Monte Carlo algorithms and asymptotic problems in nonlinear filtering

A. Budhiraja and H.J. Kushner

1. Introduction

This paper is an extension of [4], which dealt with a wide variety of approximations to optimal nonlinear filters over long time intervals, where *pathwise average* errors are of interest. Suppose that the underlying signal model is a diffusion or jump-diffusion $X(\cdot)$, or a discrete time Markov chain, with white noise corrupted observations. Then one can rarely construct optimal filters, and an approximation must be used. A common method constructs a filter for a simpler process $\tilde{X}^h(\cdot)$, and then uses that but with the actual physical observations. For example, $\tilde{X}^h(\cdot)$ can be a discretized (state or time) form of $X(\cdot)$. It is such that $\tilde{X}^h(\cdot)$ converges weakly to $X(\cdot)$ as $h \to 0$. For each h, the approximating filter $\Pi^h(\cdot)$ is a measure, and it converges weakly to the true conditional distribution process, $\Pi(\cdot)$, as $h \to 0$ [24, 22]. If the filter is to be used over a long large interval $[0, T]$, *pathwise average* errors are often preferred to the mean value. Until further notice, consider the pathwise error on $[0, T]$ for continuous and bounded $\phi(\cdot), f(\cdot)$:

$$G^{h,T}(\phi) = \frac{1}{T} \int_0^T f\left(\phi(X(t)) - \langle \Pi^h(t), \phi \rangle\right) dt. \tag{1.1}$$

Now there are two parameters, h and T. The convergence of $\Pi^h(\cdot)$ over any finite interval says nothing about the behavior of the errors as $h \to 0, T \to \infty$. Under reasonable conditions, it was shown in [4], that the pathwise errors converge (independently of how $h \to 0, T \to \infty$) in probability to a deterministic limit, which is just what one would get for the limit if the *true optimal filter* were used, an ideal result. The paper [4] dealt with a more general setup: The signal process could be an "approximate" diffusion and wide band noise was allowed. In [24] the pathwise average error was replaced by the expectation of the pathwise average error.

A common form of $\tilde{X}^h(\cdot)$ is a Markov chain [23, 22]. When the dimension is larger than three or four, such methods are not now practical. Alternatives, based on random sampling or Monte Carlo then become attractive. The topic is of great current interest; e.g., [9, 10, 11, 14, 15, 16, 17, 26, 27, 28]. Such methods are useful when the transition probabilities are hard to compute, as when the signal is the output of a system with complex dynamics, but which can be simulated. Owing to the sampling errors as well as to the other (computational and modeling) approximations that are made, it is conceivable that the long term pathwise errors will be large, even with approximations that perform well on a bounded time interval.

We extend [4] to such sampling based algorithms for discrete time processes. Similar methods can be used in continuous time [5]. The continuous time problem is of considerable interest because of its role in the Monte-Carlo solution of stochastic PDE's. But, owing to lack of space, we confine ourselves to discrete time. We call the algorithms in [4] *integration* algorithms, since the conditional distributions are computed using integrations or sum approximations. The key ingredients of the proofs are the occupation measure methods of [4]. Section 2 gives the standard form of the optimal filter, formulates the limit problem in terms of occupation measures, and states a basic result of [4] which will be needed. The problem is framed so that the proofs can be made close to those in [4]. Because only the differences in the proofs from those in [4] will be presented, we will outline the proof of the main result in [4]. A main assumption in [4] is *consistency*, which makes precise the convergence of $\tilde{X}^h(\cdot)$ (which is used as a computational approximation in the approximating filter) to $X(\cdot)$ as $h \to 0$. This is (A2.1) here. Weaker forms will be used in the sequel, depending on the form of the random sampling algorithm. The second crucial assumption in [4] is that of uniqueness of invariant measure for the Markov process $(X(\cdot), \Pi(\cdot))$. This is (A2.2) here and is central in the proofs of this paper as well.

Section 3 concerns useful forms of the sampling algorithms. The simplest is based on pure random sampling of the approximating process $\tilde{X}^h(\cdot)$. The scheme can be generalized in many ways. Common variance reduction methods such as antithetic variables and stratified sampling can be used. Combinations of integration and sampling methods are promising, since it might be most convenient to simulate some parts of the problem, but to use "integrations" over distributions of approximating processes in others. To cover such cases, we reformulate the consistency condition so that it holds for all the above examples.

Section 4 concerns importance sampling methods [9, 12, 28]. The standard form is described. But, a more intriguing form uses a measure change which depends on the next observation, which is thus used to guide the simulation on the current time interval [9, 5, 28]. The proofs differ only slightly from that given for the basic example in Section 3.

In Section 5 we consider the problem of uniqueness of the invariant measure for the Markov process $(X(\cdot), \Pi(\cdot))$. We provide a necessary and sufficient condition for this uniqueness. We also show that the approximate filters considered in this paper enjoy another important infinite horizon property. This property says that the pathwise distance, per unit time, of the optimal filter from the approximate filter is small asymptotically in time for small h.

2. The Discrete Time Model

For any discrete parameter process Z_n, define $Z_{0,n} = \{Z_i, i \leq n\}$. The signal process $X(\cdot) = \{X(n), n < \infty\}$ is assumed to be Feller–Markov and take values in a compact set G. The observations are defined by $Y(0) = 0$ and

$$\delta Y_n \equiv Y(n) - Y(n-1) = g(X(n)) + \xi(n),\ n = 1, \ldots, \tag{2.1}$$

where $\xi(n)$ are mutually independent $(0, I)-$Gaussian random variables which are independent of $X(\cdot)$, and $g(\cdot)$ is continuous.

The Bayes' rule formula for the conditional distribution of $X(n)$ given $Y_{0,n}$ can be represented in terms of an auxiliary process $\tilde{X}(\cdot)$, where $\tilde{X}(\cdot)$ has the same evolution law as that of $X(\cdot)$ but (conditioned on its possibly random initial distribution) is independent of all the other processes. Define

$$R(\tilde{X}_{0,n}, Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}(i))\delta Y_i - \frac{1}{2}\sum_{i=1}^{n}\left|g(\tilde{X}(i))\right|^2\right].$$

Then the optimal filter $\Pi(\cdot)$ can be defined by its moments:

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(0), Y_{0,n}\}}\left[\phi(\tilde{X}(n))R(\tilde{X}_{0,n}, Y_{0,n})\right]}{E_{\{\Pi(0), Y_{0,n}\}}R(\tilde{X}_{0,n}, Y_{0,n})},$$

where $\Pi(0)$ is the distribution of $X(0)$ and of $\tilde{X}(0)$. Alternatively,

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(n-1), \delta Y_n\}}\left[\phi(\tilde{X}(1))R(\tilde{X}(1), \delta Y_n)\right]}{E_{\{\Pi(n-1), \delta Y_n\}}R(\tilde{X}(1), \delta Y_n)}, \tag{2.2}$$

where with an abuse of notation, we have defined $R(x, y) = \exp\left[g(x)'y - |g(x)|^2/2\right]$. In (2.2), $\Pi(n-1)$ is the (possibly random) distribution of the initial condition $\tilde{X}(0)$. Throughout the paper, we use the notation $E_{\{\Pi(k),Y_{k,l}\}}F(\tilde{X}_{0,l-k},Y_{k,l}), l \geq k$, for the conditional expectation of a function of $\tilde{X}_{k,l},Y_{k,l}$, given the data $Y_{k,l}$ and where the (random) distribution of $\tilde{X}(\cdot)$ is $\Pi(k)$. The analogous notation will be used when approximations to $\tilde{X}(\cdot)$ are used.

One cannot usually evaluate (2.2), and some approximation is needed. Commonly, one builds a filter for a simpler Markov process $\tilde{X}^h(\cdot)$ which has values in the compact set G, and which approximates $X(\cdot)$, but in which the actual physical observations are used, as developed in [20, 23]. More particularly, proceed as follows. Define

$$R(\tilde{X}^h_{0,n},Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}^h(i))\delta Y_i - \frac{1}{2}\sum_{i=1}^{n}\left|g(\tilde{X}^h(i))\right|^2\right].$$

Then define the approximating filter $\Pi^h(\cdot)$ by its moments:

$$\langle \Pi^h(n),\phi\rangle = \frac{E_{\{\Pi^h(0),Y_{0,n}\}}\left[\phi(\tilde{X}^h(n))R(\tilde{X}^h_{0,n},Y_{0,n})\right]}{E_{\{\Pi^h(0),Y_{0,n}\}}R(\tilde{X}^h_{0,n},Y_{0,n})}, \tag{2.3}$$

or use the recursive representation:

$$\langle \Pi^h(n),\phi\rangle = \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}^h(1),\delta Y_n)}. \tag{2.4}$$

We have constructed a filter for $\tilde{X}^h(\cdot)$ but used the observations δY_n.

