \mathbb{E}[T]\langle \mu, \Lambda\rangle\lambda\|\phi\|^2 \le \frac{\lambda^2}{\delta^2}(1+\delta^2)\|\phi\|^2,\end{aligned}$$

hence

$$\mathbb{E}\frac{|\langle S^T(\mu) - \mu, \Lambda\phi\rangle|}{\langle S^T(\mu), \Lambda\rangle} = \mathbb{E}\frac{|M_T|}{D_T} \le \delta\sqrt{1+\delta^2}\|\phi\|$$

and

$$\frac{\rho}{\delta^2} \le \mathbb{E}[T] \le \frac{\rho}{\delta^2}(1+\delta^2). \qquad \square$$

REMARK 5.5. If in addition ϕ, Λ and μ are $\mathcal{F}$-measurable r.v.s, and if conditionally w.r.t. $\mathcal{F}$ the r.v.s $(\xi^1, \dots, \xi^N, \dots)$ are i.i.d. with (conditional) probability distribution μ, then the same estimate holds for conditional expectation w.r.t. $\mathcal{F}$, that is,

$$\mathbb{E}\big[|\langle \Lambda \cdot S^T(\mu) - \Lambda \cdot \mu, \phi\rangle| \big| \mathcal{F}\big] \leq 2\delta\sqrt{1+\delta^2}\|\phi\| \tag{26}$$

and

$$\frac{\rho}{\delta^2} \leq \mathbb{E}[T|\mathcal{F}] \leq \frac{\rho}{\delta^2}(1+\delta^2).$$

5.2. *Interacting particle filter.* Let μ_n^N denote the interacting particle filter (IPF) approximation of μ_n. Initially $\mu_0^N = \mu_0$, and the transition from μ_{n-1}^N to μ_n^N is described by the following diagram:

$$\mu_{n-1}^N \underset{\text{sampled prediction}}{\longrightarrow} \mu_{n|n-1}^N = S^N(Q_n\mu_{n-1}^N) \underset{\text{correction}}{\longrightarrow} \mu_n^N = \Psi_n \cdot \mu_{n|n-1}^N.$$

In practice, the particle approximation

$$\mu_{n|n-1}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_{n|n-1}^i}$$

is completely characterized by the particle system $(\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N)$, and the transition from $(\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N)$ to $(\xi_{n+1|n}^1, \dots, \xi_{n+1|n}^N)$ consists of the following steps:

Step (i) (Correction). If the normalization constant

$$c_n = \sum_{i=1}^N \Psi_n(\xi_{n|n-1}^i)$$

is positive, then for all $i = 1, \dots, N$, compute the weight

$$\omega_n^i = \frac{1}{c_n}\Psi_n(\xi_{n|n-1}^i)$$

and set

$$\mu_n^N = \sum_{i=1}^N \omega_n^i \delta_{\xi_{n|n-1}^i},$$

otherwise set $\mu_n^N = \nu$.

Step (ii) (Sampled prediction). Independently for all $i = 1, \dots, N$, generate a r.v. $\xi_{n+1|n}^i \sim Q_{n+1}\mu_n^N$, and set

$$\mu_{n+1|n}^N = S^N(Q_{n+1}\mu_n^N) = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_{n+1|n}^i}.$$

The resampling step (ii) can be easily implemented: it requires generating random variables either according to a weighted discrete probability distribution or according to the arbitrary restarting probability distribution ν.

Notice that the IPF satisfies (16).

REMARK 5.6. Without the reinitialization procedure, proposed initially by Del Moral, Jacod and Protter [10], the normalization constant c_n could take zero value, since the likelihood function Ψ_n is not necessarily positive, and $\mu_n^N = \Psi_n \cdot \mu_{n|n-1}^N$ would not be a well-defined probability distribution. By construction, the sequential particle filter defined at the end of this section does not run into this problem.

REMARK 5.7. If the nonnegative kernel R_n is mixing, then

$$\inf_{\mu\in\mathcal{P}(E)} \langle Q_n\mu, \Psi_n\rangle = \inf_{\mu\in\mathcal{P}(E)} (R_n\mu)(E) \geq \varepsilon_n^2 (R_n\mu_{n-1})(E) = \varepsilon_n^2 \langle \mu_{n|n-1}, \Psi_n\rangle,$$

hence a.s.

$$\inf_{\mu\in\mathcal{P}(E)} \langle Q_n\mu, \Psi_n\rangle > 0,$$

in view of Remark 2.1.

Without loss of generality, it is assumed that the likelihood function is bounded.

ASSUMPTION L.

$$\sup_{x\in E} \Psi_k(x) < \infty.$$

If Assumption L holds, and if for any $k \geq 1$, the nonnegative kernel R_k is mixing, then the following notation is introduced:

$$\rho_k := \frac{\sup_{x\in E} \Psi_k(x)}{\inf_{\mu\in\mathcal{P}(E)} \langle Q_k\mu, \Psi_k\rangle},$$

and in view of Remark 5.7, ρ_k is a.s. finite.

THEOREM 5.8. *If for any $k \geq 1$, Assumption* L *holds, and the nonnegative kernel R_k is mixing, then the IPF estimator satisfies*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \mu_n - \mu_n^N, \phi\rangle| \,\big|\, Y_{1:n}\big] \leq \delta_n^{\mathrm{W}} + 2\frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2} + \frac{4}{\log 3} \sum_{k=1}^{n-2} \tau_{n:k+3} \frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2 \varepsilon_{k+1}^2},$$

where for any $k \geq 1$,

$$\delta_k^{\mathrm{W}} \leq \frac{1}{\sqrt{N}} 2\rho_k.$$

The convergence result stated in Theorem 5.8 would still hold with a time dependent number of particles.

REMARK 5.9. If the transition kernel Q_{n+1} is dominated, that is, $Q_{n+1}(x,\cdot)$ is absolutely continuous w.r.t. $\lambda_{n+1} \in \mathcal{M}^+(E)$, with density $q_{n+1}(x,\cdot)$ bounded by c_{n+1} for any $x \in E$, then convergence in the weak sense of the particle filter can be used to prove convergence in total variation of the particle predictor. Indeed, using Lemma 3.5 yields

$$\mathbb{E}\big[\|\mu_{n+1|n} - Q_{n+1}\mu_n^N\| \big| Y_{1:n}\big] \le c_{n+1}\lambda_{n+1}(E) \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \mu_n - \mu_n^N, \phi\rangle| \big| Y_{1:n}\big],$$

where both $\mu_{n+1|n}$ and $Q_{n+1}\mu_n^N$ are absolutely continuous w.r.t. λ_{n+1}, and

$$\frac{d(Q_{n+1}\mu_n^N)}{d\lambda_{n+1}}(x') = \sum_{i=1}^N \omega_n^i q_{n+1}(\xi_{n|n-1}^i, x')$$

for any $x' \in E$, which can be easily computed.

REMARK 5.10. In general, it is not realistic to assume that the r.v.s ρ_k are uniformly bounded; hence it seems difficult to guarantee that convergence holds uniformly in time for a given observation sequence. On the other hand, averaging over observation sequences makes it possible to obtain convergence uniformly in time, under more realistic assumptions. Indeed, if for any $k \ge 1$, the nonnegative kernel R_k is mixing with nonrandom ε_k, and $\mathbb{E}[\rho_k]$ is finite, then

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}|\langle \mu_n - \mu_n^N, \phi\rangle| \le \delta_n + 2\frac{\delta_{n-1}}{\varepsilon_n^2} + \frac{4}{\log 3}\sum_{k=1}^{n-2} \tau_{n\,:\,k+3}\frac{\delta_k}{\varepsilon_{k+2}^2\varepsilon_{k+1}^2},$$

where for any $k \ge 1$,

$$\delta_k = \mathbb{E}[\delta_k^{\mathrm{W}}] \le \frac{1}{\sqrt{N}} 2\mathbb{E}[\rho_k].$$

REMARK 5.11. Notice that, if the nonnegative kernel R_k is mixing, then

$$\frac{\sup_{x\in E}\Psi_k(x)}{\langle \mu_{k|k-1}, \Psi_k\rangle} \le \rho_k \le \frac{\sup_{x\in E}\Psi_k(x)}{\varepsilon_k^2\langle \mu_{k|k-1}, \Psi_k\rangle},$$

and it follows from Remark 2.2 that

$$\mathbb{E}\left[\frac{\sup_{x\in E}\Psi_k(x)}{\langle \mu_{k|k-1}, \Psi_k\rangle}\Big| Y_{1:k-1}\right] = \int_F \left[\sup_{x\in E} g_k(x,y)\right]\lambda_k^F(dy),$$

hence a necessary and sufficient condition for $\mathbb{E}[\rho_k]$ to be finite is

$$\int_F \left[\sup_{x\in E} g_k(x,y)\right]\lambda_k^F(dy) < \infty.$$

COROLLARY 5.12. *If for any $k \geq 1$, the nonnegative kernel R_k is mixing with $\varepsilon_k \geq \varepsilon > 0$ and nonrandom ε, and $\mathbb{E}[\rho_k] \leq \rho$, then convergence, averaged over observation sequences, holds uniformly in time, that is,*

$$\sup_{\phi:\,\|\phi\|=1} \mathbb{E}|\langle \mu_n - \mu_n^N, \phi\rangle| \leq \left(1 + \frac{2}{\varepsilon^2} + \frac{4}{\varepsilon^6 \log 3}\right)\delta \qquad \text{with } \delta \leq \frac{1}{\sqrt{N}} 2\rho.$$

