# Approximation and Limit Results for Nonlinear Filters Over an Infinite Time Interval: Part II, Random Sampling Algorithms

Amarjit Budhiraja *
Department of Mathematics
University of Notre Dame
Notre Dame, IN 46656

Harold J. Kushner†
Division of Applied Mathematics
Brown University
Providence, RI 02912

December 11, 2000

**Abstract**

The paper is concerned with approximations to nonlinear filtering problems that are of interest over a very long time interval. Since the optimal filter can rarely be constructed, one needs to compute with numerically feasible approximations. The signal model can be a jump–diffusion, reflected or not. The observations can be taken either in discrete or continuous time. The cost of interest is the pathwise error per unit time over a long time interval. In a previous paper of the authors [2], it was shown, under quite reasonable conditions on the approximating filter and on the signal and noise processes that (as time, bandwidth, process and filter approximation, etc.) go to their limit in any way at all, the limit of the pathwise average costs per unit time is just what one would get if the approximating processes were replaced by their ideal values and the optimal filter were used. When suitable approximating filters cannot be readily constructed due to excessive computational requirements or to problems associated with a high signal dimension, approximations based on random sampling methods (or, perhaps, combinations of sampling and analytical methods) become attractive, and are the subject of a great deal of attention. This is somewhat analogous to the use of monte carlo methods for high dimensional integration problems. Owing to the sampling errors as well as to the other (computational and modeling) approximations that are made, in the filter and signal processes, it is conceivable that the long term pathwise average errors per unit time will be large, even with approximations that would perform well over some bounded time interval.

*Supported in part by contracts N00014-96-1-0276 and N00014-96-1-0279 from the Office of Naval Research and NSF grant DMI 9812857.

†Supported in part by contracts DAAH04-96-1-0075 from the Army Research office and NSF grant ECS 9703895.

The work of the previous paper is extended to a wide class of such algorithms. Under quite broad conditions, covering virtually all the cases considered to date, it is shown that the pathwise average errors converge to the same limit that would be obtained if the optimal filter were used, as time goes to infinity and the approximation parameter goes to its limit in any way at all. All the extensions (e.g., wide bandwidth observation or system driving noise) in [2] hold for our random sampling algorithms as well.

**Key words:** Nonlinear filters, numerical approximations to nonlinear filters, robustness of filters, infinite time filtering, occupation measures, pathwise average errors, random sampling algorithms

**AMS subject classification numbers:** 93E11, 60G35

# 1 Introduction

This paper is an extension of the work in [2], which was concerned with the performance of a wide variety of approximations to optimal nonlinear filters over very long time intervals, where *pathwise average* errors are of interest. Let us first briefly review the motivation for that paper. Suppose that the underlying signal model is a diffusion or jump-diffusion $X(\cdot)$ (reflected or not), or a discrete time Markov chain, with white noise corrupted observations, and the dynamics and/or the observation function are nonlinear. Then, except for some few examples, one cannot construct "finite" or computable optimal filters, and some type of approximation must be used.

A very common approximation method starts by approximating the process $X(\cdot)$ by a simpler process $\tilde{X}^h(\cdot)$ for which the optimal nonlinear filter can be constructed, and then uses that filter but with the observations being those on the actual physical process $X(\cdot)$. For example, $\tilde{X}^h(\cdot)$ might be a discretized (in state and/or in time) form of $X(\cdot)$. It is such that, as $h \to 0$, $\tilde{X}^h(\cdot)$ converges weakly to $X(\cdot)$. Let $\Pi^h(\cdot)$ denote the actual approximating filter. For each $h$, it is a measure valued process and $\Pi^h(\cdot)$ converges weakly to the true conditional distribution process as $h \to 0$. I.e., the computed expectations of any bounded and continuous function converges to the true conditional expectation [24, 22]. However, if the filter is to be used over a very long large interval $[0, T]$, the most appropriate errors are often the *pathwise average* (rather than the mathematical expectation) errors per unit time, for whatever definition of "error" is appropriate. This is the case since we work with only one long path, and the mathematical expectation over all paths might not be a useful indicator of the quality of the approximation. For specificity in this introduction, let us define the average pathwise cost or error on $[0, T]$ to be

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

where $\phi(\cdot)$ and $f(\cdot)$ are arbitrary bounded and continuous functions. Our results cover much more general forms of the cost or error function.

Now there are two parameters, $h$ and $T$. The convergence of the filter process $\Pi^h(\cdot)$ over any fixed finite interval says nothing about the behavior of the pathwise average errors as $h \to 0$ and $T \to \infty$ arbitrarily. Under reasonable conditions, it was shown in [2], that the pathwise errors converge in probability to an optimal deterministic limit, and this limit is exactly what one would get for the limit of the mathematical expectation $EG^{h,T}(\phi)$ if the *true optimal filter* were used instead of $\Pi^h(t)$. This is an ideal result. The convergence is independent of how $h \to 0$ or $T \to \infty$. For applications, it is important that $h$ and $T$ be allowed to go to their limits in an arbitrary way.

The reference [2] actually dealt with a much more general setup. The signal process was allowed to be not necessarily a diffusion but a process (possibly driven by wide bandwidth noise), which converges weakly to a diffusion as some parameter converges to its limit (e.g. the noise bandwidth goes to infinity). Wide bandwidth observation noise was also allowed. The case where the observations are taken in discrete time was also covered.

The filtering problem with observations in continuous time is often justified as being the limit, as $\Delta \to 0$, of the problem where the observations are taken at the discrete times $n\Delta, n = 1, 2, \ldots$. As $\Delta \to 0$, the problem of correlation in the observation noise can become serious. The wide bandwidth observation noise, combined with all of the other approximations, can conceivably lead to serious errors when $T$ is large. Let the observations be in discrete time with a small interval between them, the observation noise $\xi_n^\Delta$ be wide bandwidth (but whose suitably scaled sums converge weakly to a Wiener process), let the signal be only approximated by a sampled diffusion, and let the pathwise average errors over a long time interval $T$ be of interest. Then under broad conditions which fill in the above description, the desired limit result continues to hold, as all the parameters go to their limits simultaneously (inter-observation interval, observation noise bandwidth, $h$, the parameter denoting the signal approximation, $T$). The desired result is that the limit of the pathwise errors are what one would get for the optimal filter with the limit (i.e., the ideal) signal and observational model used.

We also note the reference [24], where the pathwise average error was replaced by the expectation of the pathwise average error. In [17], the asymptotics of the filter alone were dealt with, and it was presumed that the filter was the true optimal filter, not an approximation.

We now turn to the description of this paper. In [2], the approximate filter was that for an approximating process $\tilde{X}^h(\cdot)$ but with the actual physical observations on $X(\cdot)$ used. One common and convenient example is the Markov chain approximation method, where the approximating process $\tilde{X}^h(\cdot)$ is a continuous time interpolation of a Markov chain [23, 22]. When the dimension is larger than three or four, such methods can have excessively high computational requirements. Alternatives, based on random sampling or Monte Carlo then become attractive, analogous to the case of classical multidimensional integration. The topic is of considerable current interest; e.g., [3, 4, 5, 11, 12, 14, 15, 26, 27, 28].

Such methods are also of interest when the transition probabilities are very hard to compute; for example, in discrete time problems, where the signal is the output of a system with very complex dynamics, but which can be conveniently simulated. All of the issues in [2] (which were mentioned above) arise here as well, in addition to the potentially serious errors due to the random sampling and the very large time intervals. This paper extends the results in [2] to such sampling based algorithms. To distinguish the algorithms in [2] from those of this paper, we refer to the former as *integration* algorithms, since the conditional distributions are computed using integrations or summations over the distributions of the approximating processes. One must keep in mind that random sampling based filters usually require a large number of samples if they are to work well.

Appropriate analogs of the occupation measure methods in [2] are employed for the proofs. Section 2 provides the standard background for the filtering problem in continuous time. Section 3 discussed the formulation of the limit problem for the continuous time case in terms of occupation measures, and states a main result from [2] which will be used. Section 4 repeats this for the discrete time problem. In order to effectively exploit the past results, the problem is set up in such a way that the proofs are close to those for the "integration" algorithms of [2]. Thus only the differences in the proofs will be presented. In preparation for this, the structure of the proof in [2] is briefly outlined, and the points where there will be a difference noted. A fundamental assumption in [2] is the *consistency* assumption which quantifies the weak convergence of the computational approximating process $\tilde{X}^h(\cdot)$ to $X(\cdot)$ as $h \to 0$. This is (A4.1) here. It will be weakened in several ways, depending on the form of the random sampling algorithm.

Section 5 concerns a variety of forms of the sampling algorithms in discrete time. The simplest form is based on a pure random sampling of an approximating process $\tilde{X}^h(\cdot)$. The part of the convergence proof that differs from that in [2] is given. The basic scheme can be generalized in many ways. The standard variance reduction methods such as antithetic variables and stratified sampling can be used. Combinations of integration and sampling methods are often of great use, since it might be most convenient to simulate some parts of the problem, but to use "integrations" with respect to approximating variables in others. See Examples 4 and 5 in Section 5. We rephrase the consistency condition so that it covers quite general algorithms. Section 6 concerns the use of importance sampling methods for the discrete time problem [3, 9, 28]. The standard form is discussed. But, the most interesting form is where the measure change depends on the next observation, which is thus used to guide the simulation on the current time interval. Some such algorithms were used in [3, 28], as well as by the authors. The proofs for all of these cases differ only slightly from that given for the basic example in Section 5 and the differences are only discussed. Finally in Section 7 we study the continuous time analogs of the various random sampling and combined random sampling-integration algorithms studied in Sections 5 and 6. We will begin by indicating the form of the approximating filter for the case where the random samples are mutually

independent and identically distributed. We will then consider a general form of the approximating filter which would cover not only the case of such i.i.d. samples but also various variance reduction schemes and importance sampling algorithms of the type studied for discrete time problems in Sections 5 and 6.