The process $\tilde{X}^h(\cdot)$ in (2.3) and (2.4) is assumed to be independent of the other processes, given its initial distribution. We also use the following:

A2.1. A consistency assumption. For any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure Π, $(\tilde{X}^h(0),\tilde{X}^h(1))$ with the initial distribution Π^h converges weakly to $(\tilde{X}(0),\tilde{X}(1))$ with the initial distribution Π.

The process $\Pi(\cdot)$ in (2.2) is well defined even if $\Pi(0)$ is not the distribution of $X(0)$. Allowing $(X(0),\Pi(0))$ to be arbitrary, $\Psi_f(\cdot) = (X(\cdot),\Pi(\cdot))$ is Feller–Markov. We also need:

A2.2. A uniqueness assumption. There is a unique stationary process $\Psi_f(\cdot) = (X(\cdot),\Pi(\cdot))$. Denote its measure by $\bar{Q}_f(\cdot)$.

The occupation measure. For each n, define $B(n) = \sum_{i=1}^{n}\xi(i)$ and

$$\Psi^h(n,\cdot) = \left\{X(n+\cdot),\Pi^h(n+\cdot),,Y(n+\cdot)-Y(n),B(n+\cdot)-B(n)\right\},$$

$$\Psi^h_f(n,\cdot) = \{X(n+\cdot),\Pi^h(n+\cdot)\}.$$

When discussing weak convergence of processes such as $\Psi^h(n,\cdot)$,we use a "sequence" topology, as in [4]. The $\Pi^h(n)$ take values in $\mathcal{M}(G)$, the space of probability measures on G, and the weak topology is used on this space.

For a random variable Z and set A, let $I_A(Z)$ denote the indicator function of the event that $Z \in A$. Define the occupation measure $Q^{h,N}(\cdot)$ as follows. For a Borel set C in the product sequence space,

$$Q^{h,N}(C) = \frac{1}{N}\sum_{n=1}^{N} I_C(\Psi^h(n,\cdot)). \tag{2.6}$$

Define $\Psi(\cdot) = (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot))$. In the sequel, lower case letters $x(\cdot), \pi(\cdot)$, etc., are used for the canonical sample paths. Letters such as $x, y, \ldots,$ are used to denote vectors such as $x(n), y(n)$, etc. Define $\psi_f(\cdot) = (x(\cdot), \pi(\cdot))$, $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$ and $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot), w(\cdot))$. Let $F(\cdot)$ be a real–valued bounded and continuous (with probability one with respect to $\bar{Q}_f(\cdot)$) function of $\psi_f(\cdot)$. Then, the following is proved in [4].

Theorem 2.1. *Let the filtering model be as above. Assume that* (A2.2) *holds. Define the approximate filter via* (2.3), *where* $\tilde{X}^h(\cdot)$ *satisfies* (A2.1). *Then* $\{Q^{h,N}(\cdot); h > 0, N \geq 0\}$ *is tight. Let* $Q(\cdot)$ *denote a weak sense limit, as* $h \to 0$ *and* $N \to \infty$. *Let* ω *be the canonical variable on the probability space on which* $Q(\cdot)$ *is defined, and denote the sample values by* $Q^\omega(\cdot)$. *Then, for each* ω, $Q^\omega(\cdot)$ *induces a process*

$$\Psi^\omega(\cdot) = (X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot)). \tag{2.7}$$

For almost all ω *the following hold.* $(X^\omega(\cdot), \Pi^\omega(\cdot))$ *is stationary.* $B^\omega(\cdot)$ *is the sum of mutually independent* $N(0, I)$ *random variables* $\{\xi^\omega(n)\}$ *which are independent of* $X^\omega(\cdot)$. *Also*

$$\delta Y_n^\omega \equiv Y^\omega(n) - Y^\omega(n-1) = g(X^\omega(n)) + \xi^\omega(n), \tag{2.8}$$

and $X^\omega(\cdot)$ *has the transition probability function of* $X(\cdot)$. *For each integer* n *and each bounded and measurable real–valued function* $\phi(\cdot)$

$$\langle \Pi^\omega(n), \phi\rangle = \frac{E_{\{\Pi^\omega(0), Y_{0,n}^\omega\}}\left[\phi(\tilde{X}(n)) R(\tilde{X}_{0,n}, Y_{0,n}^\omega)\right]}{E_{\{\Pi^\omega(0), Y_{0,n}^\omega\}} R(\tilde{X}_{0,n}, Y_{0,n}^\omega)}. \tag{2.9}$$

Finally, in probability and where $h \to 0, N \to \infty$ *in any way at all.*

$$\frac{1}{N}\sum_{n=1}^{N} F(\Psi_f^h(n+\cdot)) = \int F(\psi_f(\cdot))\bar{Q}_f^{h,N}(d\psi_f(\cdot)) \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)). \tag{2.10}$$

Discussion of the proof. In the problems of this paper, the approximating filter will be constructed using random sampling methods or combinations of random sampling and integration methods. Since many arguments in the proofs are similar to those used in [4], we will refer to the proof in [4] as much as possible, and to concentrate on the differences. In view of that we now highlight the chief features of the proof in [4]. Analogous remarks hold for the continuous parameter model. Further details are in the reference. In [4] (see (2.6)), the measure valued random variable $Q^{h,N}(\cdot)$ was obtained as an occupation measure connected with the processes $X(\cdot), \Pi^h(\cdot), B(\cdot), Y(\cdot)$, and the same definition will be used in the sequel, but with the new definitions of $\Pi^h(\cdot)$ of this paper used.

The first step is to show that the sequence $\{Q^{h,N}(\cdot); h, N\}$ of measure valued random variables is tight. For that it suffices to show that sequence of its expectations is tight [21, Chapter 1.6]. It is enough to show that the families $\{X(n+\cdot); n \geq 0\}$,$\{B(n+\cdot)-B(n); n \geq 0\}$, $\{Y(n+\cdot)-Y(n); n \geq 0\}$, $\{\Pi^h(n+\cdot); h > 0, n \geq 0\}$ are tight. However, this is trivial in view of the compactness of the state space.

By the first equality in (2.10), the limit is determined by the weak sense limits of the occupation measures, as $N \to \infty, h \to 0$. Thus, we need to determine the sample values $Q^\omega(\cdot)$ of any weak sense limit $Q(\cdot)$. Equivalently, we need to characterize the set of processes induced by $Q^\omega(\cdot)$. The proof of the stationarity of the $(X^\omega(\cdot), \Pi^\omega(\cdot))$ in [4] will work without any change for the problems of this paper. The proofs of the representation (2.8) and that $X^\omega(\cdot)$ has the law of evolution of $X(\cdot)$ for almost all ω are also as in [4]. Thus, establishing the representation (2.9) is the only step in the proof that will be different from that in [4]. Once this step is established, (2.10) follows from the uniqueness assumption on the invariant measure $\bar{Q}_f(\cdot)$, of the joint signal and filter process.

The proof of the representation for the $\Pi^\omega(\cdot)$ will depend on the choice of $\Pi^h(\cdot)$. The following comments concerning a key detail in the proof of the representation (2.9) in [4] will be useful in providing a guide to the proofs for the cases of this paper.

For arbitrary $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot))$, and integer m define the function $A(\cdot)$ by

$$A(\psi(m)) = \langle \pi(m), \phi \rangle - \frac{E_{\{\pi(m-1),y(m)\}}\left[\phi(\tilde{X}(1))R(\tilde{X}(1), y(m))\right]}{E_{\{\pi(m-1),y(m)\}}R(\tilde{X}(1), y(m))}. \tag{2.11}$$

The aim of the proof in [4] was to show that, for almost all ω and all m, $A(\Psi^\omega(m)) = 0,$ with probability 1 which implies (2.9). This was done by showing that

$$0 = E \int Q^{\omega}(d\psi)\,[A(\psi(m))]_1^2, \tag{2.12}$$

where $[A]_1^2 = \min\{|A|^2, 1\}$. The prelimit form of the right side of (2.12) is $E\int Q^{h,N}(d\psi)\,[A(\psi(m))]_1^2$. By the definition of $Q^{h,N}(\cdot)$, this equals $E\sum_{n=1}^{N}\left[A(\Psi^h(m+n))\right]_1^2/N$, where

$$A(\Psi^h(n)) = \langle \Pi^h(n), \phi\rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}(1))R(\tilde{X}(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}(1),\delta Y_n)}. \tag{2.13}$$

Thus, to show (2.12) it suffices to show that $E[A(\Psi^h(n))]_1^2 \to 0$, uniformly in n as $h \to 0$. Finally to show the above it suffices, in view of tightness of the families $\{\Pi^h(n); h > 0, n > 0\}$, $\{\delta Y_n; n > 0\}$ and the consistency assumption (A2.1), to show that

$$E\left[\langle \Pi^h(n), \phi\rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}^h(1),\delta Y_n)}\right]_1^2 \tag{2.14}$$

converges to 0, uniformly in n as $h \to 0$. However, by the definition of $\Pi^h(n)$ via (2.4) the above expression is identically zero. Hence, (2.12) holds for any weak sense limit. Analogs of this argument will be used in the sequel.