PROOF OF THEOREM 5.8. It is sufficient to bound the local error δ_k^{W} in the weak sense and to apply Theorem 4.8. Since R_k is mixing, $\langle Q_k \mu_{k-1}^N, \Psi_k\rangle > 0$, in view of Remark 5.7. Using estimate (25) with $\Lambda = \Psi_k$, $\mu = Q_k \mu_{k-1}^N$ and $\mathcal{F} = \sigma(Y_{1:k},\, \mu_{k-1}^N)$ yields

$$\begin{aligned}
&\mathbb{E}\big[|\langle \mu_k^N - \bar{R}_k(\mu_{k-1}^N), \phi\rangle| \,\big|\, Y_{1:k},\, \mu_{k-1}^N\big] \\
&\quad = \mathbb{E}\big[|\langle \Psi_k \cdot \big(S^N(Q_k \mu_{k-1}^N)\big) - \Psi_k \cdot (Q_k \mu_{k-1}^N), \phi\rangle| \,\big|\, Y_{1:k},\, \mu_{k-1}^N\big] \qquad (27) \\
&\quad \leq \frac{2}{\sqrt{N}} \frac{\sup_{x\in E} \Psi_k(x)|}{\langle Q_k \mu_{k-1}^N, \Psi_k\rangle} \|\phi\| \leq \frac{1}{\sqrt{N}} 2\rho_k \|\phi\|.
\end{aligned}$$

□

REMARK 5.13. Let η_n^N denote the interacting particle system (IPS) approximation of η_n considered in [11] and in other works by the same authors. Initially $\eta_0^N = S^N(\eta_0)$, and the transition from η_{n-1}^N to η_n^N is described by the following diagram:

$$\eta_{n-1}^N \underset{\text{correction}}{\longrightarrow} \hat{\eta}_{n-1}^N = \Psi_{n-1} \cdot \eta_{n-1}^N \underset{\substack{\text{sampled}\\ \text{prediction}}}{\longrightarrow} \eta_n^N = S^N(Q_n \hat{\eta}_{n-1}^N).$$

Clearly $\hat{\eta}_0^N = \Psi_0 \cdot \eta_0^N = \Psi_0 \cdot S^N(\eta_0)$, and the transition from $\hat{\eta}_{n-1}^N$ to $\hat{\eta}_n^N$ is described by the following diagram:

$$\hat{\eta}_{n-1}^N \underset{\substack{\text{sampled}\\ \text{prediction}}}{\longrightarrow} \eta_n^N = S^N(Q_n \hat{\eta}_{n-1}^N) \underset{\text{correction}}{\longrightarrow} \hat{\eta}_n^N = \Psi_n \cdot \eta_n^N,$$

which involves exactly the same steps as the transition from μ_{n-1}^N to μ_n^N described above; only the initial conditions $\hat{\eta}_0^N = \Psi_0 \cdot S^N(\eta_0)$ and $\mu_0^N = \mu_0$ are different. Using the following decomposition of the global error into an initial error and a sum of local errors transported by a sequence of normalized evolution operators:

$$\hat{\eta}_n^N - \hat{\eta}_n = \sum_{k=1}^{n} \big[\bar{R}_{n:k+1}(\hat{\eta}_k^N) - \bar{R}_{n:k}(\hat{\eta}_{k-1}^N)\big] + \big[\bar{R}_{n:1}(\hat{\eta}_0^N) - \bar{R}_{n:1}(\hat{\eta}_0)\big],$$

and proceeding as in the proof of Theorem 4.8, yields under the assumptions of

Theorem 5.8,

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \hat{\eta}_n - \hat{\eta}_n^N, \phi\rangle| \big| Y_{0:n}\big]$$
$$\leq \delta_n^{\mathrm{W}} + 2\frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2} + \frac{4}{\log 3}\sum_{k=1}^{n-2} \tau_{n:k+3} \frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2 \varepsilon_{k+1}^2} + \frac{4}{\log 3}\tau_{n:3}\frac{\delta_0^{\mathrm{W}}}{\varepsilon_2^2\varepsilon_1^2},$$

where for any $k \geq 0$,

$$\delta_k^{\mathrm{W}} \leq \frac{1}{\sqrt{N}} 2\rho_k \quad \text{and} \quad \rho_0 := \frac{\sup_{x\in E} \Psi_0(x)}{\langle \mu_0, \Psi_0\rangle}.$$

Finally, notice that

$$\eta_{n+1} - \eta_{n+1}^N = Q_{n+1}\hat{\eta}_n^N - S^N(Q_{n+1}\hat{\eta}_n^N) + Q_{n+1}(\hat{\eta}_n - \hat{\eta}_n^N),$$

hence

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \eta_{n+1} - \eta_{n+1}^N, \phi\rangle| \big| Y_{0:n}\big] \leq \frac{1}{\sqrt{N}} + \sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \hat{\eta}_n - \hat{\eta}_n^N, \phi\rangle| \big| Y_{0:n}\big].$$

This proves the uniform convergence of the IPS approximation to the optimal predictor, with rate $1/\sqrt{N}$, for exactly the model considered in [9, 11], under our weaker mixing assumption.

In the proof of Theorem 5.8, if we use $\langle S^N(Q_k\mu_{k-1}^N), \Psi_k\rangle$ instead of $\langle Q_k\mu_{k-1}^N, \Psi_k\rangle$ as the denominator in (27), we see that, for the local error to be small, the empirical mean of the likelihood function over the predicted particle system should be large enough. This theoretical argument is also supported by numerical evidence, in cases where the likelihood function is localized in a small region of the state space (which typically arises when measurements are accurate). Indeed, such a region can be so small that it does not contain enough points of the predicted particle system, which automatically results in a small value of the predicted empirical mean of the likelihood function. This phenomenon is called *degeneracy of particle weights* and is a known cause of divergence of particle filters. To solve this degeneracy problem, one idea is to add a regularization step to the algorithm; the resulting filters, called regularized particle filters (RPF), are studied in the next section. Another idea is to control the predicted empirical mean

$$\langle S^N(Q_k\mu_{k-1}^N), \Psi_k\rangle = \frac{1}{N}\sum_{i=1}^N \Psi_k(\xi_{k|k-1}^i),$$

by using an adaptive number of particles. To guarantee a local error of order δ_k, independently of any lower bound assumption on the likelihood function, we choose a random number of particles,

$$N_k := \inf\left\{N : \delta_k^2 \sum_{i=1}^N \Psi_k(\xi_{k|k-1}^i) \geq \sup_{x\in E} \Psi_k(x)\right\}, \tag{28}$$

that will automatically fit the difficult case of localized likelihood functions: the resulting filter, called sequential particle filter (SPF) is studied below.

5.3. *Sequential particle filter.* Let $\mu_n^{N_n}$ denote the sequential particle filter (SPF) approximation of μ_n. Initially $\mu_0^{N_0} = \mu_0$, and the transition from $\mu_{n-1}^{N_{n-1}}$ to $\mu_n^{N_n}$ is described by the following diagram:

$$\mu_{n-1}^{N_{n-1}} \underset{\substack{\text{sequential}\\ \text{sampled}\\ \text{prediction}}}{\longrightarrow} \mu_{n|n-1}^{N_n} = S^{N_n}\bigl(Q_n\,\mu_{n-1}^{N_{n-1}}\bigr) \underset{\text{correction}}{\longrightarrow} \mu_n^{N_n} = \Psi_n \cdot \mu_{n|n-1}^{N_n}.$$

In practice, the particle approximation

$$\mu_{n|n-1}^{N_n} = \frac{1}{N_n}\sum_{i=1}^{N_n} \delta_{\xi_{n|n-1}^i}$$

is completely characterized by the particle system $(\xi_{n|n-1}^1, \ldots, \xi_{n|n-1}^{N_n})$, and the transition from $(\xi_{n|n-1}^1, \ldots, \xi_{n|n-1}^{N_n})$ to $(\xi_{n+1|n}^1, \ldots, \xi_{n+1|n}^{N_{n+1}})$ consists of the following steps:

Step (i) (Correction). For all $i = 1, \ldots, N_n$, compute the weight

$$\omega_n^i = \frac{1}{c_n}\Psi_n(\xi_{n|n-1}^i),$$

with the normalization constant

$$c_n = \sum_{i=1}^{N_n} \Psi_n(\xi_{n|n-1}^i)$$

and set

$$\mu_n^{N_n} = \Psi_n \cdot \mu_{n|n-1}^{N_n} = \sum_{i=1}^{N_n} \omega_n^i \delta_{\xi_{n|n-1}^i}.$$

Step (ii) (Sequential sampled prediction). Independently for all $i = 1, \ldots, N_{n+1}$, generate a r.v. $\xi_{n+1|n}^i \sim Q_{n+1}\mu_n^{N_n}$, where the random number N_{n+1} of particles is defined by the stopping time

$$N_{n+1} = \inf\left\{N : \delta_{n+1}^2 \sum_{i=1}^{N} \Psi_{n+1}(\xi_{n+1|n}^i) \geq \sup_{x\in E} \Psi_{n+1}(x)\right\}$$

and set

$$\mu_{n+1|n}^{N_{n+1}} = S^{N_{n+1}}\bigl(Q_{n+1}\mu_n^{N_n}\bigr) = \frac{1}{N_{n+1}}\sum_{i=1}^{N_{n+1}} \delta_{\xi_{n+1|n}^i}.$$

Exactly as for the IPF, the resampling step (ii) can be easily implemented; it only requires generating random variables according to a weighted discrete probability distribution. Notice that a.s.

$$\frac{1}{N}\sum_{i=1}^{N}\Psi_n(\xi_{n|n-1}^i)\to\langle Q_n\mu_{n-1}^{N_{n-1}},\Psi_n\rangle$$

as $N\uparrow\infty$, and if the nonnegative kernel R_n is mixing, then $\langle Q_n\mu_{n-1}^{N_{n-1}},\Psi_n\rangle>0$ in view of Remark 5.7, hence the stopping time N_n is a.s. finite. Moreover, the normalization constant c_n is positive, since

$$c_n=\sum_{i=1}^{N_n}\Psi_n(\xi_{n|n-1}^i)\geq\frac{1}{\delta_n^2}\sup_{x\in E}\Psi_n(x)>0,$$

hence $\Psi_n\cdot\mu_{n|n-1}^{N_n}$ is a well-defined probability distribution.

Notice that the SPF satisfies (16). The following theorem shows that using a random number of particles allows controling the local error independently of any lower bound assumption on the likelihood functions. The counterpart is that the computational time of the resulting algorithm is random and that the expected number of particles does depend on the integrated lower bounds of the likelihood functions.

THEOREM 5.14. *If for any $k\geq1$, the nonnegative kernel R_k is mixing and the random number N_k of particles is defined as in* (28), *then the following inequality holds*:

$$\sup_{\phi\,:\,\|\phi\|=1}\mathbb{E}\big[|\langle\mu_n-\mu_n^{N_n},\phi\rangle||Y_{1:n}\big]\leq\delta_n^{\mathrm{W}}+2\frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2}+\frac{4}{\log3}\sum_{k=1}^{n-2}\tau_{n:k+3}\frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2\varepsilon_{k+1}^2},$$

where for any $k\geq1$,

$$\delta_k^{\mathrm{W}}\leq2\delta_k\sqrt{1+\delta_k^2}$$

and

$$\frac{\rho_k}{\delta_k^2}\leq\mathbb{E}[N_k|Y_{1:k}]\leq\frac{\rho_k}{\delta_k^2}(1+\delta_k^2).$$

COROLLARY 5.15. *If for any $k\geq1$, the nonnegative kernel R_k is mixing with $\varepsilon_k\geq\varepsilon>0$ and the random number N_k of particles is defined as in* (28) *with $\delta_k\leq\delta$, then convergence holds uniformly in time, that is*,

$$\sup_{\phi\,:\,\|\phi\|=1}\mathbb{E}\big[|\langle\mu_n-\mu_n^{N_n},\phi\rangle||Y_{1:n}\big]\leq\left(1+\frac{2}{\varepsilon^2}+\frac{4}{\varepsilon^6\log3}\right)2\delta\sqrt{1+\delta^2}.$$

PROOF OF THEOREM 5.14. It is sufficient to bound the local error δ_k^{W} in the weak sense, and to apply Theorem 4.8. Since R_k is mixing, $\langle Q_k \mu_{k-1}^{N_{k-1}}, \Psi_k \rangle > 0$ in view of Remark 5.7. Using estimate (26) with $\Lambda = \Psi_k$, $\mu = Q_k \mu_{k-1}^{N_{k-1}}$ and $\mathcal{F} = \sigma(Y_{1:k}, \mu_{k-1}^{N_{k-1}})$ yields

$$\begin{aligned}
&\mathbb{E}\big[\big|\langle \mu_k^{N_k} - \bar{R}_k(\mu_{k-1}^{N_{k-1}}), \phi\rangle\big| \big| Y_{1:k}, \mu_{k-1}^{N_{k-1}}\big] \\
&\quad = \mathbb{E}\big[\big|\langle \Psi_k \cdot S^{N_k}(Q_k \mu_{k-1}^{N_{k-1}}) - \Psi_k \cdot (Q_k \mu_{k-1}^{N_{k-1}}), \phi\rangle\big| \big| Y_{1:k}, \mu_{k-1}^{N_{k-1}}\big] \\
&\quad \le 2\delta_k \sqrt{1+\delta_k^2}\, \|\phi\|. \qquad \square
\end{aligned}$$

In this section, we have proved that the IPF and its sequential variant converge uniformly in time under the mixing assumption. This theoretical argument is also supported by numerical evidence, for example, in extreme cases where the hidden state sequence satisfies a noise-free state equation. Indeed, because multiple copies are produced after each resampling step, the diversity of the particle system can only decrease along the time in such cases, and the particle system ultimately concentrates on a few points, if not a single point, of the state space. This phenomenon is called *degeneracy of particle locations* and is another known cause of divergence of particle filters. To solve this degeneracy problem, and also the problem of *degeneracy of particle weights* already mentionned, we have proposed in [23] to add a regularization step in the algorithm, so as to guarantee the diversity of the particle system along the time: the resulting filters, called regularized particle filters (RPF), are studied in the next section under the same mixing assumption.

6. Uniform convergence of regularized particle filters. The main idea consists in changing the discrete approximation μ_n^N for an absolutely continuous approximation, with the effect that in the resampling step N random variables are generated according to an absolutely continuous distribution, hence producing a new particle system with N different particle locations. In doing this, we implicitly assume that the hidden state sequence takes values in a Euclidean space $E = \mathbb{R}^m$ and that the optimal filter μ_n has a smooth density w.r.t. the Lebesgue measure, which is the case in most applications. From the theoretical point of view, this additional assumption allows obtaining strong approximations of the optimal filter, in total variation or in the L^p sense for any $p \ge 1$. In practice, this provides approximate filters which are much more stable along the time than the IPF.

Obtaining an absolutely continuous approximation is achieved by adding a regularization step in the algorithm, using a kernel method, classical in density estimation. If the regularization occurs before the correction by the likelihood function, we obtain the preregularized particle filter, the numerical analysis of which was done in [21], in the general case without the mixing assumption. An

improved version of the preregularized particle filter, called the kernel filter (KF), is proposed in [18]. If the regularization occurs after the correction by the likelihood function, we obtain the postregularized particle filter, which has been proposed in [23] and [27] and compared with the IPF in some classical tracking problems, such as bearing only tracking, or range and bearing tracking with a multiple dynamical model. The local rejection regularized particle filter (L2RPF), which generalizes both the KF and the post-RPF, is introduced in [24], where further implementation details and applications to tracking problems can be found.