# 2 Background: The Optimal Filter and Numerical Approximations: Continuous Time

**The optimal filter in continuous time.** For simplicity and specificity and until further notice, suppose that the signal process is the $\mathbb{R}^r$ valued diffusion

$$dX = p(X)dt + \sigma(X)dW, \tag{2.1}$$

where $W(\cdot)$ is a standard vector–valued Wiener process and $p(\cdot)$ and $\sigma(\cdot)$ are continuous. We suppose that the solution is unique in the weak sense for each initial condition. Furthermore, suppose throughout that there is a compact set $G$ such that $X(t) \in G$ for all $t$, if $X(0) \in G$, and we always let $X(0) \in G$. All probability measures (random or not) on $\mathbb{R}^r$ considered hereafter will be assumed to have their support contained in $G$. The observation process is

$$Y(t) = \int_0^t g(X(s))ds + B(t), \tag{2.2}$$

where $g(\cdot)$ is a continuous vector–valued function and $B(\cdot)$ is a standard vector-valued Wiener process, independent of $W(\cdot)$ and $X(0)$.

As pointed out in [2], the approximation and limit results proved there continue to hold, with minor changes in the proofs, for the case where the signal process is a jump diffusion or is a reflecting diffusion with appropriate conditions on the reflection direction. Indeed, consider the problem where the path is constrained to lie in $G$ by boundary reflection. Let $G$ have piecewise smooth boundaries, and for appropriate conditions on the reflection direction, replace (2.1) by the solution to the Skorohod problem:

$$dX = p(X)dt + \sigma(X)dW + dZ,$$

where $Z$ is the reflection term. Under suitable conditions, the solution is weak sense unique and is a strong Markov process [6]. With minimal alterations, all of the results of this paper carry over to this model, and to the extension where a jump driving process is added. However, for the sake of simplicity, we confine our work to the model (2.1).

Let $\tilde{X}(\cdot)$ be a process satisfying (2.1), and which (loosely speaking) is conditionally independent of $(X(\cdot), W(\cdot), B(\cdot))$ given its initial condition: We formalize this as follows. $\tilde{X}(\cdot)$ is a process satisfying (2.1) such that there exists a (possibly random) probability measure $\Pi^*$ on $\mathbb{R}^r$ with the properties that conditioned on $\Pi^*$, $\tilde{X}(\cdot)$ is independent of $(X(\cdot), W(\cdot), B(\cdot))$ and the conditional

distribution of $\tilde{X}(0)$ given $\Pi^*$ is $\Pi^*$. We will call $\Pi^*$ the "random initial distribution" of $\tilde{X}(\cdot)$ (i.e., the distribution of $\tilde{X}(0)$). It will vary depending on the need, and will be specified when needed.

For any process $U(\cdot)$, let $U_{a,b}$, $a \leq b$, denote the set $\{U(s), a \leq s \leq b\}$. Let $E_Z f$ denote the expectation of a function $f$ given the data (or $\sigma-$algebra) $Z$. Until further notice, let $\Pi(0)$ denote the distribution of $X(0)$, and $\Pi(t)$ the distribution of $X(t)$ given the data $Y_{0,t}$ and $\Pi(0)$. Define

$$R(\tilde{X}_{0,t}, Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}(s))dY(s) - \frac{1}{2}\int_0^t \left|g(\tilde{X}(s))\right|^2 ds\right]. \tag{2.3}$$

Using the representation of the optimal filter $\Pi(\cdot)$ as it was originally developed in [19], for each bounded and measurable real–valued function $\phi(\cdot)$, we can define the evolution of the optimal filter by

$$\int \phi(x)\Pi(t)(dx) \equiv \langle \Pi(t), \phi\rangle = \frac{E_{\{\Pi(0),Y_{0,t}\}}\left[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y_{0,t})\right]}{E_{\{\Pi(0),Y_{0,t}\}}R(\tilde{X}_{0,t}, Y_{0,t})}. \tag{2.4}$$

The notation $E_{\{\Pi(0),Y_{0,t}\}}$ denotes the expectation conditioned on the data $Y_{0,t}$ and on $\Pi(0)$ being the initial distribution of $\tilde{X}(\cdot)$ (i.e., $\Pi(0)$ is the current value of what we generically called $\Pi^*$ above). This representation is convenient for our purposes, and is equivalent to the forms used subsequently which were based on measure transformations, as in [7, 13, 25].

The Markov property of $X(\cdot)$ implies that the filter defined by (2.4) satisfies the semigroup relation:

$$\langle \Pi(t), \phi\rangle = \frac{E_{\{\Pi(t-s),Y_{t-s,t}\}}\left[\phi(\tilde{X}(s))R(\tilde{X}_{0,s}, Y_{t-s,t})\right]}{E_{\{\Pi(t-s),Y_{t-s,t}\}}R(\tilde{X}_{0,s}, Y_{t-s,t})},\ 0 < s \leq t. \tag{2.5}$$

In (2.5), $\Pi(t-s)$ is the random initial distribution of $\tilde{X}(\cdot)$. Throughout the paper, we use the notation $E_{\{\Pi(a),Y_{a,b}\}}F(\tilde{X}_{0,s}, Y_{a,b})$ for the conditional expectation, given the data $\{Y_{a,b}, \Pi(a)\}$ and where the random initial distribution for $\tilde{X}(\cdot)$ is $\Pi(a)$. The analogous notation will be used when approximations to $\tilde{X}(\cdot)$ are used.

**An approximating filter.** Except for some special cases, $\Pi(t)$ is very hard to compute for nonlinear problems. A fundamental difficulty in realizing (2.4) (in either discrete or continuous time), is that one needs to compute the evolution of a random measure on the range space of the signal process, and these measures are rarely defined by a finite dimensional parameter (of reasonable size). Thus, in applications to nonlinear problems one must use some type of approximation. Sometimes, one can effectively linearize. Otherwise, perhaps the most common method of approximation is to approximate the signal process by a simpler form for which a convenient filter can be constructed. Then the approximate filter is obtained by constructing the filter for that approximating signal process,

but using the actual physical observations. For example, the approximating filter might be that for a time and space discretization of the process (2.1). The key mathematical ideas behind such approximations and their convergence properties (over finite time intervals) are in [23], in connection with the Markov chain approximation method, a canonical form of this idea.

Let us formalize the above canonical approximation. Let $\tilde{X}^h(\cdot)$ denote the approximating process, which is used to construct the approximating filter. I.e., the approximating filter is constructed as though the true process was $\tilde{X}^h(\cdot)$, but in this filter, we use the actual physical observations defined by (2.2).

Let $\Pi^h(0)$ be an approximation to the true initial distribution of $X(0)$. The $\tilde{X}^h(\cdot)$ might be a Markov process, for example a continuous time Markov chain on a finite state space. More commonly, it is an interpolation of a discrete parameter process: I.e, there is $\delta_h > 0$ and which goes to zero as $h \to 0$ such that $\tilde{X}^h(\cdot)$ is constant on the intervals $[n\delta_h, n\delta_h + \delta_h)$ and $\tilde{X}^h(n\delta_h), n = 0, \ldots,$ is Markov. When the signal process is defined in continuous time, we always assume that $\tilde{X}^h(\cdot)$ is of one of these two forms. Furthermore, we always suppose (without loss of generality) that $\tilde{X}^h(t)$ takes values in $G$.

Define

$$R(\tilde{X}^h_{0,t}, Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}^h(s))dY(s) - \frac{1}{2}\int_0^t \left|g(\tilde{X}^h(s))\right|^2 ds\right]. \tag{2.6}$$

For Markov $\tilde{X}^h(\cdot)$, the approximating filter $\Pi^h(\cdot)$ is defined by

$$\langle \Pi^h(t), \phi\rangle = \frac{E_{\{\Pi^h(0), Y_{0,t}\}}\left[\phi(\tilde{X}^h(t))R(\tilde{X}^h_{0,t}, Y_{0,t})\right]}{E_{\{\Pi^h(0), Y_{0,t}\}}R(\tilde{X}^h_{0,t}, Y_{0,t})}, \tag{2.7}$$

and $\Pi^h(\cdot)$ satisfies the semigroup equation

$$\langle \Pi^h(t+s), \phi\rangle = \frac{E_{\{\Pi^h(t), Y_{t,t+s}\}}\left[\phi(\tilde{X}^h(s))R(\tilde{X}^h_{0,s}, Y_{t,t+s})\right]}{E_{\{\Pi^h(t), Y_{t,t+s}\}}R(\tilde{X}^h_{0,s}, Y_{t,t+s})},\ s > 0,\ t \geq 0. \tag{2.8}$$

According to our standard notation, the initial distribution of $\tilde{X}^h(\cdot)$ in (2.6) is $\Pi^h(0)$ and it is $\Pi^h(t)$ in (2.8).