3. Some Approximating Filters

In [4], the approximate filter $\Pi^h(\cdot)$ was defined by (2.4). One example is the Markov chain approximation method [20, 23], where $\tilde{X}^h(\cdot)$ is a Markov chain approximation to $\tilde{X}(\cdot)$. If the dimension is high, this can have high computational requirements and Monte Carlo becomes attractive [9, 10, 11, 14, 15, 16, 17, 26, 27, 28]. Several forms will be discussed, starting with the most basic, which uses simple random sampling to evaluate (2.4). The problem is set up so that much of the proof of [4, Theorem 5.1] (our Theorem 2.1) can be used. Then, we develop more general approximations.

Example 1. The basic random sampling filter. Let $v^h \to \infty$ be integers, and $\Pi^h(n-1)$ the estimate of the conditional distribution of $X(n-1)$, given $Y_{0,n-1}$. Given $\Pi^h(n-1)$, construct $\Pi^h(n)$ based on "random sampling," as follows. Let $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ be i.i.d samples (independent of δY_n, conditioned on $\Pi^h(n-1)$) from $\tilde{X}^h(\cdot)$, where $\tilde{X}^h(\cdot)$ has initial distribution $\Pi^h(n-1)$. We need only simulate $\tilde{X}^h(0), \tilde{X}^h(1)$. The approximating filter $\Pi^h(n)$ is defined by the sample average:

$$\langle \Pi^h(n), \phi \rangle = \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}. \tag{3.1}$$

Theorem 3.1. *Under* (A2.1) *and* (A2.2) *and the above construction of* $\Pi^h(\cdot)$*, the conclusions of Theorem* 2.1 *hold.*

Proof. The proof of Theorem 2.1 was outlined after its statement. The proof here is similar, and only the differences will be discussed. As stated earlier the only statement that needs to be verified is (2.9) for which it suffices to show that (2.12) holds and which in turn follows if the expression in (2.14) converges to 0 uniformly in n as h converges to 0. By the definition of $\Pi^h(n)$, (2.14) is

$$\begin{aligned} E \Bigg[& \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h} \\ & - \frac{E_{\{\Pi^h(n-1), \delta Y_n\}} \left[\phi(\tilde{X}^h(1)) R(\tilde{X}^h(1), \delta Y_n) \right]}{E_{\{\Pi^h(n-1), \delta Y_n\}} R(\tilde{X}^h(1), \delta Y_n)} \Bigg]_1^2 . \end{aligned} \tag{3.2}$$

We need only show that, for any bounded, continuous and real-valued $\phi(\cdot)$,

$$\begin{aligned} \lim_{h\to 0} \limsup_n E & \left[\frac{1}{v^h} \sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n) \right. \\ & \left. - E_{\{\Pi^h(n-1), \delta Y_n\}} \phi(\tilde{X}^h(1)) R(\tilde{X}^h(1), \delta Y_n) \right]_1^2 = 0. \end{aligned} \tag{3.4}$$

This holds since for each h and n, $\{\tilde{X}^{h,l,n}(\cdot), l\}$ are mutually independent, identically distributed and independent of δY_n (conditioned on $\Pi^h(n-1)$), and the mean square value (conditional on $\{\Pi^h(n-1), \delta Y_n\}$) of the functional

$$\phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n) - E_{\{\Pi^h(n-1), \delta Y_n\}} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)$$

has uniformly (in h, l, n) bounded expectation. ■

In view of (A2.1), the expression in (3.4) is asymptotically equivalent to the following, which is often more convenient.

$$E\left[\frac{1}{v^h}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)\right. \\ \left. -E_{\{\Pi^h(n-1),\delta Y_n\}}\phi(\tilde{X}(1))R(\tilde{X}(1),\delta Y_n)\right]_1^2. \tag{3.4$'$}$$

Example 2. Some generalizations of the filter in Example 1. The crucial step in the proof is showing (3.4′), the form that we will use. This convergence is essentially a consequence of (A2.1), which can be weakened considerably to get many useful extensions of the basic algorithm of Example 1.

A weaker form of (A2.1). We keep the assumption of mutual independence (conditional on $\Pi^h(n-1),\delta Y_n$) of the $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ for each h, n, and that the probability law of $\{\tilde{X}^{h,l,n}(0)\}$ is $\Pi^h(n-1)$, but we allow greater flexibility in the choice of the individual $\tilde{X}^{h,l,n}(\cdot)$. Namely, in the construction of $\Pi^h(n)$ in (3.1), the Markov family from which $\tilde{X}^{h,l,n}(\cdot)$ is sampled might depend on l, n. But, the $\tilde{X}^{h,l,n}(0)$ still form an i.i.d sample from $\Pi^h(n-1)$. To see the advantages, first rewrite the expression in (3.2) as the sum of

$$\frac{\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h} \\ -\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}, \tag{3.5}$$

and

$$\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h} \\ -\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}(1),\delta Y_n)}. \tag{3.6}$$

It is enough to work separately with the differences of the numerators and of the denominators in these expressions. To handle (3.5), use the mutual independence and the uniform bounds on the expectations of the conditional variances. To handle (3.6), we will use a revised form of (A2.1), namely:

A3.1. For each (n, h), the set $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is mutually independent and independent of δY_n, conditioned on $\Pi^h(n-1)$. Suppose that an arbitrary

Π^h replaces $\Pi^h(n-1)$ in the construction of the $\{\tilde{X}^{h,l,n}(0), l\}$. Then, as $h \to 0$, for any such sequence, and for each bounded, continuous and real valued $\Phi(\cdot)$,

$$E_{\Pi^h}\Phi(\tilde{X}^{h,l,n}(1)) - E_{\Pi^h}\Phi(\tilde{X}(1)) \to 0 \text{ uniformly in } n, l. \tag{3.7}$$

This assumption, when used for $\Phi(x) \equiv \Phi_y(x) = \phi(x)R(x,y)$, for each fixed y, leads to the desired convergence for the expression in (3.6). Even though $R(\cdot)$ is not bounded we can assume so, since $\{\delta Y_n; n \geq 1\}$ is tight. Thus, we do not need the convergence in (A3.1) for $\Phi(\cdot) = \Phi_y(\cdot)$ to hold uniformly in y.

Note that we are no longer assuming that the $\tilde{X}^{h,l,n}(\cdot)$ are all samples of the same $\tilde{X}^h(\cdot)$ process. The second part of (A3.1) will hold iff for all Π and any sequence $\{h_k, l_k, n_k\}_{k\geq 1}$ for which ($\mathcal{L}(X)$ denotes the probability law of X) $\mathcal{L}(\tilde{X}^{h_k,l_k,n_k}(0)) \Rightarrow \Pi$, as $k \to \infty$, we have that

$$\mathcal{L}(\tilde{X}^{h_k,l_k,n_k}(0), \tilde{X}^{h_k,l_k,n_k}(1)) \Rightarrow \mathcal{L}(\tilde{X}(0), \tilde{X}(1)) \text{ as } k \to \infty, \ \mathcal{L}(\tilde{X}(0)) = \Pi.$$

Drop the mutual conditional independence. Return to (3.2) and let $\Phi(\cdot)$ be bounded and continuous. Then the convergence in (3.2) is implied by the weaker consistency assumption, which can replace (A2.1) and the mutual independence in Theorem 3.1:

A3.2. For each (h,n), $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is independent of δY_n, conditioned on $\Pi^h(n-1)$, but they might not be independent in l. They are constructed subject to the following rule. Suppose that an arbitrary measure $\Pi^{h,n}$ takes the role of $\Pi^h(n-1)$ in the construction of $\{\tilde{X}^{h,l,n}(\cdot), l\}$. Then $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is constructed such that as $h \to 0$ and, for any bounded, continuous $\Phi(\cdot)$, in probability,

$$\frac{1}{v^h}\sum_{l=1}^{v^h} \Phi(\tilde{X}^{h,l,n}(1)) - E_{\{\Pi^{h,n}\}}\Phi(\tilde{X}(1)) \to 0, \text{ uniformly in } n. \tag{3.8}$$

This condition (instead of (A2.1) and the mutual independence of the samples) suffices for Theorem 3.1. This condition is useful where the samples $\{\tilde{X}^{h,l,n}(\cdot)), l \leq v^h\}$ are not mutually independent, as in variance reduction methods; e.g., antithetic variables or stratified sampling.

Variance reduction methods. The common methods for Monte Carlo variance reduction, such as stratified sampling and antithetic variables, can all be used, and we briefly discuss one form of stratified sampling to see the possibilities. Let $\Pi^h(n-1)$ be concentrated on points $\{x^{h,l,n}; l = 1, \ldots, v^h\}$, and let $\Pi^h_l(n-1)$ denote the weight that $\Pi^h(n-1)$ puts on $x^{h,l,n}$. For illustrative purposes, we will use variance reduction only on the samples of $\tilde{X}^{h,l,n}(0)$. Once these are given, the samples of $\tilde{X}^{h,l,n}(1)$ are obtained by independent sampling, using the transition probability of $\tilde{X}^{h,n}(\cdot)$.