6.1. *Notation and preliminary results.* The following notation and definitions will be used below. Throughout the end of this paper, $E = \mathbb{R}^m$. For any $\mu \in \mathcal{P}(E)$, define

$$I(\mu) := \left[\int_E |x|^{m+1} \mu(dx) \right]^{1/m+1},$$

and if μ is absolutely continuous w.r.t. the Lebesgue measure on E, with density $f = \frac{d\mu}{dx}$, define

$$I(f) = I(\mu) = \left[\int_E |x|^{m+1} f(x)\, dx \right]^{1/m+1}$$

and

$$J(f) = J\left(\frac{d\mu}{dx}\right) := \int_E \sqrt{f(x)}\, dx.$$

From the multidimensional Carlson inequality (see [16], Lemma 7) there exists a universal constant A_m such that, for any absolutely continuous $\mu \in \mathcal{P}(E)$,

$$J\left(\frac{d\mu}{dx}\right) \le A_m (I(\mu))^{m/2}. \tag{29}$$

Hence $J(\frac{d\mu}{dx})$ is finite if $I(\mu)$ is finite. Let $W^{2,1}$ denote the Sobolev space of functions defined on E, which together with their derivatives up to order two, are integrable w.r.t. the Lebesgue measure on E. Let $\|\cdot\|_{2,1}$ and $|\cdot|_{2,1}$ denote the corresponding norm and seminorm, that is,

$$\|u\|_{2,1} := \sum_{0 \le |i| \le 2} \int_E |D^i u(x)|\, dx \quad \text{and} \quad |u|_{2,1} := \sum_{|i|=2} \int_E |D^i u(x)|\, dx,$$

respectively, where for any multiindex $i = (i_1, \ldots, i_m)$ of order $|i| = i_1 + \cdots + i_m$,

$$D^i = \frac{\partial^{i_1 + \cdots + i_m}}{\partial x_1^{i_1} \cdots \partial x_m^{i_m}}.$$

Let the regularization kernel K be a symmetric probability density on E, such that

$$\int_E K(x)\,dx = 1, \qquad \int_E xK(x)\,dx = 0 \quad \text{and} \quad \alpha := \tfrac{1}{2}\int_E |x|^2 K(x)\,dx < \infty.$$

Assume also that the regularization kernel K is square integrable, that is,

$$\beta := \left[\int_E K^2(x)\,dx\right]^{1/2} < \infty$$

and that the symmetric probability density $L = \frac{K^2}{\beta^2}$ satisfies

$$\gamma := I(L) = \left[\int_E |x|^{m+1} L(x)\,dx\right]^{1/m+1} < \infty.$$

For any bandwidth $h > 0$, define the rescaled kernel

$$K_h(x) = \frac{1}{h^m} K\left(\frac{x}{h}\right)$$

for any $x \in E$.

DEFINITION 6.1. For any $\mu \in \mathcal{M}^+(E)$, the nonnegative measure $K_h * \mu$ is defined as the measure absolutely continuous w.r.t. the Lebesgue measure, with density

$$\frac{d(K_h * \mu)}{dx}(x) = \int_E K_h(x - x')\mu(dx'),$$

where $*$ denotes the convolution operator.

Notice that convolution by K_h preserves the total mass, that is, $(K_h * \mu)(E) = \mu(E)$ for any $\mu \in \mathcal{M}^+(E)$, hence $K_h * \bar{\mu}$ is the normalized nonnegative measure (i.e., probability distribution) associated with the unnormalized nonnegative measure $K_h * \mu$. Moreover, $K_h * \mu$ approximates μ in the following sense.

LEMMA 6.2. *Let $\mu \in \mathcal{M}^+(E)$ be absolutely continuous w.r.t. the Lebesgue measure, with density $\frac{d\mu}{dx} \in W^{2,1}$. Then*

$$\|K_h * \mu - \mu\| \le \alpha h^2 \left|\frac{d\mu}{dx}\right|_{2,1}.$$

The proof for the one-dimensional case can be found in [13], Theorem 7.1, and for the multidimensional case in [29], Chapter I, Lemma 4.4, or in [16], Proposition 4.

Given a sample $(\xi^1, \ldots, \xi^N)$ from an unknown probability distribution $\mu \in \mathcal{P}(E)$ with a smooth density, and given a nonnegative function Λ which can

be evaluated at any point of E, we are interested in approximating the posterior probability distribution $\Lambda \cdot \mu$ by a probability distribution with a smooth density. By construction, the approximation error can be estimated in a strong sense such as the total variation. Of course, more sample points are needed to approximate a whole density than are needed to simply approximate moments, as in the weak sense approximation. Typically, the sample size depends on the dimension, which is the usual *curse of dimensionality*. However, getting an approximation of the whole density is usually worth the effort, as it allows getting meaningful information, for example, confidence regions.

The classical density estimation theory (see, e.g., [30] for the L^2 theory and [13] for the L^1 theory), has extensively dealt with the problem of estimating μ alone, which reduces to our problem in the particular case where Λ is constant. The solution consists in regularizing the empirical probability distribution associated with the sample and provides kernel-type estimators $K_h * S^N(\mu)$. Minimization of the mean errors $\mathbb{E}\|K_h * S^N(\mu) - \mu\|$ or $\mathbb{E}\|K_h * S^N(\mu) - \mu\|_2^2$, in the L^1 or L^2 sense, over the bandwidth h and the regularization kernel K has also been studied. Similarly, in our more general setting, we propose two estimators for $\Lambda \cdot \mu$: the preregularized estimator $\Lambda \cdot (K_h * S^N(\mu))$ where regularization occurs before the correction by Λ, and the postregularized estimator $K_h * (\Lambda \cdot S^N(\mu))$ where regularization occurs after the correction by Λ. We immediately see that the preregularized estimator consists in applying the correction by Λ to the classical density estimator $K_h * S^N(\mu)$ and that both estimators reduce to the classical density estimator when Λ is constant. Consequently, we will focus below on estimating the mean error for the postregularized estimator, and results for the preregularized estimator will follow immediately. In Proposition 6.3, we consider the mean error between the unnormalized nonnegative measures $K_h * (\Lambda S^N(\mu))$ and $\Lambda\mu$. In the general case, the error between the corresponding probability distributions $K_h * (\Lambda \cdot S^N(\mu))$ and $\Lambda \cdot \mu$ will then be derived using estimate (7), but in some particular cases we may derive some sharper bounds. In what follows, we only state some bounds without trying to optimize over the bandwidth h or the regularization kernel K.

PROPOSITION 6.3. *Let $\mu \in \mathcal{P}(E)$ be absolutely continuous w.r.t. the Lebesgue measure, with density $\frac{d\mu}{dx} \in W^{2,1}$, and let Λ be a nonnegative bounded measurable function defined on E, with bounded derivatives up to order two. Then*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big|\big\langle K_h * \big(\Lambda S^N(\mu)\big) - \Lambda\mu, \phi\big\rangle\big| \le \frac{1}{\sqrt{N}} \sup_{x\in E} \Lambda(x) + \alpha h^2 \left|\Lambda \frac{d\mu}{dx}\right|_{2,1}.$$

If in addition $I(\mu)$ is finite, then

$$\mathbb{E}\big\| K_h * \big(\Lambda S^N(\mu)\big) - \Lambda\mu \big\| \le \frac{\beta A_m}{\sqrt{N h^m}} (I(\mu) + h\gamma)^{m/2} \sup_{x\in E} \Lambda(x) + \alpha h^2 \left|\Lambda \frac{d\mu}{dx}\right|_{2,1}.$$

The proof is based on the following decomposition of the error into variation and bias errors:

$$K_h * (\Lambda S^N(\mu)) - \Lambda\mu = K_h * (\Lambda S^N(\mu)) - K_h * (\Lambda\mu) + K_h * (\Lambda\mu) - \Lambda\mu.$$

Under the assumptions, the nonnegative measure $\Lambda\mu$ is absolutely continuous w.r.t. the Lebesgue measure, with density $\Lambda\frac{d\mu}{dx} \in W^{2,1}$, such that

$$\left|\Lambda\frac{d\mu}{dx}\right|_{2,1} \leq \sup_{u\in W^{2,1}} \frac{|\Lambda u|_{2,1}}{\|u\|_{2,1}} \left\|\frac{d\mu}{dx}\right\|_{2,1},$$

and Lemma 6.2 can be used to bound the bias error. The following lemma is used to bound the variation error.

LEMMA 6.4. *Let $\mu \in \mathcal{P}(E)$, and let Λ be a nonnegative bounded measurable function defined on E. Then*

$$\sup_{\phi:\,\|\phi\|=1} \mathbb{E}\left|\langle K_h * (\Lambda S^N(\mu)) - K_h * (\Lambda\mu), \phi\rangle\right| \leq \frac{1}{\sqrt{N}} \sup_{x\in E} \Lambda(x).$$

If in addition $I(\mu)$ is finite, then

$$\mathbb{E}\left\| K_h * (\Lambda S^N(\mu)) - K_h * (\Lambda\mu)\right\| \leq \frac{\beta A_m}{\sqrt{Nh^m}} \big(I(\mu) + h\gamma\big)^{m/2} \sup_{x\in E} \Lambda(x).$$

PROOF. Using Lemma 5.1 and the symmetry of the regularization kernel, yields

$$\begin{aligned}
&\mathbb{E}\left|\langle K_h * (\Lambda S^N(\mu)) - K_h * (\Lambda\mu), \phi\rangle\right| \\
&\quad = \mathbb{E}\left|\langle S^N(\mu) - \mu, \Lambda(K_h * \phi)\rangle\right| \\
&\quad \leq \frac{1}{\sqrt{N}} \|\Lambda(K_h * \phi)\| \leq \frac{1}{\sqrt{N}} \|\phi\| \sup_{x\in E} \Lambda(x),
\end{aligned}$$

for any bounded measurable test function ϕ defined on E, which proves the estimate in the weak sense.