When $\tilde{X}^h(\cdot)$ is piecewise constant with $\tilde{X}^h(n\delta)$ being Markov, then the approximating filter is defined by (2.7) and (2.8), but where $t$ and $s$ are integral multiples of $\delta$, and $\Pi^h(\cdot)$ is constant on the intervals $[n\delta, n\delta + \delta)$. Thus, the evolution of $\Pi^h(\cdot)$ can be written in recursive form in general. We see that, by Bayes' rule, (2.7) and (2.8) are filters for the $\tilde{X}^h(\cdot)$ process, but with the actual observations $Y_{n\delta, n\delta+\delta}$ used at step $n$. The conditions for convergence of this Markov chain approximation method are in [20, 23]. The following is the essential condition.

**A2.1. A consistency assumption.** We assume that for any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure $\Pi$, $\tilde{X}^h(\cdot)$

with the initial distribution $\Pi^h$ converges weakly to $\tilde{X}(\cdot)$ with the initial distribution $\Pi$.

By the fact that $X(\cdot)$ is a Feller process, (A2.1) is equivalent to the following: For any sequence $\Pi^h$ and any $q(\cdot)$ which is a bounded, continuous and real–valued function on the Skorohod space $D[G;0,\infty)$ (the space of $G$-valued functions which are right continuous and have left hand limits and with the Skorohod topology),

$$E_{\Pi^h} q(\tilde{X}^h(\cdot)) - E_{\Pi^h} q(\tilde{X}(\cdot)) \to 0, \tag{2.9}$$

as $h \to 0$. The proof that (2.9) holds under (A2.1) uses an argument by contradiction. Suppose that it were false. Then, there is a sequence $\Pi^h$ and a $\rho > 0$ such that the absolute value of the left side of (2.9) is greater than $\rho$. By taking a subsequence, we can suppose, without loss of generality, that there is $\Pi$ such that $\Pi^h \Rightarrow \Pi$. Then, rewrite the left side of (2.9) as

$$\left[E_{\Pi^h} q(\tilde{X}^h(\cdot)) - E_{\Pi} q(\tilde{X}(\cdot))\right] + \left[E_{\Pi} q(\tilde{X}(\cdot)) - E_{\Pi^h} q(\tilde{X}(\cdot))\right].$$

The first term goes to zero by (A2.1) and the second by the Feller property of $X(\cdot)$, thus leading to a contradiction. The converse is proved in a similar manner.

# 3 Occupation Measures: Continuous Time

We now provide the definitions which are needed for the formulation of the limit and robustness results. The methods are based on occupation measure arguments.

**Assumptions and definitions.** The measure valued process $\Pi(\cdot)$ is well defined by (2.4) no matter what the initial condition $\Pi(0)$ is, even if it is *not* the distribution of $X(0)$, or if it is random but *independent* of $(W(\cdot), B(\cdot))$. For example, we might build a filter with an incorrect initial value $\Pi(0)$. This more general interpretation will be important for the sequel. Thus, we can speak of the pair $(X(\cdot), \Pi(\cdot))$ as having an arbitrary initial condition. We say that the process $(X(\cdot), \Pi(\cdot))$ is stationary if the distribution of $(X(t+\cdot), \Pi(t+\cdot))$ does not depend on $t$. From the Feller–Markov property of $X(\cdot)$ and the semigroup relation (2.5) it is easy to show that $(X(\cdot), \Pi(\cdot))$ is a Feller–Markov process. Since it takes values in a compact state space there exists at least one stationary process. Let $\bar{Q}(\cdot)$ denote the measure of the *joint* process $\Psi(\cdot)$ $= (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot), W(\cdot))$, where $(X(\cdot), \Pi(\cdot))$ is stationary. Let $\bar{Q}_f(\cdot)$ denote the measure of the stationary joint process $(X(\cdot), \Pi(\cdot))$.

We make the following key assumption throughout.

**A3.1. A uniqueness assumption.** The process $(X(\cdot), \Pi(\cdot))$ has a unique stationary measure.

The importance of the uniqueness of the stationary joint process was shown in [24]. Some discussion of the uniqueness of $\bar{Q}_f(\cdot)$ is in [2, Section 7], where there is also a discussion of the filtering interpretation of the stationary process. For each $t \geq 0$, define the shifted process $\Psi_f^h(t,\cdot) = \big(\big(X(t+\cdot), \Pi^h(t+\cdot)\big)$ and the centered and/or shifted processes

$$\Psi^h(t,\cdot) = \big(\Psi_f^h(t,\cdot), Y(t+\cdot) - Y(t), B(t+\cdot) - B(t), W(t+\cdot) - W(t)\big)\,.$$

**The path spaces.** The vector-valued processes such as $X(\cdot), Y(\cdot)$, $B(\cdot), \tilde{X}(\cdot)$, and so forth, will take values in the path space $D[I\!R^k; 0, \infty)$; i.e., in the space of $I\!R^k$-valued functions which are right continuous and have left hand limits (CADLAG), with the Skorohod topology [1, 8] for the appropriate value of $k$.

Let $\mathcal{M}(G)$ denote the space of measures on $G$, with the weak topology. Let $m_n(\cdot)$ and $m(\cdot)$ be in $\mathcal{M}(G)$. Recall that $m_n(\cdot)$ converges weakly to $m(\cdot)$ if for each bounded and continuous function $\phi(\cdot)$ on $G$, $\langle m_n, \phi\rangle \to \langle m, \phi\rangle$. Let $\{\phi_i(\cdot)\}$ be a set of continuous functions which are dense (in the topology of uniform convergence) in the set of bounded and continuous functions on $G$. Then weak convergence is equivalent to the metric convergence

$$d(m_n, m) = \sum_i 2^{-i} |\langle m_n - m, \phi_i\rangle| \to 0.$$

The optimal filter $\Pi(t)$ and its approximations $\Pi^h(t)$ at each time $t$ take values in $\mathcal{M}(G)$. The process $\Pi(\cdot)$ and its approximations will take values in the space $D[\mathcal{M}(G); 0, \infty)$, also with the Skorohod topology used.

For a random variable $Z$ and set $A$, let $I_A(Z)$ denote the indicator function of the event that $Z \in A$. Let $C$ be a measurable set in the product path space of $\Psi^h(t,\cdot)$. Define the *occupation measure* $Q^{h,T}(\cdot)$ by

$$Q^{h,T}(C) = \frac{1}{T}\int_0^T I_C(\Psi^h(t,\cdot))dt. \tag{3.1}$$

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 $x(t), y(t)$, etc. Define $\psi_f(\cdot) = (x(\cdot), \pi(\cdot))$, $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$ and $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot), w(\cdot))$.

The random measures $Q^{h,T}(\cdot)$ defined by (3.1) take values in the space of measures on the product path space

$$\mathcal{M}\left(D[I\!R^k; 0, \infty) \times D[\mathcal{M}(G); 0, \infty)\right)$$

for the appropriate value of $k$ (which is the sum of the dimensions of $x, y, b, w$).

**An error or cost function.** Let $F(\cdot)$ be a real–valued function on $D[G; 0, \infty) \times D[\mathcal{M}(G); 0, \infty)$ which is measurable and continuous (w.p.1 with respect to the measure $\bar{Q}_f(\cdot)$). Owing to the compactness of $G$, we can suppose that $F(\cdot)$ is bounded. As in [2], we are concerned with the asymptotic (pathwise) behavior

of the sample averages $\int_0^T F(\Psi_f^h(t,\cdot))dt/T$ as $h \to 0$ and $T \to \infty$. Let $Q_f^{h,T}(\cdot)$ denote the $(X(\cdot), \Pi^h(\cdot))-$marginal of $Q^{h,T}(\cdot)$, i.e. for arbitrary measurable set $C'$ in the product path space of $\Psi_f^h(t,\cdot)$

$$Q_f^{h,T}(C') = \frac{1}{T}\int_0^T I_{C'}(\Psi_f^h(t,\cdot))dt.$$

By the definition of the occupation measure, we can write

$$\frac{1}{T}\int_0^T F(\Psi_f^h(t,\cdot))dt = \int F(\psi_f(\cdot))Q_f^{h,T}(d\psi_f(\cdot)). \tag{3.2}$$

The representation (3.2) shows that the asymptotic values of the left hand side can be obtained from the limits of the set of occupation measures $Q_f^{h,T}$, as $h \to 0$ and $T \to \infty$.

It was shown in [2] that for a broad class of approximate filters

$$\frac{1}{T}\int_0^T F(\Psi_f^h(t,\cdot))dt \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)) \tag{3.3}$$

in probability. It was also shown that

$$\frac{1}{T}\int_0^T F(X(t+\cdot), \Pi(t+\cdot)))dt \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)) \tag{3.3'}$$

where $\Pi(\cdot)$ in (3.3') is the true optimal filter. Note that via an application of dominated convergence theorem we can replace the expressions on the left sides of (3.3) and (3.3′) by their expected values. These results say that sample pathwise average errors of many types will converge to the same stationary value that one would get if the true optimal filter were used, and the pathwise average error were replaced by its expectation. This desired result is formalized in the theorem below.

The convergence is in the sense of probability, and holds as $T \to \infty$ and $h \to 0$ in any way at all. The arbitrariness of the way that $T \to \infty$ and $h \to 0$ is crucial in applications. It is important that the approximation is good for all small $h$, not depending on $T$, if $T$ is large enough.