If the $v^h \Pi^h_l(n-1)$ were all integers, then the optimal (zero sampling variance) sampling of $\tilde{X}^{h,l,n}(0)$ takes the initial point $x^{h,l,n}$ exactly $v^h \Pi^h_l(n-1)$ times. If some $v^h \Pi^h_l(n-1)$ is not an integer, one tries to approximate this as well as possible. The following is a typical method. First take $x^{h,l,n}$ exactly $[v^h \Pi^h_l(n-1)]$ (the integer part) times. We must still choose $\delta v^{h,n} = \sum_l \delta v^{h,n}_l$ points, where $\delta v^{h,n}_l = v^h \Pi^h_l(n-1) - [v^h \Pi^h_l(n-1)]$. The "residual frequency" of $x^{h,l,n}$ is $\delta v^{h,n}_l / \delta v^{h,n}$. Divide the set $\{x^{h,l,n}, l\}$ into disjoint subsets $S^{h,n}_i, i = 1, \ldots$, where $S^{h,n}_i$ has $[\delta v^{h,n,i}]$ points, where

$$\delta v^{h,n,i} = \sum_{l \in S^{h,n}_i} \delta v^{h,n}_l .$$

Allocate $[\delta v^{h,n,i}]$ points to subset i, and select these randomly (with replacement) from $S^{h,n}_i$, where $x^{h,l,n} \in S^{h,n}_i$ has weight $\delta v^{h,n}_l / \delta v^{h,n,i}$. Since $\bar{v}^{h,n} \equiv \delta v^{h,n} - \sum_i [\delta v^{h,n,i}] \geq 0$, we still need to allocate $\bar{v}^{h,n}$ points, if this is positive. Generally, the division into subgroups is done such that $\bar{v}^{h,n}/v^h$ is small. If it is positive, either repeat the above procedure, or just use random selection. The above construction can be put in the framework of Example 2 and condition (A3.2) holds. The technique can make a noticeable difference in applications.

Example 3. We would like to use algorithms that allow useful combinations of random sampling and integration methods. This will require an alteration in (A3.1) or (A3.2). To motivate the new assumption, consider an example for which (A3.2) holds. Let $\tilde{X}^{h,n}(\cdot)$ satisfy (A3.1). Given approximations $\Pi^h(j)$ for $j = 1, 2, \cdots, n-1$, suppose that $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ are samples of $\tilde{X}^{h,n}(\cdot)$ and are conditionally independent of δY_n given $\Pi^h(n-1)$. Define $\Pi^h(n)$ via (3.1). If the samples are mutually independent (conditioned on $\Pi^h(n-1)$) and $\tilde{X}^{h,n}(0)$ has distribution $\Pi^h(n-1)$, then (A3.1) (hence (A3.2)) holds. Theorem 3.1 can be proved under a weaker condition than (A3.1) or (A3.2), which allows greater flexibility. To motivate a useful form, rewrite Example 2 as follows.

For each h and n, define a measure (on $G \times G$) valued random variable $P^{h,n}_{\Pi^h(n-1)}$ as follows. Let $P^{h,n}_{\Pi^h(n-1)}(A)$ be the fraction of the samples $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ in $A \subset G \times G$. In particular, $P^{h,n}_{\Pi^h(n-1)}\{B \times G\}$ is the fraction of the samples $\tilde{X}^{h,l,n}(0)$ in the set B. Condition (A3.2) is equivalent to, uniformly in n,

$$\lim_{h\to 0} E\left[\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1))\right]_1^2 = 0. \tag{3.9}$$

The critical condition is thus the convergence in (3.9). The way that (3.9) is written in terms of a random measure, $P^{h,n}_{\Pi^h(n-1)}$, suggests approximate filters that need not be based exclusively on Monte Carlo. The $P^{h,n}_{\Pi^h(n-1)}$ might be determined in part by random sampling and in part analytically.

A generalization of $P^{h,n}_{\Pi^h(n-1)}$ and the approximating filter. Motivated by (3.9), consider the following general form of the filter and consistency condition. Define $\{\Pi^h(n); n \geq 1\}$ recursively as follows. Given $\Pi^h(n-1)$, let $P^{h,n}_{\Pi^h(n-1)}$ be a measure–valued random variable on the sample space $G \times G$, which is conditionally independent of δY_n given $\Pi^h(n-1)$. Define $\Pi^h(n)$ by

$$\langle \Pi^h(n), \phi\rangle = \frac{\int \phi(x(1))R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}{\int R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}. \tag{3.10}$$

We will use the following weaker consistency condition.

A3.3. For each bounded, continuous $\Phi(\cdot)$

$$\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1)), \tag{3.11}$$

converges to 0, in probability, uniformly in n as $h \to 0$..

The following useful result follows from the above comments.

Theorem 3.2. *Theorem* 3.1 *holds for the* $\Pi^h(\cdot)$ *constructed above, if* (A3.3) *replaces* (A2.1) *and the mutual independence of the samples.*

(A3.3) can be used for a large variety of approximation methods. For example, in the form (2.4), $P^{h.n}_{\Pi^h(n-1)}$ would be the measure of $(\tilde{X}^h(0), \tilde{X}^h(1))$ with $\tilde{X}^h(0)$ having the (random) distribution $\Pi^h(n-1)$. The conditions for the convergence for the classical Markov chain approximation, the random sampling method above and various combinations of them, either in the same or in different time intervals can all be put into the form of (3.11) for appropriate choices of $P^{h.n}_{\Pi^h(n-1)}$. Importance sampling methods can also fit into the scheme and improve the filter performance.

Example 4. An application of (A3.3): Combined random sampling and integration. Consider the following common case Let $X(n) =$

$b(X(n-1), \zeta(n-1))$, where $b(\cdot)$ is bounded and continuous and the $\{\zeta(n)\}$ are i.i.d. (distribution function P_ζ, with compact support), and independent of $X(0)$. First, suppose that $\Pi(n-1)$ is the actual conditional distribution of $X(n-1)$, given $Y_{0,n-1}$. Then the optimal $\Pi(n)$ is defined by (2.2). If the computation on the right side of (2.2) is hard it needs to be approximated. The computational difficulties might be due to the computation of the transition probability of $\{X(n)\}$, or to integrations over a continuous state space that are required to evaluate (2.2). Let $\Pi^h(n)$ the estimate of the conditional distribution given $Y_{0,n}$.

Let $\Pi^h(n-1)$ be given. The computation of $\Pi^h(n)$ can be done by a direct simulation (Example 1) or by combined "simulation-integration" or perhaps even by an "integration" method, as we now show. First approximate P_ζ by a computationally more convenient P^h_ζ, where $P^h_\zeta \Rightarrow P_\zeta$. Then, approximate $b(\cdot)$ by $b_h(\cdot)$ such that $\lim_{h\to 0}\sup_{x,\zeta}|b(x,\zeta) - b_h(x,\zeta)| = 0$. If the associated integrations are not hard, one can use (2.3) to define $\Pi^h(n)$, where

$$\tilde{X}^h(1) = b_h(\tilde{X}^h(0), \zeta^h), \tag{3.13a}$$

In (3.13a), $\tilde{X}^h(0)$ has distribution $\Pi^h(n-1)$ and ζ^h has distribution P^h_ζ. If the supports are finite, then the integrations are summations. This $\tilde{X}^h(\cdot)$ process satisfies (A2.1). Hence Theorem 2.1 holds for $\Pi^h(\cdot)$ if (A2.2) holds. Alternatively, use Monte Carlo (Example 1). The sampling is "conditionally independent" of the past, given $\Pi^h(n-1)$. Take v^h independent samples from $\Pi^h(n-1)$ and P^h_ζ, call them $\tilde{X}^{h,n,l}(0)$ and $\zeta^{h,n,l}$, $l \le v^h$, resp. Then get $\Pi^h(n)$ from (3.1) and use

$$\tilde{X}^{h,n,l}(1) = b_h(\tilde{X}^{h,n,l}(0), \zeta^{h,n,l}). \tag{3.13b}$$

Combinations of the above approaches are also useful. E.g., if the support of the P^h_ζ is a reasonable finite set, one can sample from $\Pi^h(n-1)$, but "integrate" over the noise for each sample of the initial condition. Alternatively, one might discretize the state space such that the support of the $\tilde{X}^h(0)$ is confined to a finite set, and integrate with respect to the "initial" measure $\Pi^h(n-1)$, but simulate the noise. For each combination, there is a $P^{h,n}_{\Pi^h(n-1)}$ such that (A3.3) holds, provided that the discretization of the space converges to the whole space in an appropriate manner and, the number of samples goes to infinity as $h \to 0$. The construction of $P^{h,n}_{\Pi^h(n-1)}$ is not hard and the details are omitted. Another potentially useful combination of integration and simulation, based on the Markov chain approximation method [23] is in [5]. Random sampling is used in the part of the state space where the analytic computations are too hard, and the combined approach is shown to satisfy our conditions.