The proof of the estimate in total variation is classical. By definition,

$$\left\| K_h * (\Lambda S^N(\mu)) - K_h * (\Lambda\mu)\right\| = \int_E |f_h^N(x) - f_h(x)|\, dx,$$

where

$$f_h^N(x) = \frac{d(K_h * (\Lambda S^N(\mu)))}{dx}(x) = \frac{1}{N}\sum_{i=1}^N K_h(x - \xi^i)\Lambda(\xi^i),$$

$$f_h(x) = \frac{d(K_h * (\Lambda\mu))}{dx}(x) = \int_E K_h(x - x')\Lambda(x')\mu(dx'),$$

and it follows from the proof of Lemma 5.1 that

$$\mathbb{E}|f_h^N(x) - f_h(x)| \le \frac{1}{\sqrt{N}}\Big[\int_E K_h^2(x-x')\Lambda^2(x')\mu(dx')\Big]^{1/2}$$
$$\le \frac{\beta}{\sqrt{Nh^m}}\Big[\int_E L_h(x-x')\mu(dx')\Big]^{1/2}\sup_{x\in E}\Lambda(x),$$

where the symmetric probability density $L = \frac{K^2}{\beta^2}$ satisfies $K_h^2 = \frac{\beta^2}{h^m}L_h$. Therefore

$$\mathbb{E}\big\|K_h * \big(\Lambda S^N(\mu)\big) - K_h * (\Lambda\mu)\big\| \le \frac{\beta}{\sqrt{Nh^m}} J\left(\frac{d(L_h * \mu)}{dx}\right)\sup_{x\in E}\Lambda(x).$$

Using the Minkowski inequality yields

$$I(L_h * \mu) = \Big[\int_E\int_E |x' + hu|^{m+1}L(u)\mu(dx')\,du\Big]^{1/m+1}$$
$$\le \Big[\int_E |x'|^{m+1}\mu(dx')\Big]^{1/m+1} + h\Big[\int_E |u|^{m+1}L(u)\,du\Big]^{1/m+1}$$
$$\le I(\mu) + hI(L),$$

and using estimate (29) yields

$$J\left(\frac{d(L_h * \mu)}{dx}\right) \le A_m(I(L_h * \mu))^{m/2} \le A_m\big(I(\mu) + h\gamma\big)^{m/2}. \qquad \square$$

REMARK 6.5. If the probability distribution μ is absolutely continuous w.r.t. the Lebesgue measure, with probability density f, then the following more precise bound holds:

$$J\left(\frac{d(L_h * \mu)}{dx}\right) = J(L_h * f) \le J(f) + J(|L_h * f - f|),$$

where $J(|L_h * f - f|)$ goes to zero when $h \downarrow 0$ (see [16], Proposition 8).

REMARK 6.6. If in addition ϕ, Λ and μ are $\mathcal{F}$-measurable r.v.s, and if conditionally w.r.t. $\mathcal{F}$ the r.v.s $(\xi^1, \ldots, \xi^N)$ are i.i.d. with (conditional) probability distribution μ, then the same estimates hold for conditional expectation w.r.t. $\mathcal{F}$, that is,

$$\mathbb{E}\big[\big|\langle K_h * \big(\Lambda S^N(\mu)\big) - \Lambda\mu, \phi\rangle\big|\big|\mathcal{F}\big]$$
$$\le \Big[\frac{1}{\sqrt{N}}\sup_{x\in E}\Lambda(x) + \alpha h^2\Big|\Lambda\frac{d\mu}{dx}\Big|_{2,1}\Big]\|\phi\| \tag{30}$$

and

$$\mathbb{E}\big[\big\|K_h * \big(\Lambda S^N(\mu)\big) - \Lambda\mu\big\|\big|\mathcal{F}\big]$$
$$\le \frac{\beta A_m}{\sqrt{Nh^m}}\big(I(\mu) + h\gamma\big)^{m/2}\sup_{x\in E}\Lambda(x) + \alpha h^2\Big|\Lambda\frac{d\mu}{dx}\Big|_{2,1}. \tag{31}$$

6.2. *Preregularized particle filter.* Let $\mu_n^{N,h}$ denote the preregularized particle filter (pre-RPF) approximation of μ_n. Initially $\mu_0^{N,h} = \mu_0$, and the transition from $\mu_{n-1}^{N,h}$ to $\mu_n^{N,h}$ is described by the following diagram:

$$\mu_{n-1}^{N,h} \underset{\text{sampled prediction}}{\longrightarrow} \mu_{n|n-1}^{N,h} = S^N(Q_n \mu_{n-1}^{N,h}) \underset{\text{preregularized correction}}{\longrightarrow} \mu_n^{N,h} = \Psi_n \cdot (K_h * \mu_{n|n-1}^{N,h}).$$

In practice, the particle approximation

$$\mu_{n|n-1}^{N,h} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_{n|n-1}^i},$$

is completely characterized by the particle system $\{\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N\}$, and the transition from $\{\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N\}$ to $\{\xi_{n+1|n}^1, \dots, \xi_{n+1|n}^N\}$ consists of the following steps.

Step (i) (Preregularized correction). If the normalization constant

$$c_n = \sum_{i=1}^{N} (K_h * \Psi_n)(\xi_{n|n-1}^i)$$

is positive, then set

$$\mu_n^{N,h}(dx) = \big(\Psi_n \cdot (K_h * \mu_{n|n-1}^{N,h})\big)(dx) = \frac{1}{c_n} \sum_{i=1}^{N} \Psi_n(x) K_h(x - \xi_{n|n-1}^i)\, dx,$$

otherwise set $\mu_n^{N,h} = \nu$.

Step (ii) (Sampled prediction). Independently for all $i = 1, \dots, N$, generate $\xi_{n+1|n}^i \sim Q_{n+1}\, \mu_n^{N,h}$, and set

$$\mu_{n+1|n}^{N,h} = S^N(Q_{n+1} \mu_n^{N,h}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_{n+1|n}^i}.$$

The resampling step (ii) requires generating random variables either according to the arbitrary restarting probability distribution ν, or according to a complex probability distribution known up to a normalization constant, which can be done with a rejection algorithm (see [12]), or with the more efficient local rejection algorithm (see [18]), if the kernel K has compact support. In any case, the implementation is less straightforward than for the IPF.

Notice that the pre-RPF satisfies (16).

ASSUMPTION R (Regularity of the Markov kernel). For any $\mu \in \mathcal{P}(E)$, the probability distribution $Q_k \mu$ is absolutely continuous w.r.t. the Lebesgue measure, with density in $W^{2,1}$, and

$$D_k := \sup_{\mu \in \mathcal{P}(E)} \left| \frac{d(Q_k \mu)}{dx} \right|_{2,1} < \infty.$$

REMARK 6.7. Assumption R is equivalent to supposing that the probability distribution $Q_k(x,\cdot)$ is absolutely continuous w.r.t. the Lebesgue measure, with density $q_k(x,\cdot)$ in $W^{2,1}$ for any $x \in E$, and

$$D_k = \sup_{x\in E} |q_k(x,\cdot)|_{2,1} < \infty.$$

ASSUMPTION M (Existence of moments).

$$I_k := \sup_{\mu\in\mathcal{P}(E)} I(Q_k\mu) < \infty.$$

REMARK 6.8. Assumption M is equivalent to supposing that

$$I_k = \sup_{x\in E}\left[\int_E |x'|^{m+1} Q_k(x,dx')\right]^{1/m+1} < \infty.$$

Alternatively, if the Markov kernel Q_k is mixing, and $I(\mu_{k|k-1})$ is finite, then

$$I_k \le \frac{1}{\varepsilon_k^{2/m+1}} I(\mu_{k|k-1}) < \infty.$$

THEOREM 6.9. *If for any $k \ge 1$, Assumptions* L *and* R *hold, and the nonnegative kernel R_k is mixing, then the pre-RPF estimator satisfies*

$$\sup_{\phi\,:\,\|\phi\|=1} \mathbb{E}\big[|\langle \mu_n - \mu_n^{N,h}, \phi\rangle| \big| Y_{1:n}\big] \le \delta_n^{\mathrm{W}} + 2\frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2} + \frac{4}{\log 3}\sum_{k=1}^{n-2} \tau_{n:k+3} \frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2\varepsilon_{k+1}^2},$$

where for any $k \ge 1$,

$$\delta_k^{\mathrm{W}} \le \left[\frac{1}{\sqrt{N}} + \alpha h^2 D_k\right] 2\rho_k.$$

If in addition for any $k \ge 1$, Assumption M *holds, then*

$$\mathbb{E}\big[\|\mu_n - \mu_n^{N,h}\| \big| Y_{1:n}\big] \le \delta_n^{\mathrm{TV}} + \frac{2}{\log 3}\sum_{k=1}^{n-1} \tau_{n:k+2} \frac{\delta_k^{\mathrm{TV}}}{\varepsilon_{k+1}^2},$$

where for any $k \ge 1$,

$$\delta_k^{\mathrm{TV}} \le \left[\frac{\beta A_m}{\sqrt{Nh^m}}(I_k + h\gamma)^{m/2} + \alpha h^2 D_k\right] 2\rho_k.$$

The convergence result stated in Theorem 6.9 would still hold with a time dependent bandwidth, and with a time dependent number of particles.