Let $\phi(\cdot)$ be a bounded, continuous and real–valued function. A special case of (3.3) is the convergence of the mean square error

$$\begin{aligned} G^{h,T}(\phi) \equiv \frac{1}{T}\int_0^T \left[\langle \Pi^h(t), \phi\rangle - \phi(X(t))\right]^2 dt \\ \to \int \left[\langle \pi(0), \phi\rangle - \phi(x(0))\right]^2 \bar{Q}_f(d\psi_f(\cdot)) \end{aligned} \tag{3.4}$$

in the sense of probability as $h \to 0$ and $T \to \infty$ in any way at all. The right side of (3.4) is what one would also get as the limit if the true optimal filter were used (and even with an expectation of the pathwise average used) [2]. In

this sense there is pathwise asymptotic optimality of the approximating filter over the infinite time interval.

The following is the main background theorem from [2].

**Theorem 3.1.[2, Theorem 3.2.]** *Let the filtering model be as in Section 2. Assume the uniqueness condition* (A3.1). *Define the approximate filter* $\Pi^h(\cdot)$ *via* (2.7) *where* $\tilde{X}^h(\cdot)$ *satisfies the consistency condition* (A2.1). *Then, for every sequence* $\{h_k, T_k\}_{k\geq 1}$ *such that* $h_k \to 0$ *and* $T_k \to \infty$ *as* $k \to \infty$, *the family* $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ *is tight. Extract a weakly convergent subsequence with weak sense limit denoted by* $Q(\cdot)$, *a measure–valued random variable. Let* $Q^\omega(\cdot)$ *denote the sample values of* $Q(\cdot)$. $Q^\omega(\cdot)$ *induces a process, denoted by*

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

*Here, the* $\omega$ *indexes the process, not the sample paths of the process. For almost all* $\omega$ *the following hold. The processes* $(B^\omega(\cdot), W^\omega(\cdot))$ *are independent standard Wiener, with respect to which* $\{X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot)\}$ *are non anticipative.*

$$dY^\omega = g(X^\omega)dt + dB^\omega, \tag{3.5}$$

$$dX^\omega = p(X^\omega)dt + \sigma(X^\omega)dW^\omega. \tag{3.6}$$

*For each bounded and measurable real–valued function* $\phi(\cdot)$

$$\langle \Pi^\omega(t), \phi\rangle = \frac{E_{\{\Pi^\omega(0), Y^\omega_{0,t}\}}\left[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y^\omega_{0,t})\right]}{E_{\{\Pi^\omega(0), Y^\omega_{0,t}\}}R(\tilde{X}_{0,t}, Y^\omega_{0,t})}. \tag{3.7}$$

*Equivalently, for all* $t, s$*:*

$$\langle \Pi^\omega(t+s), \phi\rangle = \frac{E_{\{\Pi^\omega(t), Y^\omega_{t,t+s}\}}\left[\phi(\tilde{X}(t+s))R(\tilde{X}_{0,s}, Y^\omega_{t,t+s})\right]}{E_{\{\Pi^\omega(t), Y^\omega_{t,t+s}\}}R(\tilde{X}_{0,s}, Y^\omega_{t,t+s})}. \tag{3.8}$$

$(X^\omega(\cdot), \Pi^\omega(\cdot))$ *is the unique stationary process and hence its distribution does not depend on* $\omega$ *or on the chosen convergent subsequence. Finally,* (3.3) *holds in probability as* $h \to 0$ *and* $T \to \infty$ *in any way at all, for any bounded and measurable real–valued function* $F(\cdot)$ *which is continuous almost everywhere with respect to* $\bar{Q}_f(\cdot)$.

# 4 The Discrete Time Problem

Now, we review the discrete time form of the results in the previous section. Let all processes be defined in discrete time. The signal process $X(\cdot) = \{X(n), n < \infty\}$ is assumed to be Feller–Markov and takes values in the compact set $G$. The observations are defined by $Y(0) = 0$ and

$$\delta Y_n \equiv Y(n) - Y(n-1) = g(X(n)) + \xi(n),\ n = 1, \ldots, \tag{4.1}$$

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 true conditional distribution of $X(n)$ given $Y_{0,n}$ can be represented in terms of an auxiliary process $\tilde{X}(\cdot)$ as for the continuous time case in Section 2, 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 $\tilde{X}(0)$. Alternatively,

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(n-1), \delta Y_n\}}\left[\phi(\tilde{X}(1))R(\tilde{X}(1), \delta Y_n)\right]}{E_{\{\Pi(n-1), \delta Y_n\}}R(\tilde{X}(1), \delta Y_n)}. \tag{4.2}$$

Analogously to the continuous time observation case, except in some special cases one cannot evaluate (4.2), and it is generally necessary to approximate it in some way. The approximation problems and methods are similar to those in Section 3. For example, the sequence $X(\cdot)$ might be samples of a diffusion process taken at discrete instants $0, \Delta, 2\Delta, \ldots$, and to evaluate (4.2) the transition function must then be computed, at least approximately. We could approximate the solution to the Fokker–Planck equation on $[0, \Delta]$ by the Markov chain approximation method, or by other numerical means. Even if $X(\cdot)$ is actually originally defined as a discrete time process, it might be too hard to do the necessary integrations in (4.2) with the true transition function, and an appropriate approximation might be needed. In such cases, one often uses an approach which is analogous to what was done in the continuous time case. Namely, build a filter for a simpler Markov process $\tilde{X}^h(\cdot)$ (in discrete time here) which has values in the compact set $G$, and which approximates $X(\cdot)$, but use the actual physical observations.

This procedure is formalized as follows. For $n = 1, \ldots,$ 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:

$$\langle \Pi^h(n), \phi\rangle = \frac{E_{\{\Pi^h(0), Y_{0,n}\}}\left[\phi(\tilde{X}^h(n))R(\tilde{X}^h_{0,n}, Y_{0,n})\right]}{E_{\{\Pi^h(0), Y_{0,n}\}}R(\tilde{X}^h_{0,n}, Y_{0,n})}. \tag{4.3}$$

With an abuse of notation, define

$$R(x,y) = \exp\left[g(x)'y - |g(x)|^2/2\right].$$

Then, one has the following recursive representation for $\Pi^h(\cdot)$.

$$\langle \Pi^h(n), \phi\rangle = \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}^h(1),\delta Y_n)}. \tag{4.4}$$

Equations (4.3) and (4.4) correspond to the filter which models the signal process via $\tilde{X}^h(\cdot)$ but uses the actual observations $\delta Y_n = Y(n) - Y(n-1)$

The process $\tilde{X}^h(\cdot)$ in (4.4) is assumed to be independent of the other processes, given its initial distribution. We also use the analog of the basic consistency assumption (A2.1), which is the sense of approximation of $X(\cdot)$ by $\tilde{X}^h(\cdot)$:

**A4.1. A consistency assumption.** For any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure $\Pi$, $\tilde{X}^h(\cdot)$ with the initial distribution $\Pi^h$ converges weakly to $\tilde{X}(\cdot)$ with the initial distribution $\Pi$.

It is easy to see that if $\Pi^h(0)$ converges weakly to the distribution of $X(0)$, then $(X^h(\cdot), \Pi^h(\cdot))$ converges weakly to the true (signal, filter) pair $(X(\cdot), \Pi(\cdot)$. Analogous to the remark below (A2.1), (A4.1) is equivalent to the following condition: Let $q(\cdot)$ be a bounded and continuous function on the path space. Then, for any sequence $\Pi^h$ as $h \to 0$

$$E_{\Pi^h} q(\tilde{X}^h(\cdot)) - E_{\Pi^h} q(\tilde{X}(\cdot)) \to 0. \tag{4.5}$$

The above assumption implies that if $\tilde{X}^h(0)$ converges weakly (for any subsequence of values of $h$) with limit distribution $\Pi(0)$, then the sequence $\{\tilde{X}^h(n); n \geq 1\}$ converges weakly to $\{X(n); n \geq 1\}$ with initial distribution $\Pi(0)$. In fact, in view of the semigroup property of the filter , we need only deal with $\{\tilde{X}^h(0), \tilde{X}^h(1)\}$.

Analogously to the situation in Section 3, (4.2) is well defined even if $\Pi(0)$ is not the initial distribution of $X(\cdot)$. Allowing the initial condition $(X(0), \Pi(0))$ to be arbitrary, the discrete time process $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$ is Feller–Markov. We now write the discrete time analog of the key uniqueness assumption:

**A4.2. 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 define the analog of $\Psi^h(t,\cdot)$, namely:

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

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

Define the canonical elements of the path spaces $\psi(\cdot)$ and $\psi_f(\cdot)$ analogously, as done in Section 3.

The Skorohod topology is replaced by a "sequence" topology, as follows. The $\Pi^h(n)$ still take values in $\mathcal{M}(G)$, and the weak topology is still used on this space. Let $d^\pi(\cdot)$ and $d^k(\cdot)$ denote the metrics on $\mathcal{M}(G)$ (induced by the weak topology) and on $I\!R^k$, resp., where $k$ is the sum of the dimensions of $X(n), B(n)$ and $Y(n)$. Let $d_0(\cdot)$ denote the product metric. Let $a(\cdot) = \{(a(1), \ldots\}$ and $b(\cdot) = \{(b(1), \ldots\}$ be sequences with the $a(n)$ and $b(n)$ taking values in the product space $\mathcal{M}(G) \times I\!R^k$. Then the metric on the product path (sequence) space is

$$d(a(\cdot), b(\cdot)) = \sum_{n=0}^{\infty} 2^{-n} \left[d_0(a(n), b(n)) \wedge 1\right].$$

Define the occupation measure $Q^{h,N}(\cdot)$ by: for a Borel set $C$ in the product sequence space,

$$Q^{h,N}(C) = \frac{1}{N} \sum_{n=1}^{N} I_C(\Psi^h(n, \cdot)). \tag{4.6}$$

Analogously to the definitions in Section 3, define $\Psi(\cdot) = (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot))$. 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 discrete time analog of Theorem 3.1 is proved in [2].