4. Importance Sampling Methods

Importance sampling methods are used to improve the performance of Monte Carlo algorithms [12, 13]. Suppose that we wish to estimate $Ef(X)$ via Monte Carlo, where $f(\cdot)$ is bounded and continuous and X has distribution P. The simplest estimate has the form $\sum_{i=1}^{n} f(X_i)/n$ where $\{X_i\}$ are mutually independent and chosen at random from P. Suppose that we know that values of X in a set A have the dominant effect on $Ef(X)$, where A has small probability. Then large n would be needed to get a good estimate. Let Q be mutually absolutely continuous with respect to P, and where $Q(A)$ has a "moderate" value. Then for appropriate choice of Q, the unbiased estimate $\sum_{i=1}^{n} f(X_i)[dP/dQ](X_i)/n$ has a much smaller variance than the original estimate, where now the X_i are drawn at random and independently from Q [12, 13]). Importance sampling samples more (using Q) in the region which prior information suggests is more important. It has also been used to improve nonlinear filtering algorithms that use random sampling [9, 28]. We next discuss the general idea and show that the proof of convergence is covered by what has already been done. In the next example, we describe importance sampling on a typical interval $[n-1, n]$ and it does not use the next observation δY_n. This next observation can provide useful information to guide the sampling. We only illustrate some possibilities. Variance reduction methods can be added as can combined sampling–integration methods.

Example 5: Importance sampling. Let us start with the setup in Example 1. Let P^h_{n-1} denote the probability law of $\tilde{X}^h(\cdot) = (\tilde{X}^h(0), \tilde{X}^h(1))$ when $\Pi^h(n-1)$ is the measure of $\tilde{X}^h(0)$. Construct a measure $M^{h,n}$, which is mutually absolutely continuous with respect to P^h_{n-1}. In fact, we can allow $M^{h,n}$ to be random in the sense that it is not a deterministic function of P^h_{n-1}. (See the comments at the end of the section.) In this case, $M^{h,n}$ is assumed to be a.s. mutually absolutely continuous with respect to P^h_{n-1}. For example, $M^{h,n}$ might depend on some other observations which are not used for the filter. Now, take v_h samples, called $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$, from $M^{h,n}$ (or from the current sample of, if it is random). Suppose that they are mutually independent, conditioned on $\Pi^h(n-1), \delta Y_n$.

Define the likelihood ratio (the Radon–Nikodym derivative) and its value on the random path $\tilde{X}^{h,k,n}(\cdot)$ by

$$L^{h,n} = \frac{dP^h_{n-1}}{dM^{h,n}}, \quad L^{h,k,n} = \frac{dP^h_{n-1}}{dM^{h,n}} \left(\tilde{X}^{h,k,n}(\cdot) \right). \tag{4.1}$$

We will use the following assumption.

A4.1.

$$\sup_{h,n} E\frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^h(0),\tilde{X}^h(1))R^2(\tilde{X}^h(1),\delta Y_n) < \infty, \tag{4.2}$$

where $\tilde{X}^h(\cdot)$ in (4.2) has the distribution P^h_{n-1} (conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$).

Define the approximate filter $\Pi^h(\cdot)$ to be:

$$\langle \Pi^h(n),\phi\rangle = \frac{\sum_{l=1}^{v^h} L^{h,k,n}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{\sum_{l=1}^{v^h} L^{h,k,n}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}, \tag{4.3}$$

Theorem 4.1. *Assume* (A2.1), (A2.2) *and* (A4.1) *and the filter form* (4.3). *Then Theorem* 3.1 *holds.*

Proof. To prove the theorem, it suffices to show that

$$E\left[\frac{1}{v^h}\sum_{k=1}^{v_h}\Big(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)\Big)\right]^2 \tag{4.4}$$

converges to zero uniformly in n, as $h \to 0$, where $\Phi(x,y) = \phi(x)R(x,y)$ and $\phi(\cdot)$ is any bounded and continuous function.

The random variables $\tilde{X}^{h,k,n}(\cdot)$ have distribution $M^{h,n}$ and are conditionally mutually independent, given $\Pi^h(n-1), \delta Y_n$. Note that

$$\begin{aligned} &E^{M^{h,n}}_{\Pi^h(n-1),\delta Y_n} L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) \\ &\qquad = E^{P^h_{n-1}}_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n) = E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n). \end{aligned} \tag{4.5}$$

The superscripts in the first two terms denote the measure and the subscripts the conditioning data. The first equality in (4.5) follows by the definition of the Radon–Nikodym derivative. The second equality is simply a statement of the fact that all we need to know about $\tilde{X}^h(\cdot)$ to compute the expectation is its initial distribution and the one step law of evolution. Using (4.5) and the mutual conditional independence, (4.4) can be rewritten as

$$\frac{1}{(v^h)^2}E\sum_{k=1}^{v_h}\Big[L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n). - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)\Big]^2. \tag{4.6}$$

The expectation of a typical summand of (4.6) is bounded by

$$2EL^{h,k,n}\Big[L^{h,k,n}\Phi^2(\tilde{X}^{h,k,n}(1),\delta Y_n)\Big] + 2E\Big|E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)\Big|^2. \tag{4.7}$$

The right hand term of (4.7) is bounded uniformly in n, h. Using the definition of the Radon–Nikodym derivative, the left hand term is

$$O(1) E L^{h,n} \left[L^{h,n} R^2(\tilde{X}^{h,n}(1), \delta Y_n) \right],$$

where $L^{h,n}$ and $\tilde{X}^{h,n}(1)$ are the canonical values of $L^{h,k,n}$ and $\tilde{X}^{h,k,n}(1)$, resp. This can be bounded by

$$O(1) E \left[\frac{dP^h_{n-1}}{dM^{h,n}} \left(\tilde{X}^h(0), \tilde{X}^h(1) \right) R^2(\tilde{X}^h(1), \delta Y_n) \right],$$

where $\tilde{X}^h(\cdot)$ is as in (A4.1). By (A4.2), the above expression is $O(1)$, which yields the theorem. ■

Dropping the mutual independence. As in Example 2, where we relaxed the condition on mutual independence by using (A3.2), we now give a similar condition which can be used to incorporate variance reduction methods and importance sampling. Let $M^{h,n}$ be as before. Also let $\{\tilde{X}^{h,l,n}(\cdot); l \leq v^h\}$ be as before, except that they need not be (conditionally) mutually independent. Assume (A4.2) instead of (A4.1).

A4.2. Let $\Phi(x, y) = \phi(x) R(x, y)$, where $\phi(\cdot)$ is a bounded and continuous function. Then, uniformly in n as $h \to 0$,

$$E \left[\frac{1}{v^h} \sum_{k=1}^{v_h} \left(L^{h,k,n} \Phi(\tilde{X}^{h,k,n}(1), \delta Y_n) - E_{\Pi^h(n-1), \delta Y_n} \Phi(\tilde{X}^h(1), \delta Y_n) \right) \right]_1^2 \to 0.$$

The following extends Theorem 4.1.

Theorem 4.2. *Theorem 3.1 holds for the filter defined by* (4.3) *under* (A2.1), (A2.2) *and* (A4.2).

Remark on (A4.1) and (A4.2). Assumptions (A4.1) and (A4.2) hold for many examples of interest. Often the desired change of measure is such that

$$\sup_{h,n} \sup_{x,y \in G} \frac{dP^h_{n-1}}{dM^{h,n}}(x, y) < \infty.$$

Then (A4.1) is clearly satisfied. Typically importance sampling is with respect to initial condition only and Π^h_{n-1} is a discrete probability measure. Then a natural choice of $M^{h,n}$ is the measure obtained by multiplying the probabilities of the points by appropriate weights which are uniformly bounded in h, n. An important case where the uniform boundedness does not hold, but where (A4.1) can be verified, is in Example 7 of [5], which concerns observation dependent importance sampling. For all these situations, if the sampling from $M^{h,n}$ using some variance reduction scheme of the types discussed earlier, then (A4.2) holds.

Of special interest is where the importance sampling is with respect to the initial condition only. To illustrate this case, we consider Example 4, where the sampling filter without importance sampling is given by (3.1) with $\tilde{X}^{h,n,l}(\cdot)$ defined via (3.13b). Then P^h_{n-1} can be identified with the measure $\Pi^h(n-1) \times P^h_\zeta$ in that this measure determines that of $(\tilde{X}^h(0), \tilde{X}^h(1))$. Now, let us write $M^{h,n}$ in a similar manner; namely, $M^{h,n} = M^{h,n}_0 \times P^h_\zeta$, where $M^{h,n}_0$ is a random measure on G. In this case, the measure transformation is over the initial condition only and we have

$$L^{h,n} = \frac{d\Pi^h(n-1)}{dM^{h,n}_0} \text{ and } L^{h,k,n} = \frac{d\Pi^h(n-1)}{dM^{h,n}_0}(\tilde{X}^{h,k,n}(0)).$$

Observation dependent importance sampling. The idea of applying importance sampling to the initial condition can be enhanced by the use of the next observation δY_n to determine the $M^{h,n}_0$ [5, 9, 28]. The idea and the demonstration that a large class of such algorithms satisfy (A4.1) is developed in [5]. There is no space for a full development. But the following is a rough description for the signal model of Example 4, where $\tilde{X}^h(\cdot)$ is defined by (3.13). Data and examples of such a procedure can be found in [28]. Again, (A4.1) and the mutual absolute continuity are the only conditions (in addition to (A2.1) and (A2.2)) that need to be verified for Theorem 4.1 to hold.