PROOF OF THEOREM 6.9. To prove the estimate in the weak sense, it is sufficient to bound the local error δ_k^{W} in the weak sense, and to apply Theorem 4.8.

Since R_k is mixing, $\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle > 0$ in view of Remark 5.7. Using estimate (6) yields

$$\begin{aligned}|\langle \mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h}), \phi\rangle| &= \left|\left\langle \Psi_k \cdot \left(K_h * S^N(Q_k\mu_{k-1}^{N,h})\right) - \Psi_k \cdot (Q_k\mu_{k-1}^{N,h}), \phi\right\rangle\right| \\ &\le \frac{|\langle K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}, \Psi_k\phi\rangle|}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle} \\ &\quad + \frac{|\langle K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle|}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle}\|\phi\|\end{aligned}$$

for any bounded measurable test function ϕ defined on E. By definition,

$$\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle \ge \inf_{\mu\in\mathcal{P}(E)}\langle Q_k\mu, \Psi_k\rangle.$$

Using estimate (30) with $\Lambda \equiv 1$, $\mu = Q_k\mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(Y_{1:k},\ \mu_{k-1}^{N,h})$, yields

$$\begin{aligned}&\mathbb{E}\left[|\langle K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}, \Psi_k\phi\rangle| \,\middle|\, Y_{1:k},\ \mu_{k-1}^{N,h}\right] \\ &\quad\le \left[\frac{1}{\sqrt{N}} + \alpha h^2\left|\frac{d(Q_k\mu_{k-1}^{N,h})}{dx}\right|_{2,1}\right]\|\Psi_k\phi\| \\ &\quad\le \left[\frac{1}{\sqrt{N}} + \alpha h^2 D_k\right]\|\phi\|\sup_{x\in E}\Psi_k(x).\end{aligned}$$

To prove the estimate in total variation, it is sufficient to bound the local error δ_k^{TV} in total variation, and to apply Theorem 4.6. Using estimate (7) yields

$$\begin{aligned}\left\|\mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h})\right\| &= \left\|\Psi_k \cdot \left(K_h * S^N(Q_k\mu_{k-1}^{N,h})\right) - \Psi_k \cdot (Q_k\mu_{k-1}^{N,h})\right\| \\ &\le 2\frac{\sup_{x\in E}\Psi_k(x)}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle}\left\|K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}\right\|.\end{aligned}$$

Using estimate (31) with $\Lambda \equiv 1$, $\mu = Q_k\mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(Y_{1:k},\ \mu_{k-1}^{N,h})$, yields

$$\begin{aligned}&\mathbb{E}\left[\left\|K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}\right\| \,\middle|\, Y_{1:k},\ \mu_{k-1}^{N,h}\right] \\ &\quad\le \frac{\beta A_m}{\sqrt{Nh^m}}\left(I(Q_k\mu_{k-1}^{N,h}) + h\gamma\right)^{m/2} + \alpha h^2\left|\frac{d(Q_k\mu_{k-1}^{N,h})}{dx}\right|_{2,1} \\ &\quad\le \frac{\beta A_m}{\sqrt{Nh^m}}(I_k + h\gamma)^{m/2} + \alpha h^2 D_k.\end{aligned}$$

□

In the pre-RPF, the correction is applied directly to a regularized probability distribution, hence each point in the support of the regularized density is updated, and in principle the *degeneracy of particle weights* which occurs when the correction is applied to a discrete probability distribution, as in the IPF, is now avoided. This intuition is supported by numerical evidence, and by the following theorem, which shows that it is possible to control the local error, averaged over observation sequences, independently of any lower bound assumption on the likelihood functions (notice that the a.s. bounds of Theorem 6.9 still depend on the integrated lower bounds of the likelihood functions).

THEOREM 6.10. *If for any* $k \geq 1$, *Assumptions* R *and* M *hold*, *and the nonnegative kernel* R_k *is mixing with nonrandom* ε_k, *then the pre-RPF estimator satisfies*

$$\mathbb{E}\|\mu_n - \mu_n^{N,h}\| \leq \delta_n + \frac{2}{\log 3} \sum_{k=1}^{n-1} \tau_{n:k+2} \frac{\delta_k}{\varepsilon_{k+1}^2},$$

where for any $k \geq 1$,

$$\delta_k \leq \frac{2}{\varepsilon_k^2} \left[\frac{\beta A_m}{\sqrt{Nh^m}} (I_k + h\gamma)^{m/2} + \alpha h^2 D_k \right].$$

COROLLARY 6.11. *If for any* $k \geq 1$, *the nonnegative kernel* R_k *is mixing with* $\varepsilon_k \geq \varepsilon > 0$ *and nonrandom* ε, *and Assumptions* R *and* M *hold with* $D_k \leq D$ *and* $I_k \leq I$, *then convergence*, *averaged over observation sequences*, *holds uniformly in time*, *that is*,

$$\mathbb{E}\|\mu_n - \mu_n^{N,h}\| \leq \left(1 + \frac{2}{\varepsilon^4 \log 3}\right) \delta,$$

with

$$\delta \leq \frac{2}{\varepsilon^2} \left[\frac{\beta A_m}{\sqrt{Nh^m}} (I + h\gamma)^{m/2} + \alpha h^2 D \right].$$

Both the SPF (see Theorem 5.14) and the pre-RPF allow bounding the error independently of any lower bound assumption on the likelihood functions, and the computational time of both algorithms is random [recall that a rejection is needed in the resampling step (ii) of the pre-RPF].

PROOF OF THEOREM 6.10. It is sufficient to bound the local error $\mathbb{E}[\delta_k^{\mathrm{TV}}]$ in total variation, averaged over observation sequences, and to apply Theorem 4.6. If the nonnegative kernel R_k is mixing, then

$$\langle Q_k \mu_{k-1}^{N,h}, \Psi_k \rangle \geq \varepsilon_k^2 \langle \mu_{k|k-1}, \Psi_k \rangle,$$

in view of Remark 5.7, with nonrandom ε_k, hence using inequality (7) yields

$$\begin{aligned}\|\mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h})\| &\le 2\int_E \frac{\Psi_k(x)}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle}|\mu_k|(dx)\\ &\le \frac{2}{\varepsilon_k^2}\int_E \frac{\Psi_k(x)}{\langle \mu_{k|k-1}, \Psi_k\rangle}|\mu_k|(dx),\end{aligned}$$

where

$$\mu_k = K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}.$$

It follows from Remark 2.2 that

$$\mathbb{E}\left[\frac{\Psi_k(x)}{\langle \mu_{k|k-1}, \Psi_k\rangle}\middle| Y_{1:k-1},\ \mu_{k|k-1}^{N,h},\ \mu_{k-1}^{N,h}\right] = \mathbb{E}\left[\frac{\Psi_k(x)}{\langle \mu_{k|k-1}, \Psi_k\rangle}\middle| Y_{1:k-1}\right] = 1,$$

hence

$$\mathbb{E}[\|\mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h})\||Y_{1:k-1},\ \mu_{k|k-1}^{N,h},\ \mu_{k-1}^{N,h}] \le \frac{2}{\varepsilon_k^2}|\mu_k|(E).$$

Using estimate (31) with $\Lambda \equiv 1$, $\mu = Q_k\,\mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(\mu_{k-1}^{N,h})$, yields

$$\begin{aligned}\mathbb{E}[|\mu_k|(E)|\mu_{k-1}^{N,h}] &= \mathbb{E}[\|K_h * S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}\||\mu_{k-1}^{N,h}]\\ &\le \frac{\beta A_m}{\sqrt{Nh^m}}\big(I(Q_k\mu_{k-1}^{N,h}) + h\gamma\big)^{m/2} + \alpha h^2\left|\frac{d(Q_k\mu_{k-1}^{N,h})}{dx}\right|_{2,1}\\ &\le \frac{\beta A_m}{\sqrt{Nh^m}}(I_k + h\gamma)^{m/2} + \alpha h^2 D_k.\end{aligned}$$

□

REMARK 6.12. In the same way as for the SPF (see Theorem 5.14), one could think of using a random number of particles so as to avoid any lower bound assumption on the likelihood functions. However, for the pre-RPF, it is not sufficient to evaluate the quantity $\Psi_k(\xi_{k|k-1}^i)$ for each simulated particle, as in (28), but one has to evaluate the integral

$$(K_h * \Psi_k)(\xi_{k|k-1}^i) = \int_E \Psi_k(x)K_h(x - \xi_{k|k-1}^i)\,dx,$$

instead. This evaluation is in general very costly, which makes the idea of using a random number of particles for the pre-RPF rather unpractical.