**Theorem 4.1.** *Let the filtering model be as above. Assume that* (A4.2) *holds. Define the approximate filter via* (4.3)*, where we assume that the auxiliary process* $\tilde{X}^h(\cdot)$ *satisfies* (A4.1)*. Then* $\{Q^{h,N}(\cdot); h > 0, N \geq 0\}$ *is tight. Let* $Q(\cdot)$ *denote a weak sense limit, always as* $h \to 0$ *and* $N \to \infty$*. Let* $\omega$ *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* $\omega$*,* $Q^\omega(\cdot)$ *is a measure on the product path* (*sequence*) *space. It induces a process*

$$\Psi^\omega(\cdot) = (X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot)). \tag{4.7}$$

*For almost all* $\omega$ *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*

$$\delta Y_n^\omega \equiv Y^\omega(n) - Y^\omega(n-1) = g(X^\omega(n)) + \xi^\omega(n), \tag{4.8}$$

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

$$\langle \Pi^\omega(n), \phi \rangle = \frac{E_{\{\Pi^\omega(0), Y_{0,n}^\omega\}} \left[\phi(\tilde{X}(n)) R(\tilde{X}_{0,n}, Y_{0,n}^\omega)\right]}{E_{\{\Pi^\omega(0), Y_{0,n}^\omega\}} R(\tilde{X}_{0,n}, Y_{0,n}^\omega)}. \tag{4.9}$$

*Finally,*

$$\frac{1}{N}\sum_{n=1}^{N} F(\Psi_f^h(n+\cdot)) = \int F(\psi_f(\cdot))\bar{Q}_f^{h,N}(d\psi_f(\cdot)) \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)) \tag{4.10}$$

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

**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 [2], in the sequel we will try to use as much of the proof in [2] as possible, and to concentrate on the differences. In view of that we now highlight the chief features of the proof in [2]. We comment on the discrete parameter case, but analogous remarks hold for the continuous parameter model. Further details are in the reference. In [2] (see (4.6)), 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 what follows, but with the new definitions of $\Pi^h(\cdot)$ of this paper used.

The first step in the proof of Theorem 4.1 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]. In order to show that, 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, showing that is trivial in view of the compactness of the state space. We note that in the continuous time case the proof of tightness of these processes involves a little more work.

By the first equality in (4.10), 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 [2] will work without any change for the problems of this paper. Furthermore the proofs of the representation (4.8) and that $X^\omega(\cdot)$ has the law of evolution of $X(\cdot)$ for almost all $\omega$ will be no different than the analogous arguments in [2], and similarly for the continuous parameter case. Thus, establishing the representation (4.9) becomes the only step in the proof that will be different from that in [2]. Once this step is established, (4.10) follows readily from the uniqueness assumption on the invariant measure of the joint signal and filter process.

The proof of the representation for the $\Pi^\omega(\cdot)$ will differ slightly, depending on the choice of $\Pi^h(\cdot)$. The following comments concerning a key detail in the proof of the representation (4.9) in [2] 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

$$A(\psi(m)) = \langle \pi(m), \phi \rangle - \frac{E_{\{\pi(m-1),y(m)\}} \left[ \phi(\tilde{X}(1)) R(\tilde{X}(1), y(m)) \right]}{E_{\{\pi(m-1),y(m)\}} R(\tilde{X}(1), y(m))}. \tag{4.11}$$

The aim of the proof in [2] was to show that, for almost all $\omega$ and all $m$,

$$A(\Psi^{\omega}(m)) = 0, \text{ with probability 1,} \tag{4.12}$$

which implies (4.9). This was done by showing that

$$0 = E \int Q^{\omega}(d\psi) \left[ A(\psi(m)) \right]_1^2, \tag{4.13}$$

where we define

$$[A]_1^2 = \min\{|A|^2, 1\}. \tag{4.14}$$

The prelimit form of the right side of (4.13) is

$$E \int Q^{h,N}(d\psi) \left[ A(\psi(m)) \right]_1^2, \tag{4.15}$$

which, by the definition of $Q^{h,N}(\cdot)$, equals

$$\frac{1}{N} E \sum_{n=1}^{N} \left[ A(\Psi^h(m+n)) \right]_1^2, \tag{4.16}$$

where

$$A(\Psi^h(n)) = \langle \Pi^h(n), \phi \rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}} \left[ \phi(\tilde{X}(1)) R(\tilde{X}(1), \delta Y_n) \right]}{E_{\{\Pi^h(n-1),\delta Y_n\}} R(\tilde{X}(1), \delta Y_n)}. \tag{4.17}$$

In order to show (4.13) it suffices to show

$$E[A(\Psi^h(n))]_1^2 \to 0,$$

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 (A4.1), to show that

$$E \left[ \langle \Pi^h(n), \phi \rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}} \left[ \phi(\tilde{X}^h(1)) R(\tilde{X}^h(1), \delta Y_n) \right]}{E_{\{\Pi^h(n-1),\delta Y_n\}} R(\tilde{X}^h(1), \delta Y_n)} \right]_1^2 \tag{4.18}$$

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

However in view of the definition of $\Pi^h(n)$ via (4.4) the above expression is identically zero, which implies that (4.13) holds for any weak sense limit. An analog of this argument will be used in the next section.

# 5 Some Approximating Filters of Interest: Discrete Time

In [2], the approximate filter $\Pi^h(\cdot)$ was defined by the analytical formula (4.4) for the discrete time problem, and by (2.7) for the continuous time problem. One example is the Markov chain approximation method, where the auxiliary process $\tilde{X}^h(\cdot)$ is a Markov chain approximation to $\tilde{X}(\cdot)$. When the dimension is high, such methods can have excessively high computational requirements. Alternatives, based on random sampling or Monte Carlo then become attractive, analogous to the case of classical multidimensional integration [3, 4, 5, 11, 12, 14, 15, 26, 27, 28]. In this section, several forms of this approach will be discussed. We start with the simplest form, which uses unsophisticated random sampling to evaluate the right hand side of (4.4). The problem is set up so that much of the proof of [2, Theorem 5.1] (this is Theorem 4.1 above) can be used. After treating this simple (but canonical) case, we then move on to more general approximations, pointing out at each instance the crucial condition required for the analog of Theorem 4.1 to hold.

**Example 1. The basic "sampling" filter.** Let $v^h$ be a sequence of integers which goes to infinity as $h \to 0$. Let $\Pi^h(n-1)$ denote the estimate of the conditional distribution of $X(n-1)$, given $Y_{0,n-1}$. Given $\Pi^h(n-1)$, we now construct $\Pi^h(n)$ based on "random sampling." Let $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ be i.i.d samples (which are independent of $\delta Y_n$, conditioned on $\Pi^h(n-1)$) from $\tilde{X}^h(\cdot)$, where $\tilde{X}^h(\cdot)$ satisfies the consistency condition (A4.1) and has the initial distribution $\Pi^h(n-1)$. One need only simulate samples of $\tilde{X}^h(0), \tilde{X}^h(1)$.

The filter $\Pi^h(n)$ is defined by the sample average:

$$\langle \Pi^h(n), \phi \rangle = \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}, \tag{5.1}$$

which yields our estimate $\Pi^h(n)$ of the conditional distribution of $X(n)$, given $Y_{0,n}$.

**Theorem 5.1.** *Under* (A4.1) *and* (A4.2) *and the above construction of* $\Pi^h(\cdot)$*, the conclusions of Theorem* 4.1 *hold.*

**Proof.** The basic steps in the proof of Theorem 4.1 were outlined after the statement of that theorem. The proof of the current theorem is similar, and we will only concern ourselves with the differences.

The set $\{Q^{h,N}(\cdot); h > 0, T < \infty\}$ is obviously tight since each of the families $\{X(n+\cdot); n \ge 0\}$,$\{B(n+\cdot) - B(n); n \ge 0\}$, $\{Y(n+\cdot) - Y(n); n \ge 0\}$, $\{\Pi^h(n+\cdot); h > 0, n \ge 0\}$ is tight. Let $Q(\cdot) = \{Q(n), n = 0, 1, ..\}$ denote the limit of a weakly convergent subsequence, and denote the samples by $Q^\omega(\cdot)$. Then $Q^\omega(\cdot)$ induces a process $\Psi^\omega(\cdot)$ as in (4.7), and we need to identify the components. The stationarity of $(X^\omega(\cdot), \Pi^\omega(\cdot))$ is proved as in [2], with no change. Similarly,

the characterization (4.8), the properties of $B^{\omega}(\cdot)$ and the fact that $X^{\omega}(\cdot)$ has the transition function of $X(\cdot)$ is done exactly as in [2].