Suppose that $\Pi^h(n-1)$ is concentrated on the v^h points $\{x^{h,l,n}, l \le v^h\}$, with the $l-$th point having probability $\Pi^h_l(n-1)$. The path emanating from some of the $x^{h,l,n}$ might be "poor" predictors of the observation δY_n in the sense that the conditional density

$$p\left\{\delta Y_n \middle| X(n) = b_h(x^{h,l,n}, \xi^h)\right\}$$

is very small with a high probability. For some other points $x^{h,l,n}$, this value might be high with a reasonable probability. It seems reasonable to explore the paths emanating from the more promising initial points more fully, if this can be done without (asymptotically) biasing the procedure. The main problem is that we do not know (apart from the values of the $\Pi^h(n-1)$) which are the more promising points, and how much more promising they are. The "weights" for the importance sampling are to be determined by an exploratory sampling procedure, after which the sampling to get the next estimate $\Pi^h(n)$ will be done. Such algorithms are sometimes quite useful [28] in that the total computation for a filter with comparable accuracy can be less than what is needed for a direct method such as that in Example 1.

The procedure starts by getting a "typical" value of $b_h(x^{h,l,n}, \xi^h)$. The word "typical" is used loosely here. The aim is to get some preliminary approximation to the "predictive values" of the trajectories emanating from the point, given the next observation. This "typical value" might be an estimate of the mean value, or it might be a simple sample value or an average of several sample values. We call these the "indicator" values, and denote them by $\hat{X}^{h,l,n}(1), l \leq v^h$. Then the "predictive power" of this indicator value is computed, and the associated weights used to get the importance sampling measure for the final computation of $\Pi^h(n)$. See Example 7 in Section 6 of [5] for full details.

5. On the Uniqueness of the Invariant Measure:

A key assumption, (A2.2), was that the process $(X(\cdot), \Pi(\cdot))$ has a unique invariant measure. Obtaining conditions for this uniqueness is a difficult problem. We now provide a necessary and sufficient condition for uniqueness and give some examples where this condition can be verified. The proof of the necessary and sufficient condition shows an important "infinite horizon" property (Theorem 5.2) of the approximate filters of this paper, namely, that the pathwise distance, per unit time, of the optimal filter from the approximate filter is small asymptotically in time for small h.

We also discuss the weakening of (A2.2), using an idea of [19]. which shows that under appropriate conditions uniqueness holds in a certain class of invariant measures. (The more general case is in [3].) Finally we show that under this weakened hypothesis, errors per unit time for the optimal filter converge in probability to the expected limit. This result does not follow from the works of Kunita[18, 19], which study expected errors per unit time. The results will be given for discrete time. The statements and proofs in continuous time are almost identical.

Let $X(\cdot) = \{X(n) : n < \infty\}$ be, as in previous sections, a Feller-Markov process taking values in a compact set G. The observations are given by (2.1). Without loss of generality, consider the process $(X(\cdot), Y(\cdot))$ defined canonically on the path (sequence) space, $(G \times I\!\!R^m)^{\mathcal{N}}$, where m is the dimension of the observation and $\mathcal{N}$ the set of natural numbers. Let P_μ be the probability measure on the path space of $(X(\cdot), Y(\cdot))$ under which $X(0)$ has the distribution μ. With an abuse of notation, we use the notation P_x when μ is a point mass concentrated at $x \in G$. If the filter is initialized at $\Pi(0)=\nu$ (whether or not it is the distribution of $X(0)$), we denote the process by $\Pi_\nu(\cdot)$. Following the notation of Section 2, $\Pi_\nu(\cdot)$ is defined by

$$\langle \Pi_\nu(n), \phi\rangle = \frac{E_{\nu,Y_{0,n}}\left[\phi(\tilde{X}(n))R(\tilde{X}_{0,n}, Y_{0,n})\right]}{E_{\nu,Y_{0,n}}\left[R(\tilde{X}_{0,n}, Y_{0,n})\right]},$$

where $\phi(\cdot)$ is a bounded and measurable function, and $\tilde{X}(\cdot)$ is as in Section 2. Recall that $\Pi_\nu(0)=\nu$ is the distribution of $\tilde{X}(0)$. We begin with the following definition.

Definition 5.1. We say that the *"filter forgets its initial condition"* if for all $\nu \in \mathcal{M}(G)$, for all $x \in G$ and continuous bounded $\phi(\cdot)$ on G:

$$\frac{1}{N}\sum_{n=1}^{N} |\langle \Pi_\nu(n), \phi\rangle - \langle \Pi_x(n), \phi\rangle|$$

converges to zero in P_x-probability as $n \to \infty$.

One of our main results (Corollary 5.3) is that if $X(\cdot)$ has a unique invariant measure then the filter forgets its initial condition if and only if $(X(\cdot), \Pi(\cdot))$ has a unique invariant measure. The following gives one side of the assertion.

Theorem 5.1. *Suppose that $X(\cdot)$ has a unique invariant measure and the filter forgets its initial condition. Then the pair $(X(\cdot), \Pi(\cdot))$ has a unique invariant measure.*

Proof. Since $X(\cdot)$ is Feller-Markov with compact state space and g is continuous, it follows that $(X(\cdot), \Pi(\cdot))$ is Feller-Markov with a compact state space. Hence $(X(\cdot), \Pi(\cdot))$ has at least one invariant measure. Now suppose that ρ_1 and ρ_2 are invariant measures for the pair $(X(\cdot), \Pi(\cdot))$. We will show that for a measure determining class of functions f on $G \times \mathcal{M}(G)$

$$\int_{G\times\mathcal{M}(G)} f(x,\alpha)\rho_1(dx,d\alpha) = \int_{G\times\mathcal{M}(G)} f(x,\alpha)\rho_2(dx,d\alpha), \tag{5.1}$$

where (x,α) denotes a generic element of $G \times \mathcal{M}(G)$. It suffices to consider the class

$$\begin{aligned} \mathcal{C} \doteq \; & \{f \in C_b(G\times\mathcal{M}(G)) | \text{there exist } \; k \geq 1, \;\; \phi, \phi_1, \cdots, \phi_k \in C_b(G); \\ & H \in C_b^2(I\!\!R^k) \text{ such that } \;\; f(x,\alpha) = \phi(x)H(\langle\alpha,\phi_1\rangle, \cdots, \langle\alpha,\phi_k\rangle)\}, \end{aligned}$$

where for a metric space S, $C_b(S)$ denotes the space of bounded and continuous functions and $C_b^2(\mathbb{R}^k)$ is the space of functions on $\mathbb{R}^k$ which is continuous and bounded together with its partial derivatives up to second order. Now fix $f \in \mathcal{C}$ and let $\phi, \phi_1, \cdots, \phi_k$ and H be as in the definition of $\mathcal{C}$. Then there exists a C (depending on f) such that

$$\sup_{x\in G} |f(x,\alpha_1) - f(x,\alpha_2)| \leq C \sum_{i=1}^{k} |\langle \alpha_1, \phi_i\rangle - \langle \alpha_2, \phi_i\rangle|.$$

Let μ denote the unique invariant measure of $X(\cdot)$, and let μ_1, μ_2 be regular conditional probability functions such that $\rho_i(dx, d\alpha) = \mu_i(x, d\alpha)\mu(dx)$; $i = 1, 2$. The left side of (5.1) equals, for all N and $i = 1$,

$$\frac{1}{N}\sum_{j=1}^{N} \int_G \left[\int_{\mathcal{M}(G)} E_x \left[f(X(j), \Pi_\alpha(j))\right] \mu_i(x, d\alpha)\right] \mu(dx),$$

while the right hand side of (5.1) equals the same expression for $i = 2$. Thus

$$\begin{aligned}
&\left|\int_{G\times\mathcal{M}(G)} f(x,\alpha)\rho_1(dx,d\alpha) - \int_{G\times\mathcal{M}(G)} f(x,\alpha)\rho_2(dx,d\alpha)\right| \\
&\quad \leq \frac{1}{N}\sum_{j=1}^{N} \int_G \int_{\mathcal{M}(G)} \int_{\mathcal{M}(G)} E_x \left|f(X(j),\Pi_\alpha(j)) - f(X(j),\Pi_\beta(j))\right| \mu_1(x,d\alpha)\mu_2(x,d\beta)\mu(dx) \\
&\quad \leq C \int_G \int_{\mathcal{M}(G)} \int_{\mathcal{M}(G)} \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{k} E_x \left|\langle \Pi_\alpha(j)), \phi_i\rangle - \langle \Pi_\beta(j)), \phi_i\rangle\right| \mu_1(x,d\alpha)\mu_2(x,d\beta)\mu(dx).
\end{aligned} \tag{5.2}$$

The assumption on forgetting of initial condition implies that for all x, α, β,

$$\lim_N \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{k} E_x \left|\langle \Pi_\alpha(j)), \phi_i\rangle - \langle \Pi_\beta(j)), \phi_i\rangle\right| = 0.$$

The dominated convergence theorem implies that the right side of (5.2) converges to 0, which proves (5.1) and the theorem. ■

We now prove the infinite horizon property for the approximating filters which was mentioned at the beginning of this section. A consequence is the converse of Theorem 5.1.