6.3. *Post-regularized particle filter.* Let $\mu_n^{N,h}$ denote the post-regularized particle filter (post-RPF) approximation of μ_n. Initially $\mu_0^{N,h} = \mu_0$, and the transition from $\mu_{n-1}^{N,h}$ to $\mu_n^{N,h}$ is described by the following diagram:

$$\mu_{n-1}^{N,h} \underset{\text{sampled prediction}}{\longrightarrow} \mu_{n|n-1}^{N,h} = S^N(Q_n\mu_{n-1}^{N,h}) \underset{\text{postregularized correction}}{\longrightarrow} \mu_n^{N,h} = K_h * (\Psi_n \cdot \mu_{n|n-1}^{N,h}).$$

In practice, the particle approximation

$$\mu_{n|n-1}^{N,h} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_{n|n-1}^i}$$

is completely characterized by the particle system $\{\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N\}$, and the transition from $\{\xi_{n|n-1}^1, \dots, \xi_{n|n-1}^N\}$ to $\{\xi_{n+1|n}^1, \dots, \xi_{n+1|n}^N\}$ consists of the following steps.

Step (i) (Post-regularized correction). If the normalization constant

$$c_n = \sum_{i=1}^{N} \Psi_n(\xi_{n|n-1}^i)$$

is positive, then for all $i = 1, \dots, N$, compute the weight

$$\omega_n^i = \frac{1}{c_n}\Psi_n(\xi_{n|n-1}^i)$$

and set

$$\mu_n^{N,h}(dx) = \big(K_h * (\Psi_n \cdot \mu_{n|n-1}^{N,h})\big)(dx) = \sum_{i=1}^{N} \omega_n^i K_h(x - \xi_{n|n-1}^i)\, dx,$$

otherwise set $\mu_n^{N,h} = K_h * \nu$.

Step (ii) (Sampled prediction). Independently for all $i = 1, \dots, N$, generate a r.v. $\xi_{n+1|n}^i \sim Q_{n+1}\mu_n^{N,h}$ and set

$$\mu_{n+1|n}^{N,h} = S^N(Q_{n+1}\mu_n^{N,h}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_{n+1|n}^i}.$$

The resampling step (ii) can be easily implemented: it requires generating random variables according to a mixture of rescaled kernels, where the mixing probability distribution is either a weighted discrete probability distribution or the arbitrary restarting probability distribution ν.

Notice that the post-RPF satisfies (16).

ASSUMPTION L″.

$$\sup_{u \in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}} < \infty.$$

REMARK 6.13. If Ψ_k is bounded, with bounded derivatives up to order two, then

$$\sup_{u \in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}} < \infty.$$

If Assumption L″ holds, and if for any $k \geq 1$ the nonnegative kernel R_k is mixing, then the following notation is introduced:

$$\rho_k'' := \frac{\sup_{u \in W^{2,1}} (|\Psi_k u|_{2,1} / \|u\|_{2,1})}{\inf_{\mu \in \mathcal{P}(E)} \langle Q_k \mu, \Psi_k \rangle},$$

and in view of Remark 5.7, ρ_k'' is a.s. finite.

ASSUMPTION R″ (Additional regularity of the Markov kernel). For any $\mu \in \mathcal{P}(E)$, the probability distribution $Q_k \mu$ is absolutely continuous w.r.t. the Lebesgue measure, with density in $W^{2,1}$, and

$$D_k'' := \sup_{\mu \in \mathcal{P}(E)} \left\| \frac{d(Q_k \mu)}{dx} \right\|_{2,1} < \infty.$$

REMARK 6.14. Assumption R″ is equivalent to supposing that the probability distribution $Q_k(x, \cdot)$ is absolutely continuous w.r.t. the Lebesgue measure, with density $q_k(x, \cdot)$ in $W^{2,1}$ for any $x \in E$, and

$$D_k'' = \sup_{x \in E} \|q_k(x, \cdot)\|_{2,1} < \infty.$$

THEOREM 6.15. *If for any $k \geq 1$, Assumptions* L″ *and* R″ *hold and the nonnegative kernel R_k is mixing, then the post-RPF estimator satisfies*

$$\sup_{\phi : \|\phi\| = 1} \mathbb{E}\big[|\langle \mu_n - \mu_n^{N,h}, \phi \rangle| \big| Y_{1:n} \big] \leq \delta_n^{\mathrm{W}} + 2 \frac{\delta_{n-1}^{\mathrm{W}}}{\varepsilon_n^2} + \frac{4}{\log 3} \sum_{k=1}^{n-2} \tau_{n:k+3} \frac{\delta_k^{\mathrm{W}}}{\varepsilon_{k+2}^2 \varepsilon_{k+1}^2},$$

where for any $k \geq 1$,

$$\delta_k^{\mathrm{W}} \leq \frac{1}{\sqrt{N}} 2 \rho_k + \alpha h^2 D_k'' \rho_k''.$$

If in addition for any $k \geq 1$, Assumption M *holds, then*

$$\mathbb{E}\big[\|\mu_n - \mu_n^{N,h}\| \big| Y_{1:n} \big] \leq \delta_n^{\mathrm{TV}} + \frac{2}{\log 3} \sum_{k=1}^{n-1} \tau_{n:k+2} \frac{\delta_k^{\mathrm{TV}}}{\varepsilon_{k+1}^2},$$

where for any $k \geq 1$,

$$\delta_k^{\mathrm{TV}} \leq \left[\frac{1}{\sqrt{N}} + \frac{\beta A_m}{\sqrt{N h^m}} (I_k + h\gamma)^{1/2} \right] \rho_k + \alpha h^2 D_k'' \rho_k''.$$

PROOF. The proof is similar to the proof of Theorem 6.9 except that estimates (30) and (31) are used here with $\Lambda = \Psi_k$. By definition,

$$K_h * \big(\Psi_k \cdot S^N (Q_k \mu_{k-1}^{N,h}) \big) = \begin{cases} \dfrac{K_h * (\Psi_k S^N (Q_k \mu_{k-1}^{N,h}))}{\langle S^N (Q_k \mu_{k-1}^{N,h}), \Psi_k \rangle}, & \text{if } \langle S^N (Q_k \mu_{k-1}^{N,h}), \Psi_k \rangle > 0, \\ K_h * \nu, & \text{otherwise,} \end{cases}$$

and since convolution by K_h preserves the total mass,

$$\langle K_h * (\Psi_k S^N(Q_k \mu_{k-1}^{N,h})), 1\rangle = \langle S^N(Q_k \mu_{k-1}^{N,h}), \Psi_k\rangle.$$

Finally, recall that the bounds in estimates (6) and (7) do not depend on the restarting probability distribution.

To prove the estimate in the weak sense, it is sufficient to bound the local error δ_k^{W} in the weak sense, and to apply Theorem 4.8. Since R_k is mixing, $\langle Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle > 0$ in view of Remark 5.7. Using estimate (6) yields

$$\begin{aligned}
|\langle \mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h}), \phi\rangle| &= |\langle K_h * (\Psi_k \cdot S^N(Q_k \mu_{k-1}^{N,h})) - \Psi_k(Q_k \mu_{k-1}^{N,h}), \phi\rangle| \\
&\le \frac{|\langle K_h * (\Psi_k S^N(Q_k \mu_{k-1}^{N,h})) - \Psi_k(Q_k \mu_{k-1}^{N,h}), \phi\rangle|}{\langle Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle} \\
&\quad + \frac{|\langle S^N(Q_k \mu_{k-1}^{N,h}) - Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle|}{\langle Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle} \|\phi\|,
\end{aligned}$$

for any bounded measurable test function ϕ defined on E. By definition,

$$\langle Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle \ge \inf_{\mu \in \mathcal{P}(E)} \langle Q_k \mu, \Psi_k\rangle.$$

Under Assumption R$''$,

$$\left| \Psi_k \frac{d(Q_k \mu)}{dx} \right|_{2,1} \le \left\| \frac{d(Q_k \mu)}{dx} \right\|_{2,1} \sup_{u \in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}} \le D_k'' \sup_{u \in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}}.$$