The main difference is in the proof of (4.9). Proceeding as illustrated for Theorem 4.1, to identify $\Pi^{\omega}(\cdot)$ we need only to show (4.13). Analogously to the procedure in Section 4, this is done by showing that the expression in (4.18) converges to 0, uniformly in $n$ as $h \to 0$. By using the definition of $\Pi^h(n)$, (4.18) can be rewritten as

$$E\left[\frac{\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h} - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}^h(1),\delta Y_n)}\right]_1^2. \tag{5.2}$$

Owing to the properties of the $|\cdot|_1^2$ metric defined by (4.14), we can work with the numerators and denominators separately, and it is only necessary to show that, for arbitrary bounded and continuous $\phi(\cdot)$,

$$E\left[\frac{1}{v^h}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n) - E_{\{\Pi^h(n-1),\delta Y_n\}}\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)\right]_1^2 \tag{5.4}$$

goes to zero, uniformly in $n$, as $h \to 0$. But, this clearly holds since for each $h$ and $n$, $\{\tilde{X}^{h,l,n}(\cdot), l\}$ are mutually independent, identically distributed and independent of $\delta Y_n$ (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. ■

We remark that in the above proof we found it convenient to work with the expression in (5.4), however in view of the consistency condition (A4.1) on $\tilde{X}^h(\cdot)$, showing that the expression in (5.4) goes to zero, uniformly in $n$, as $h \to 0$ is equivalent to showing the same for the expression in (5.4′) below.

$$E\left[\frac{1}{v^h}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n) - E_{\{\Pi^h(n-1),\delta Y_n\}}\phi(\tilde{X}(1))R(\tilde{X}(1),\delta Y_n)\right]_1^2 \tag{5.4′}$$

**Example 2. Some generalizations of the filter in Example 1.** As can be observed from the proof of Theorem 5.1, the crucial step is the establishing of

convergence of the expression in (4.18) or, equivalently, of (5.4′), the form that we will use. This convergence is essentially the consequence of the consistency condition (A4.1). However, as we will indicate in the following discussion, this consistency condition can be weakened considerably. This leads to many useful extensions of the basic form of the "sampling" algorithm of Example 1.

**A weaker form of the consistency assumption (A4.1).** We retain the assumption of mutual independence (conditional on $\Pi^h(n-1), \delta Y_n$) of the $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ for each $h, n$, and that the probability law of $\{\tilde{X}^{h,l,n}(0)\}$ is $\Pi^h(n-1)$, but allow more flexibility in the choice of the individual $\tilde{X}^{h,l,n}(\cdot)$. Namely, in the construction of $\Pi^h(n)$ in (5.1), the Markov family from which $\tilde{X}^{h,l,n}(\cdot)$ is sampled may differ for different $l, n$. However the initial conditions $\tilde{X}^{h,l,n}(0)$ still form an i.i.d sample from $\Pi^h(n-1)$. To see the possibilities, write the expression in the brackets in (5.2) as the sum of the two terms:

$$\begin{aligned}&\frac{\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}\\ &\quad-\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h},\end{aligned} \tag{5.5}$$

and

$$\begin{aligned}&\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}\\ &\quad-\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\left[\phi(\tilde{X}(1))R(\tilde{X}^h(1),\delta Y_n)\right]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}(1),\delta Y_n)}.\end{aligned} \tag{5.6}$$

Owing to the use of the $[\cdot]_1^2$ metric defined by (4.14), it is enough to work separately with the differences of the numerators and of the denominators in each of (5.5) and (5.6). Then, to handle (5.5), use the mutual independence and the uniform bounds on the expectations of the conditional variances. To handle (5.6), we will use a revised form of the consistency assumption (A4.1), which is:

**A5.1.** For each $(n, h)$, the set $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is mutually independent and independent of $\delta Y_n$, conditioned on $\Pi^h(n-1)$. Suppose that an arbitrary $\Pi^h$ 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 function $\Phi(\cdot)$,

$$E_{\Pi^h}\Phi(\tilde{X}^{h,l,n}(1)) - E_{\Pi^h}\Phi(\tilde{X}(1)) \to 0 \tag{5.7}$$

uniformly in $n$ and $l$.

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 (5.6). Observe that even though $R(\cdot)$ is not bounded we can, without loss of generality, assume so

since the family $\{\delta Y_n; n \geq 1\}$ is tight. For this reason and the fact that we use the metric (4.14), we don't need the convergence in (A5.1) 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 (A5.1) will hold iff for all $\Pi$ 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

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

where $\tilde{X}(0)$ has the law $\Pi$.

**Dropping the mutual conditional independence.** Return to the expression (5.2). Let $\Phi(\cdot)$ be bounded and continuous. Then the convergence in (5.2) is implied by the even weaker consistency assumption, which can replace (A4.1) and the mutual independence in Theorem 5.1:

**A5.2.** For each $(h, n)$, $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is independent of $\delta Y_n$, 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 $\Pi^{h,n}$ (on $G$) takes the role of $\Pi^h(n-1)$ in the construction of the $\{\tilde{X}^{h,l,n}(\cdot), l\}$. Then the associated process $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is constructed such that as $h \to 0$ and, for any bounded, continuous and real valued function $\Phi(\cdot)$,

$$\frac{1}{v^h} \sum_{l=1}^{v^h} \Phi(\tilde{X}^{h,l,n}(1)) - E_{\{\Pi^{h,n}\}} \Phi(\tilde{X}(1)) \to 0, \tag{5.8}$$

in probability, uniformly in $n$.

It is clear that this condition (instead of (A4.1) and mutual independence of samples) suffices for Theorem 5.1 to hold for the corresponding $\{\Pi^h(n)\}$. The usefulness of this condition lies in the cases where the samples $\{\tilde{X}^{h,l,n}(\cdot)), l \leq v^h\}$ for fixed $h, n$ are not mutually independent. It is of particular value when the random sampling incorporates some variance reduction method where the samples are not mutually independent; e.g., antithetic variables or stratified sampling such as discussed next.

**Variance reduction methods.** The standard methods for variance reduction in Monte Carlo, such as stratified sampling and antithetic variables, can all be used here and in the subsequent algorithms and examples. We comment on one form of stratified sampling. 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 $x^{h,l,n}$.

In this example, we only use variance reduction to get the samples of the initial values $\tilde{X}^{h,l,n}(0)$. Once these are given, the sample values of $\tilde{X}^{h,l,n}(1)$ are obtained by sampling independently, using the transition probability of the approximating Markov process $\tilde{X}^{h,n}(\cdot)$. So we concentrate on the initial values for fixed $n$.

If the $v^h\Pi_l^h(n-1)$ were all integers, then the best sampling of the values of the $\tilde{X}^{h,l,n}(0)$ would be to take the initial point $x^{h,l,n}$ exactly $v^h\Pi_l^h(n-1)$ times, since then the variance of the sampling error of the *initial condition* would be zero. Clearly all the $v^h\Pi_l^h(n-1)$ would not usually be integers, but one tries to approximate the ideal as well as possible. One common approach is the following. First take the point $x^{h,l,n}$ exactly $[v^h\Pi_l^h(n-1)]$ (the integer part) times. After this step, the "residual" number of points remaining to be chosen is

$$\delta v^{h,n} = \sum_l \delta v_l^{h,n},$$

where

$$\delta v_l^{h,n} = \left(v^h\Pi_l^h(n-1) - [v^h\Pi_l^h(n-1)]\right).$$

The "residual frequency" of point $x^{h,l,n}$ is $\delta v_l^{h,n}/\delta v^{h,n}$. Now, divide the set $\{x^{h,l,n}, l\}$ into disjoint subsets $S_i^{h,n}, i = 1, \ldots$. The set $S_i^{h,n}$ has $\delta v^{h,n,i}$ points where

$$\delta v^{h,n,i} = \sum_{l \in S_i^{h,n}} \delta v_l^{h,n}.$$

Allocate $[\delta v^{h,n,i}]$ initial points to subset $i$, and then select these points randomly (with replacement) from $S_i^{h,n}$, where the point $x^{h,l,n} \in S_i^{h,n}$ is given the weight $\delta v_l^{h,n}/\delta v^{h,n,i}$. 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, if the division into subgroups is done properly, $\bar{v}^{h,n}/v^h$ will be either zero or small. If it is positive, either repeat the above procedure to allocate the remaining $\bar{v}^{h,n}$ points, or just select $\bar{v}^{h,n}$ points randomly from the original $v^h$ points with appropriately modified weights.

It is easy to see that the above construction can be put in the framework of Example 2 and condition (A5.2) holds.

The grouping into subsets might be done by dividing the points according to their "geographic location," if this is meaningful.

**Example 3.** We would like to treat algorithms that use combinations of random sampling and integration methods in a general way. This will require an alteration in the consistency condition (A5.1) or (A5.2). In order to motivate the form which it will take, we first consider an example for which (A5.2) is

satisfied. Let $\tilde{X}^{h,n}(\cdot)$ be processes satisfying (A5.1). Having defined the approximate filter $\Pi^h(j)$ 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 that they are conditionally independent of $\delta Y_n$ given $\Pi^h(n-1)$. Define $\Pi^h(n)$ via (5.1). 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 condition (A5.1) (and thus (A5.2)) is satisfied. Theorem 5.1 can be proved under weaker consistency conditions than (A5.1) or (A5.2), which allow great and useful flexibility in constructing the filter. To motivate a useful general form, let us first rewrite Example 2 in the following suggestive way.