Theorem 5.2. *Suppose that the pair* $(X(\cdot), \Pi(\cdot))$ *admits a unique invariant measure,* ρ. *Let* $\{\Pi^h(\cdot)\}_{h>0}$ *be a collection of* $\mathcal{M}(G)$ *valued processes such that* $\Pi^h(j)$ *is* $\sigma\{Y(i) : 1 \leq i \leq j\}$ *measurable for all* j, h. *Also assume* *that for all continuous and bounded functions* F *on* $G \times \mathcal{M}(G)$,

$$\lim_{h\to 0, N\to\infty} \frac{1}{N}\sum_{j=1}^{N} F(X(j), \Pi^h(j)) = \int_{G\times\mathcal{M}(G)} F(x,\alpha)\rho(dx, d\alpha).$$

where the convergence s in probability. Then, for all bounded and continuous ϕ, *in probability,*

$$\lim_{h\to 0, N\to\infty} \frac{1}{N}\sum_{j=0}^{N} |\langle \Pi^h(j), \phi\rangle - \langle \Pi(j), \phi\rangle|^2 = 0.$$

Remark. *The hypothesis on the* $\Pi^h(\cdot)$ *are satisfied for all the approximating filters in this paper, under (A2.2).*
Proof. Note that

$$\begin{aligned} E\frac{1}{N}\sum_{j=1}^{N} |\langle \Pi^h(j), \phi\rangle - \langle \Pi(j), \phi\rangle|^2 &= \frac{1}{N}\sum_{j=1}^{N} E\langle \Pi^h(j), \phi\rangle^2 + \frac{1}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 \\ &- \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi^h(j), \phi\rangle\langle \Pi(j), \phi\rangle. \end{aligned}$$

Thus using the hypothesis on the approximating filter and noting that the optimal filter also satisfies the hypothesis, we have

$$\begin{aligned} & \lim_{h\to 0, N\to\infty} E\frac{1}{N}\sum_{j=1}^{N} |\langle \Pi^h(j), \phi\rangle - \langle \Pi(j), \phi\rangle|^2 \\ = & \lim_{h\to 0, N\to\infty} \left(\frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 - \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi^h(j), \phi\rangle\langle \Pi(j), \phi\rangle \right) \\ = & \lim_{h\to 0, N\to\infty} \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 - \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi^h(j), \phi\rangle\phi(X(j)) \\ = & \lim_{N\to\infty} \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 - \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle\phi(X(j)) \\ = & \lim_{N\to\infty} \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 - \frac{2}{N}\sum_{j=1}^{N} E\langle \Pi(j), \phi\rangle^2 = 0. \end{aligned}$$

This proves the theorem. ■

The following result is an immediate corollary.

Corollary 5.3 *If* $X(\cdot)$ *has a unique invariant measure, the filter forgets its initial condition if and only if* $(X(\cdot), \Pi(\cdot))$ *has a unique invariant measure.*

Proof. If the filter forgets its initial condition, then $(X(\cdot), \Pi(\cdot))$ has a unique invariant measure (Theorem 5.1). Now let $(X(\cdot), \Pi(\cdot))$ have a unique invariant measure, let $\nu \in \mathcal{M}(G)$ and $x \in G$ be arbitrary, and let ϕ be a continuous and bounded function on G. We need to show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} E_x \left(|\langle \Pi_\nu, \phi \rangle - \langle \Pi_x, \phi \rangle|^2 \right) = 0.$$

This is a consequence of Theorem 5.2 since $\Pi_\nu(\cdot)$ is an approximating filter (which does not depend on h) satisfying the hypothesis in Theorem 5.2 and under P_x, $\Pi_x(\cdot)$ is the optimal filter. ■

In view of the above Corollary it becomes important to obtain conditions under which the filter forgets its initial condition. This is a difficult problem in general. However, recently results in [1, 2, 7, 8, 25] have identified some important classes where the filter indeed forgets its initial condition. Finally, we consider a possible weakening of the basic assumption (A2.2). We begin by recalling the following result from [19].

Theorem 5.4 [19]. *Let* $X(\cdot)$ *have a unique invariant measure* μ. *Suppose that for all bounded and continuous functions* ϕ *on* G

$$\limsup_{j\to\infty} E_\mu |E\left[\phi(X(j))|X(0)\right] - \langle \mu, \phi \rangle| = 0. \tag{5.3}$$

Then there is only one invariant measure, ρ, for $(X(\cdot), \Pi(\cdot))$ *which satisfies*

$$\int_{G\times\mathcal{M}(G)} f(x)F(\alpha)\rho(dx, d\alpha) = \int_{G\times\mathcal{M}(G)} \langle \alpha, f \rangle F(\alpha)\rho(dx, d\alpha),$$

for all bounded and continuous functions f, F on G *and* $\mathcal{M}(G)$ *respectively.*

Now let $\Pi^h(\cdot)$ be an approximating filter and let $Q_f^{h,N}$ be as in the previous sections, occupation measures constructed using $\Pi^h(\cdot)$. Let Q_f^ω be an arbitrary weak sense limit of $Q_f^{h,N}$. Assume now that $X(\cdot)$ has a unique invariant measure μ and that (5.3) holds. Suppose now, we can verify that

$$\int f(x(0))F(\pi(0))dQ^\omega(x(\cdot), \pi(\cdot)) = \int \langle \pi(0), f \rangle F(\pi(0))dQ^\omega(x(\cdot), \pi(\cdot)), \tag{5.4}$$

a.e. ω for all continuous and bounded f, F. Then we can apply the above stated uniqueness result to conclude that Q^w is non-random and is independent of the weakly convergent subsequence. It will then follow that the occupation measures converge weakly to $\tilde{Q}_f$, where $\tilde{Q}_f$ is the stationary measure on the path space of $(X(\cdot), \Pi(\cdot))$ corresponding to the initial law ρ of Theorem 5.4. This will in turn imply that convergence in (2.10) will hold for such an approximating filter.

In order to show (5.4) it suffices to show that

$$\lim_{h\to 0, N\to\infty} E\left|\frac{1}{N}\sum_{j=1}^{N} f(X(j))F(\Pi^h(j)) - \frac{1}{N}\sum_{j=1}^{N}\langle \Pi^h(j), f\rangle F(\Pi^h(j))\right|^2 = 0. \tag{5.5}$$

(5.5) is difficult to establish and its study will be done elsewhere due to space limitations. However, we will consider the instructive case where $\Pi^h(\cdot) \equiv \Pi(\cdot)$.

Theorem 5.5. *Suppose that* $X(\cdot)$ *has a unique invariant measure* μ, *and* (5.3) *holds. Then for all continuous and bounded* F

$$\frac{1}{N}\sum_{n=1}^{N} F(\Psi_f(n+\cdot)) = \int F(\psi_f(\cdot))\bar{Q}_f^N(d\psi_f(\cdot)) \to \int F(\psi_f(\cdot))\tilde{Q}_f(d\psi_f(\cdot)),$$

where Ψ_f *is as in Theorem* 2.1 *with* $\Pi^h(\cdot)$ *replaced by the optimal filter* $\Pi(\cdot)$, Q_f^N is defined using $\Pi(\cdot)$ *instead of* $\Pi^h(\cdot)$, *and* $\tilde{Q}_f$ *is the stationary measure on the path space of* $(X(\cdot), \Pi(\cdot))$ *corresponding to the initial law* ρ *of Theorem* 5.4.

Remark: Note that we are not assuming the uniqueness of the invariant measure of $(X(\cdot), \Pi(\cdot))$. The above result shows in particular that for bounded and continuous f

$$\frac{1}{N}\sum_{j=1}^{N} |f(X(j)) - \langle \Pi(j), f\rangle|^2 \tag{5.6}$$

converges in probability to

$$\int_{G\times\mathcal{M}(G)} |f(x) - \langle \alpha, f\rangle|^2 d\rho(x, \pi) \tag{5.7}$$

as $N \to \infty$. It was shown in [18] that the expected value of (5.6) converges to (5.7), however the convergence in probability can not be deduced from there.