Using estimate (30) with $\Lambda = \Psi_k$, $\mu = Q_k \mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(Y_{1:k},\, \mu_{k-1}^{N,h})$, yields

$$\begin{aligned}
&\mathbb{E}\big[|\langle K_h * (\Psi_k S^N(Q_k \mu_{k-1}^{N,h})) - \Psi_k(Q_k \mu_{k-1}^{N,h}), \phi\rangle| \,|\, Y_{1:k},\, \mu_{k-1}^{N,h}\big] \\
&\quad \le \left[\frac{1}{\sqrt{N}} \sup_{x \in E} \Psi_k(x) + \frac{1}{2} \alpha h^2 \left| \Psi_k \frac{d(Q_k \mu_{k-1}^{N,h})}{dx} \right|_{2,1} \right] \|\phi\| \\
&\quad \le \left[\frac{1}{\sqrt{N}} \sup_{x \in E} \Psi_k(x) + \frac{1}{2} \alpha h^2 D_k'' \sup_{u \in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}} \right] \|\phi\|.
\end{aligned}$$

Using estimate (24) with $\phi = \Psi_k$, $\mu = Q_k \mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(Y_{1:k},\, \mu_{k-1}^{N,h})$, yields

$$\mathbb{E}\big[|\langle S^N(Q_k \mu_{k-1}^{N,h}) - Q_k \mu_{k-1}^{N,h}, \Psi_k\rangle| \,|\, Y_{1:k},\, \mu_{k-1}^{N,h}\big] \le \frac{1}{\sqrt{N}} \sup_{x \in E} \Psi_k(x).$$

To prove the estimate in total variation, it is sufficient to bound the local error δ_k^{TV} in total variation and to apply Theorem 4.6. Since convolution by K_h

preserves the total mass, using estimate (7) yields

$$\begin{aligned}\|\mu_k^{N,h} - \bar{R}_k(\mu_{k-1}^{N,h})\| &= \|K_h * (\Psi_k \cdot S^N(Q_k\mu_{k-1}^{N,h})) - \Psi_k \cdot (Q_k\mu_{k-1}^{N,h})\| \\ &\le \frac{\|K_h * (\Psi_k S^N(Q_k\mu_{k-1}^{N,h})) - \Psi_k(Q_k\mu_{k-1}^{N,h})\|}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle} \\ &\quad + \frac{|\langle S^N(Q_k\mu_{k-1}^{N,h}) - Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle|}{\langle Q_k\mu_{k-1}^{N,h}, \Psi_k\rangle}.\end{aligned}$$

Using estimate (31) with $\Lambda = \Psi_k$, $\mu = Q_k\mu_{k-1}^{N,h}$ and $\mathcal{F} = \sigma(Y_{1:k}, \mu_{k-1}^{N,h})$, yields

$$\begin{aligned}&\mathbb{E}[\|K_h * (\Psi_k S^N(Q_k\mu_{k-1}^{N,h})) - \Psi_k(Q_k\mu_{k-1}^{N,h})\| | Y_{1:k}, \mu_{k-1}^{N,h}] \\ &\quad \le \frac{\beta A_m}{\sqrt{Nh^m}}\big(I(Q_k\mu_{k-1}^{N,h}) + h\gamma\big)^{m/2} \sup_{x\in E}\Psi_k(x) + \alpha h^2 \left|\Psi_k \frac{d(Q_k\mu_{k-1}^{N,h})}{dx}\right|_{2,1} \\ &\quad \le \frac{\beta A_m}{\sqrt{Nh^m}}(I_k + h\gamma)^{m/2} \sup_{x\in E}\Psi_k(x) + \alpha h^2 D_k'' \sup_{u\in W^{2,1}} \frac{|\Psi_k u|_{2,1}}{\|u\|_{2,1}}.\end{aligned}$$

□

Acknowledgments. The first author gratefully thanks Marion Baudry, Natacha Caylus and Arnaud Guyader for their careful reading of earlier versions of this work.

REFERENCES

[1] ATAR, R. (1998). Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain. *Ann. Probab.* **26** 1552–1574.

[2] ATAR, R., VIENS, F. and ZEITOUNI, O. (1999). Robustness of Zakai's equation via Feynman–Kac representations. In *Stochastic Analysis*, *Control*, *Optimization and Applications*: *A Volume in Honor of Wendell H. Fleming* (W. M. McEneaney, G. G. Yin and Q. Zhang, eds.) 339–352. Birkhäuser, Boston.

[3] ATAR, R. and ZEITOUNI, O. (1997). Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.* **33** 697–725.

[4] ATAR, R. and ZEITOUNI, O. (1997). Lyapunov exponents for finite state nonlinear filtering. *SIAM J. Control Optim.* **35** 36–55.

[5] BIRKHOFF, G. (1967). *Lattice Theory*, *Colloquium Publications*, 3rd ed. Amer. Math. Soc., Providence, RI.

[6] BUDHIRAJA, A. S. and OCONE, D. L. (1997). Exponential stability of discrete-time filters for bounded observation noise. *Systems Control Lett.* **30** 185–193.

[7] BUDHIRAJA, A. S. and OCONE, D. L. (1999). Exponential stability in discrete-time filtering for nonergodic signals. *Stochastic Process. Appl.* **82** 245–257.

[8] DA PRATO, G., FUHRMAN, M. and MALLIAVIN, P. (1999). Asymptotic ergodicity of the process of conditional law in some problem of nonlinear filtering. *J. Funct. Anal.* **164** 356–377.

[9] 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.

[10] DEL MORAL, P., JACOD, J. and PROTTER, P. (2001). The Monte Carlo method for filtering with discrete-time observations. *Probab. Theory Related Fields* **120** 346–368.

[11] DEL MORAL, P. and MICLO, L. (2000). Branching and interacting particle systems approximations of Feynman–Kac formulae with applications to nonlinear filtering. *Séminaire de Probabilités XXXIV*. *Lecture Notes in Math.* **1729** 1–145. Springer, Berlin.

[12] DEVROYE, L. (1986). *Nonuniform Random Variate Generation*. Springer, New York.

[13] DEVROYE, L. (1987). *A Course on Density Estimation*. Birkhäuser, Boston.

[14] DOUCET, A., DE FREITAS, N. and GORDON, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

[15] GORDON, N. J., SALMOND, D. J. and SMITH, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. on Radar and Signal Processing* **140** 107–113.

[16] HOLMSTRÖM, L. and KLEMELÄ, J. (1992). Asymptotic bounds for the expected L_1 error of a multivariate kernel density estimator. *J. Multivariate Anal.* **42** 245–266.

[17] HOPF, E. (1963). An inequality for positive integral linear operators. *J. Math. Mechanics* **12** 683–692.

[18] HÜRZELER, M. and KÜNSCH, H. R. (1998). Monte Carlo approximations for general state space models. *J. Comput. Graph. Statist.* **7** 175–193.

[19] LE GLAND, F. and MEVEL, L. (2000). Basic properties of the projective product, with application to products of column-allowable nonnegative matrices. *Math. Control Signals Systems* **13** 41–62.

[20] LE GLAND, F. and MEVEL, L. (2000). Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems* **13** 63–93.

[21] LE GLAND, F., MUSSO, C. and OUDJANE, N. (1998). An analysis of regularized interacting particle methods for nonlinear filtering. In *Preprints of the Third IEEE European Workshop on Computer—Intensive Methods in Control and Data Processing*, *Prague 1998* (J. Rojíček, M. Valečkova, M. Kárný and K. Warwick, eds.) 167–174.

[22] 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.

[23] MUSSO, C. and OUDJANE, N. (1998). Regularization schemes for branching particle systems as a numerical solving method of the nonlinear filtering problem. In *Proceedings of the Irish Signals and Systems Conference*, *Dublin 1998*.

[24] MUSSO, C., OUDJANE, N. and LE GLAND, F. (2001). Improving regularized particle filters. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 247–271. Springer, New York.

[25] NEVEU, J. (1975). *Discrete-Parameter Martingales*. North-Holland, Amsterdam.

[26] OCONE, D. L. and PARDOUX, E. (1996). Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM J. Control Optim.* **34** 226–243.

[27] OUDJANE, N. and MUSSO, C. (1999). Multiple model particle filter. In *17ème Colloque GRETSI*, *Vannes 1999* 681–684.

[28] OUDJANE, N. and RUBENTHALER, S. (2003). Stability and uniform particle approximation of nonlinear filters in case of nonergodic signal. Prépublication PMA–786, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris VI. Available at www.proba.jussieu.fr/mathdoc/textes/PMA-786.pdf.

[29] RAVIART, P.-A. (1985). An analysis of particle methods. *Numerical Methods in Fluid Dynamics*, *Como 1983*. *Lecture Notes in Math.* **1127** 243–324. Springer, Berlin.
[30] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

IRISA/INRIA
CAMPUS DE BEAULIEU
35042 RENNES CÉDEX
FRANCE
E-MAIL: legland@irisa.fr

EDF, DIVISION R&D
1 AVENUE DU GÉNÉRAL DE GAULLE
92141 CLAMART CÉDEX
FRANCE
E-MAIL: Nadia.Oudjane@edf.fr