For each $h$ and $n$, define a measure (on the sample space $G \times G$) 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 $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ that are in the Borel set $A \subset G \times G$. In particular, $P^{h,n}_{\Pi^h(n-1)}\{B \times G\}$ is the fraction of the samples $\tilde{X}^{h,l,n}(0)$ which are in the set $B$. Since this $P^{h,n}_{\Pi^h(n-1)}$ is just the "sampling occupation measure," condition (A5.2) is equivalent to

$$E\left[\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1))\right]^2_1 \to 0, \tag{5.9}$$

uniformly in $n$ as $h \to 0$.

Thus the crucial condition becomes the convergence of the expression in (5.9). The advantage of writing the condition in the form (5.9) is that, being written in terms of a random measure $P^{h,n}_{\Pi^h(n-1)}$, it suggests other choices of approximate filters that need not be based exclusively on Monte Carlo or random sampling. For example, as seen in Example 4 below, $P^{h,n}_{\Pi^h(n-1)}$ might be determined partly by random sampling and partly analytically.

**A generalization of $P^{h,n}_{\Pi^h(n-1)}$ and the approximating filter.** Motivated by the suggestiveness of (5.9), we now consider the following general form of the approximate filter and the consistency condition. Let $\{\Pi^h(n); n \geq 1\}$ be defined recursively as follows. Having defined $\Pi^h(n-1)$, 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 $\delta Y_n$ given $\Pi^h(n-1)$. Define $\Pi^h(n)$ by

$$\langle \Pi^h(n), \phi \rangle = \frac{\int \phi(x(1))R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}{\int R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}. \tag{5.10}$$

We will need the following consistency condition.

**A5.3.** For each bounded, continuous and real valued function $\Phi(\cdot)$, as $h \to 0$,

$$\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1)) \to 0, \tag{5.11}$$

in probability, uniformly in $n$.

We now have the following useful result whose proof follows from the above comments.

**Theorem 5.2.** *Theorem* 5.1 *holds for the above constructed* $\Pi^h(\cdot)$ *if in the assumptions of that theorem the consistency condition* (A5.3) *replaces* (A4.1) *and the mutual independence of the samples.*

**Remarks.** Although not needed, it will often be the case that

$$E_{\Pi^h(n-1)}P^{h,n}_{\Pi^h(n-1)}\{B \times G\} = \Pi^h(n-1)(B). \tag{5.12}$$

The advantage of (A5.3) is that it can be used for a large variety of approximation methods. For example, in the form (4.4), $P^{h,n}_{\Pi^h(n-1)}$ 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 frames can all be put into the form of (5.11) for appropriate choices of $P^{h,n}_{\Pi^h(n-1)}$. Importance sampling methods can also be fit into the same scheme and used to improve the performance of the filter, as shown in the next section.

**Example 4. An application of (A5.3): Combined random sampling and integration.** Consider the following commonly used model. Let $X(n) = b(X(n-1), \zeta(n-1))$, where $b(\cdot)$ is bounded and continuous and the $\{\zeta(n)\}$ are mutually independent, identically distributed (with distribution function $P_\zeta$, 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 (4.2). If the computation on the right side of (4.2) is not possible, as is usually the case, it would be approximated in some way. The difficulties in evaluating the right side might be due to the problem of computing the one step transition probability of the Markov process $\{X(n)\}$, or to the actual integrations over a possibly continuous state space that are required to evaluate (4.2). As usual, let $h$ denote the approximation parameter for the actual practical filter, and $\Pi^h(n)$ the estimate of the conditional distribution given $Y_{0,n}$.

Let $\Pi^h(n-1)$ be given. We wish to compute $\Pi^h(n)$. This can be done by a direct simulation as in Example 1 or by combined "simulation-integration" or perhaps even by a pure "integration" method. These possibilities will be illustrated. Suppose that we approximate $P_\zeta$ by $P^h_\zeta$, which might have a (computationally) more convenient support and is such that $P^h_\zeta \Rightarrow P_\zeta$. In addition, approximate $b(\cdot)$ by a measurable function $b_h(\cdot)$ such that

$$\lim_{h\to 0}\sup_{x,\zeta}|b(x,\zeta) - b_h(x,\zeta)| = 0.$$

If the associated integrations are convenient to carry out, one can use (4.3)

to define $\Pi^h(n)$, where we define

$$\tilde{X}^h(1) = b_h(\tilde{X}^h(0), \zeta^h), \tag{5.13a}$$

In (5.13a), $\tilde{X}^h(0)$ has distribution $\Pi^h(n-1)$ and $\zeta^h$ has distribution $P^h_\zeta$. If the support of $P^h_\zeta$ is finite and $b_h(\cdot)$ takes only finitely many values, then the integrations reduce to summations. The $\tilde{X}^h(\cdot)$ process thus defined satisfies the consistency condition (A4.1). Hence Theorem 4.1 holds for $\Pi^h(\cdot)$ if (A4.2) holds.

Alternatively, one can simply use Monte Carlo as in Example 1. All sampling below is "conditionally independent" of the past, given $\Pi^h(n-1)$. Take $v^h$ independent samples from $\Pi^h(n-1)$ and from $P^h_\zeta$, call them $\tilde{X}^{h,n,l}(0)$ and $\zeta^{h,n,l}$, $l \le v^h$, resp. Then use the formula

$$\tilde{X}^{h,n,l}(1) = b_h(\tilde{X}^{h,n,l}(0), \zeta^{h,n,l}), \tag{5.13b}$$

and (5.1) to get $\Pi^h(n)$.

Combinations of the above two approaches might be worthwhile also. For example, if the support of the $P^h_\zeta$ is a (not too big) finite set, then one can sample from $\Pi^h(n-1)$, but "integrate" over the noise for each sample of the initial condition, by doing the necessary summations. One would normally try to choose the support of $P^h_\zeta$ such that the integrals are well approximated for an appropriate set of functions $\phi(\cdot)$. On the other hand, one might discretize the state space such that the support of the $\tilde{X}^h(0)$ (i.e., of each of the $\Pi^h(n)$) is confined to a finite set $G_h$, and integrate with respect to the "initial" measure $\Pi^h(n-1)$, but simulate the noise. For each of these combinations, there is a $P^{h,n}_{\Pi^h(n-1)}$ such that (A5.3) holds, provided that the discretization of the space converges to the whole space in an appropriate manner and, for the part of the computation which involves random sampling, 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.

**Example 5: A Markov chain approximation method. A sampled diffusion model.** In this example, we illustrate a potentially useful combination of integration and simulation. Suppose that the signal process $X(n)$ is a sample from a diffusion process $X(\cdot)$ at discrete time $n$. Suppose that $X(\cdot)$ solves an Itô equation with a unique weak sense solution for each initial condition, and has continuous drift and diffusion coefficients. Then the exact filter (4.2) involves getting the probability distribution of $\tilde{X}(1)$ (which solves the same Itô equation) with the correct initial distribution $\Pi(n-1), n = 1, \ldots$ . One could try to solve the Fokker-Planck equation by some numerical method. This is not easy when there are degeneracies. The Markov chain approximation method [23] is a general and powerful approach, which converges under the specified conditions, even with quite weak (reflecting) boundary reflections and jumps added. The following discussion uses the idea of the Markov chain approximation, without going into excessive details.

For each $h$, let $\{X_n^h\}$ be a discrete parameter Markov chain on a finite state space $G_h \subset G$, and let $p^h(x,z)$ denote the one step transition probabilities. For $\delta_h > 0$, define the continuous time interpolation $\tilde{X}^h(\cdot)$ by $\tilde{X}^h(t) = X_n^h$ on the interval $[n\delta_h, n\delta_h + \delta_h)$. Suppose, without loss of generality, that $1/\delta_h$ is an integer, and assume that $\tilde{X}^h(\cdot)$ satisfies the consistency assumption (A4.1). The use of such chains in the construction of approximate filters is now quite common. See [20, 23, 22]. The references [20, 23] give straightforward and automatic ways of constructing such chains.

In [20, 23] and in current usage in applications, the process $\tilde{X}^h(\cdot)$ is used as in the algorithm (4.4). But, it can also be used as the basic simulated process in (5.1). In order to demonstrate the possibilities, we now illustrate an interesting combination of these two schemes which is rather different from the combinations illustrated by Example 4. We work with a single observation interval at a time, and for concreteness we discuss the method for the time interval $[0,1]$. Let $\Pi^h = \Pi^h(0)$ denote the approximation to the distribution of $X(0)$. We need to approximate the distribution of $\tilde{X}(1)$. This is done by either computing or estimating the distribution of $X_{1/\delta_h}^h$, where $X_0^h$ has the distribution $\Pi^h$.

To estimate or compute the distribution of $X_{1/\delta_h}^h$, we recursively estimate or compute the distribution of the $X_m^h$ for $m = 1, 2, \ldots, 1/\delta_h$. The motivation behind the combined integration/monte carlo procedure to be described is that in some regions of the state space, it might be easier to use one method and in other regions the other method. For illustrative purposes, we suppose that $G$ is divided into disjoint subsets $G_1$ and $G_2$, and define $G_{h,i} = G_h \cap G_i$. Suppose that it is easy to compute the transition probabilities $p^h(x,z)$ for $x \in$ some neighborhood of $G_1$, but harder for $x$ outside of that neighborhood. We suppose that it is feasible to run simulations of the process for any selected initial condition. For example, $p^h(x,z)$ might be given implicitly as the output of a complicated physical mechanism for which the transition probability is hard to compute when $x$ is in $G_2$ but which can be simulated. We try to exploit this situation by simulating where convenient and integrating where that is convenient. The division into subsets $G_i$ and the associated "difficulty" of some procedure in $G_i$ is meant to be suggestive. We will sometimes be "integrating" when in $G_2$ and sometimes simulating when in $G_1$. But, the major part of each type of computation will be done in the region where it is advantageous.