Proof of Theorem 5.5: It suffices to show that for all continuous and bounded f, F defined on G and $\mathcal{M}(G)$ respectively

$$\lim_{N\to\infty}\left|\frac{1}{N}\sum_{j=1}^{N}\left[f(X(j)) - \langle \Pi(j), f\rangle\right] F(\Pi(j))\right|^2 = 0.$$

The left side can be written, up to an asymptotically negligible term, as

$$\frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1} E\left(\left[f(X(j)) - \langle \Pi(j), f\rangle\right] F(\Pi(j)) \left[f(X(k)) - \langle \Pi(k), f\rangle\right] F(\Pi(k))\right). \tag{5.8}$$

Now, for $j > k$ the expression inside the summation can be written as:

$$E\left(\left[E\left[f(X(j) \mid X(k), Y_{0,j}\right] - \langle \Pi(j), f\rangle\right] F(\Pi(j)) \left[f(X(k)) - \langle \Pi(k), f\rangle\right] F(\Pi(k))\right])$$

which, for some constant C, is bounded above by

$$CE\left|E\left[f(X(j) \mid X(k), Y_{0,j}\right] - \langle \Pi(j), f\rangle\right|.$$

Thus to prove the theorem it suffices to show that:

$$\lim_{N\to\infty}\frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1} E\left[E\left[f(X(j) \mid X(k), Y_{0,j}\right] - \langle \Pi(j), f\rangle\right]^2 = 0. \tag{5.9}$$

Next observe that

$$E\left[E\left[f(X(j) \mid X(k), Y_{0,j}\right] - \langle \Pi(j), f\rangle\right]^2 = E\left[E\left[f(X(j) \mid X(k), Y_{0,j}\right]\right]^2 - E\left[\langle \Pi(j), f\rangle^2\right]. \tag{5.10}$$

Also, $\Pi(\cdot)$ is a Feller-Markov process with semigroup $S(j)$: $S(j)F(\alpha) \doteq E_\alpha F(\Pi_\alpha(j))$, where F is an arbitrary bounded measurable function on $\mathcal{M}(G)$ and $\alpha \in \mathcal{M}(G)$.

In terms of this semigroup the expectations on the right of (5.10) are:

$$E\left[\left(S(j-k)F_f\right)\left(\delta_{X(k)}\right)\right] - \left(S(j)F_f\right)(\Pi(0)),$$

where $\Pi(0)$ is the (nonrandom) probability distribution of $X(0)$ and $F_f(\alpha) \doteq \langle \alpha, f\rangle^2$ is a continuous and bounded function on $\mathcal{M}(G)$. Thus, in view of (5.9) and (5.10) it suffices to show that

$$\lim_{N\to\infty}\frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1} E\left[\left(S(j-k)F_f\right)\left(\delta_{X(k)}\right)\right] - \frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1}\left(S(j)F_f\right)(\Pi(0)) = 0. \tag{5.11}$$

Now, by [18] there exists a unique invariant measure, M, for $\Pi(\cdot)$. To complete the proof we will show that both terms in (5.11) converge to $\langle M, F_f\rangle$. The second term equals:

$$\frac{2}{N^2}\sum_{j=1}^{N}(j-1)\left(S(j)F_f\right)(\Pi(0)).$$

Note that the sequence of measures $\{P_N\}$ on $\mathcal{M}(G)$ defined by

$$P_N(A)\doteq\frac{2}{N^2}\sum_{j=1}^{N}(j-1)\left(S(j)I_A\right)(\Pi(0)),\quad A\in\mathcal{B}(\mathcal{M}(G)),$$

is tight. Furthermore if $\overline{P}$ is a weak sense limit along the subsequence P_{N_k} then, for any continuous and bounded function F on $\mathcal{M}(G)$,

$$\begin{aligned}\langle\overline{P},F\rangle &= \lim_{k\to\infty}\frac{2}{N_k^2}\sum_{j=1}^{N_k}(j-1)(S(j)F)(\Pi(0))\\ &= \frac{2}{N_k^2}\sum_{j=1}^{N_k}(j-1)(S(j+1)F)(\Pi(0))=\langle\overline{P},S(1)F\rangle,\end{aligned}$$

where the last step follows by noting that since $\Pi(\cdot)$ is a Feller semigroup, $S(1)F$ is continuous and bounded. Thus $\overline{P}$ is an invariant measure for $\Pi(\cdot)$. Hence $\overline{P}=M$. This shows that the second term in (5.11) converges to $\langle M, F_f\rangle$. Finally consider the first term of (5.11). It clearly equals

$$\frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1}\int_G\left(S(j-k)F_f\right)(\delta_x)\nu_k(dx),$$

where ν_k is the law of $X(k)$. Once more define a sequence of measures:

$$\tilde{P}_N(A)\doteq\frac{2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{j-1}\int_G\left(S(j-k)I_A\right)(\delta_x)\nu_k(dx);\quad A\in\mathcal{B}(\mathcal{M}(G)).$$

Clearly, the sequence is also tight and the same argument as used above shows that if $\tilde{P}$ is a weak sense limit then for any continuous and bounded function F on $\mathcal{M}(G)$ $\langle\tilde{P},F\rangle=\langle\tilde{P},S(1)F\rangle$. This fact, as before, implies that $\tilde{P}_n$ converges weakly to M and thus the first term in (5.11) also converges to $\langle M, F_f\rangle$. ■

References

[1] R. Atar. Exponential stability for nonlinear filtering of diffusion processes in non-compact domain. To appear in *Annals of Probability.*

[2] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Ann. Inst. Henri Poincaré*, 33:697–725,1997.

[3] A. Bhatt, A. Budhiraja and R. Karandikar. Markov property and ergodicity of the nonlinear filter. University of Notre Dame Report. Submitted for publication, 1999.

[4] A. Budhiraja and H.J. Kushner. Approximation and limit results for nonlinear filters over an infinite time interval. Lefschetz Center for Dynamical Systems Report. To appear SIAM J. on Control and Optimization, 1999.

[5] A. Budhiraja and H.J. Kushner. Approximation and limit results for nonlinear filters over an infinite time interval: Part II, random sampling algorithms. Lefschetz Center for Dynamical Systems Report. To appear SIAM J. on Control and Optimization, 1999.

[6] A. Budhiraja and H.J. Kushner. A nonlinear filtering algorithm based on an approximation of the conditional distribution. To appear: IEEE Trans. on Automatic Control, 1999.

[7] A. Budhiraja and H.J. Kushner. Robustness of nonlinear filters over the infinite time interval. *SIAM J. Control and Opt.*,36:1618–1637, 1998.

[8] A. Budhiraja and D. Ocone. Exponential stability in discrete time filtering for non-ergodic signals. *Stochastic processes and their applications*, 82:245–257, 1999.

[9] J. Carpenter, P. Clifford, and P. Fearnhead. An improved particle filter for non-linear problems. Preprint. Statistics Dept., Univ. of Oxford., 1998.

[10] D. Crisan and T.Lyons. Nonlinear filtering and measure valued processes. *Probability Theory and Related Fields*, 109:217–244, 1997.

[11] P. Del Moral and G. Salat. Filtrage non-linéaire résolution particuliere à la monte carlo. *C.R. Acad. Sci., Paris, Ser. I, Math*, 320:1147–1152, 1997.

[12] G.S. Fishman. *Monte Carlo.* Springer, Berlin and New York, 1995.

[13] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London, 1996.

[14] N.J. Gordon, D.J. Salmond, and C. Ewing. Bayesian state estimation for tracking and guidance using the bootstrap filter. *Journal of Guidance, Control and Dynamics*, 18:1434–1443, 1995.

[15] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/nonGaussian Baysian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

[16] M. Isard and A. Blake. Contour tracking by stochastic propogation of conditional density. In *Proc. European Conf. on Computer Vision*, pages 343–356, 1996.

[17] G. Kitagawa. Monte carlo filter and smoother for non Gaussian nonlinear state space models. *J. of Computational and Graphical Statistics*, 5:1–25, 1996.

[18] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of Markov processes. *Journal of Multivariate Analysis*, 1:365–393, 1971.

[19] H. Kunita. Ergodic properties of nonlinear filtering processes. *Spatial Stochastic Processes*, Eds. Alexander, K.C. and Watkins, J.C., 233–256,1991.

[20] H.J. Kushner. *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations.* Academic Press, New York, 1977.

[21] H.J. Kushner. *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, volume 3 of *Systems and Control.* Birkhäuser, Boston, 1990.

[22] H.J. Kushner. Robustness and convergence of approximations to nonlinear filters for jump–diffusions. *Computational and Applied Math.*, 16:153–183, 1997.

[23] H.J. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, Berlin and New York, 1992.

[24] H.J. Kushner and H. Huang. Approximation and limit results for nonlinear filters with wide bandwidth observation noise. *Stochastics Stochastics Rep.*, 16:65–96, 1986.

[25] F. Le Gland and L. Mevel. Geometric ergodicity in hidden Markov models. To be published, 1996.

[26] P. Müller. Monte carlo integration in general dynamic models. *Contemporay Mathematics*, 115:145–162, 1991.

[27] M-S. Oh. Monte carlo integration via importance sampling: Dimensionality effect and an adaptive algorithm. *Contemporary mathematics*, 115:165–187, 1991.

[28] M.K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. Preprint. Math. Dept., University of Oxford., 1997.