Let us divide the unit interval is divided into $n_h$ (an integer) subintervals, with $\epsilon = \delta_h k_h$. Thus, $1/\delta_h = k_h n_h$. For notational simplicity, we work with the interval $[0,1]$, but the method is identical for all the intervals $[n, n+1]$. Let $\mu_m^h$ denote the estimate of the distribution of $X_{mk_h}^h$. with values $\mu_m^h(x), x \in G_h$, where we start with $\mu_0^h = \Pi^h$. The time interval $\epsilon$ should be small enough such that the paths starting in $G_{h,i}$ stay close to it with a high probability on that interval. One needs to be careful if $\epsilon$ is allowed to go to zero as $h \to 0$ (which we do not do), since it is well known from simulations that the procedure can degenerate unless $v^h$ goes to infinity fast enough (and at a rate which depends on how fast $\epsilon \to 0$). In fact, there is little loss of generality or practicality in

fixing $\epsilon$ to be a small constant.

Suppose that $\mu_m^h$ is given. Then $\mu_{m+1}^h$ is computed as follows. First, do the analytic computation using $p^h(x,z)$ to get the part of $\mu_{m+1}^h(z)$ which is due to the initial states in $G_{h,1}$. Namely, compute

$$\sum_{x\in G_{h,1}} P\left\{X_{(m+1)k_h}^h = z \middle| X_{mk_h}^h = x\right\} \mu_m^h(x). \tag{5.14a}$$

The part of $\mu_{m+1}^h(z)$ which is due to initial states in $G_{h,2}$ is obtained by simulation. To simulate, we sample a total of $v^h$ points (where $v^h \to \infty$) in $G_{h,2}$ with relative probabilities $\{\mu_m^h(x), x \in G_{h,2}\}$. Denote the samples by $\{X_0^{h,l,m}, l \le v^h\}$. The sampling of the $\{X_0^{h,l,m}, l \le v^h\}$ can be done either with replacement or, preferably, using a variance reduction method such as the one based on stratified sampling which was described at the end of Example 2. For each of these "initial values" $X_0^{h,l,m}, l \le v^h$,, simulate at random a path of the chain for $k_h$ steps. Let $\{X_k^{h,l,m}, k \le k_0\}$ denote the sample values.

Then the part of the estimate of $\mu_{m+1}^h(z)$ which is due to initial states in $G_{h,2}$ is

$$\left[\sum_{x\in G_{h,2}} \mu_m^h(x)\right] \frac{1}{v^h} \sum_l I_{\{X_{k_h}^{h,l,m}=z\}}. \tag{5.14b}$$

The sum of (5.14a) and (5.14b) is $\mu_{m+1}^h(z)$. If $h \to 0$ and $\Pi^h$ converges weakly to, say, $\Pi$, it follows from the consistency condition (A4.1) for the $\tilde{X}^h(\cdot)$ process that $\mu_{n_h}^h(\cdot)$ converges weakly to the distribution of $\tilde{X}(1)$, which corresponds to $\tilde{X}(0)$ having distribution $\Pi$.

To identify the measure $P_{\Pi^h(n-1)}^{h,n}$ in (A5.3) (for our example $n = 1$) note the following. It is the measure on $G_h \times G_h$, with initial distribution given by our combination of $\mu_0^h(x)$ for $x \in G_{h,1}$ and the sampling distribution for $x \in G_{h,2}$. The distribution of the terminal value, conditioned on the initial distribution, is computed by repeating the updating procedure outlined above $1/\epsilon$ times.

We note that variance reduction methods can be employed in the sampling of the random paths themselves.

# 6 More Examples: Importance Sampling Methods for Discrete Time Models

Importance sampling methods are in common use to improve the performance of Monte Carlo algorithms (see, for example [9, 10]). They have also been used to improve the quality of nonlinear filtering algorithms that use random sampling [3, 28]. When used over an infinite time interval, the robustness and convergence questions raised earlier in the paper remain important. In the next example, we discuss the general idea of importance sampling and show how the associated proof of convergence is covered by what has already been done for

setups such as that in Example 1 of Section 5. In this next example, we describe the importance sampling on a typical interval $[n-1, n]$ and it does not use the next observation $\delta Y_n$. This next observation can provide useful information to guide the sampling. There are many intriguing possibilities, and one form of such a use is discussed in Example 7. We only illustrate some possibilities. There are numerous possible variations, and the choice of the better ones is still a matter of research. With all of the variations, variance reduction methods can be used, as can combined sampling–integration methods.

**Example 6: The basic idea of importance sampling.** Return to the setup used in Example 1 of Section 5. 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 $\tilde{X}^h(0)$. For each $h$ and $n$, let $M^{h,n}$ denote a random measure which is a.s. mutually absolutely continuous with respect to $P^h_{n-1}$. For each $h$ and $n$, let $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ be mutually conditionally independent, conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$, and with the distribution $M^{h,n}$. Define the likelihood ratio (the Radon–Nikodym derivative) and its value on the random path $\tilde{X}^{h,k,n}(\cdot)$ by

$$L^{h,n} = \frac{dP^h_{n-1}}{dM^{h,n}}, \quad L^{h,k,n} = \frac{dP^h_{n-1}}{dM^{h,n}}\left(\tilde{X}^{h,k,n}(\cdot)\right). \tag{6.1}$$

We introduce the following assumption.

**A6.1.**

$$\sup_{h,n} E\frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^h(1))R^2(\tilde{X}^h(1), \delta Y_n) < \infty, \tag{6.2}$$

where $\tilde{X}^h(\cdot)$ in (6.2) 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:

$$\langle \Pi^h(n), \phi\rangle = \frac{\sum_{l=1}^{v^h} L^{h,k,n}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} L^{h,k,n}R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}, \tag{6.3}$$

**Theorem 6.1.** *Assume* (A4.1), (A4.2) *and* (A6.1) *and the filter form* (6.3). *Then Theorem* 5.1 *holds.*

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

$$E\left[\frac{1}{v^h}\sum_k \left(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1), \delta Y_n) - E_{\Pi^h(n-1), \delta Y_n}\Phi(\tilde{X}^h(1), \delta Y_n)\right)\right]^2 \tag{6.4}$$

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

We can write

$$E_{M^{h,n},P^h_{n-1},\delta Y_n} L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) = \\ E_{M^{h,n},P^h_{n-1},\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n) = E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n), \tag{6.5}$$

where as before the subscript in the expectation is the conditioning data and $\tilde{X}^h(\cdot)$ in the second and the third expression has the law (conditioned on $M^{h,n}, P^h_{n-1}, \delta Y_n)$ $P^h_{n-1}$. The first equality 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.

Under $M^{h,n}$, the samples are mutually independent, conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$. By the above facts, in (6.4) can be rewritten as

$$EE_{M^{h,n},P^h_{n-1},\delta Y_n}\left[\frac{1}{v^h}\sum_k \left(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n). - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)\right)\right] \\ = \frac{1}{(v^h)^2}EE_{M^{h,n},P^h_{n-1},\delta Y_n}\sum_k\left[L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n). - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)\right]^2. \tag{6.6}$$

The right hand side of (6.6) is

$$O(1)\frac{1}{v^h}EE_{M^{h,n},P^h_{n-1},\delta Y_n}\left[L^{h,n}R^2(\tilde{X}^h(1),\delta Y_n)\right].$$

The last term, in turn can be bounded by

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

where $\tilde{X}^h(\cdot)$ is as in (A6.1). The above expression is easily seen to be $O(1/v^h)$ by (6.2). ■

**An extension: Dropping the mutual independence.** As in example 2 of Section 5, where we relaxed the condition on mutual independence of samples by instead assuming (A5.2), we can formulate a similar condition here which can be used to incorporate variance reduction methods along with importance sampling. More precisely, let $M^{h,n}$ be as before. Also let $\{\tilde{X}^{h,l,n}(\cdot); l \le v^h\}$ be as before, except that they need not be (conditionally) mutually independent. Instead of assuming (A6.1), assume (A6.2) below.

**A6.2.** Let $\Phi(x,y) = \phi(x)R(x,y)$, where $\phi(\cdot)$ is a bounded and continuous real valued function. Then

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

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

The following extension of Theorem 6.1 can now be stated.

**Theorem 6.2.** *Theorem 5.1 holds for the filter defined by* (6.3) *under* (A4.1), (A4.2) *and* (A6.2).

Of special interest is the case where the importance sampling is with respect to the initial condition only. To illustrate this case, we consider the special case of Example 4 of Section 5, where the sampling filter without importance sampling is given by (5.1) with $\tilde{X}^{h,n,l}(\cdot)$ defined via (5.13b). 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

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

In the next example, we see that the idea of applying importance sampling to the initial condition can be enhanced by the use of the next observation $\delta Y_n$ to determine the $M^{h,n}_0$.

**Example 7: Observation dependent importance sampling.** Appropriate measure transformations $M^{h,n}$ (or $M^{h,n}_0$ as defined at the end of the above example) can improve the estimates quite a bit [3, 28]. Better $M^{h,n}$ will depend on the next observation $\delta Y_n$ and we will illustrate the point via the signal model of Example 4 of Section 5, where $\tilde{X}^h(\cdot)$ is defined by (5.13). Data and examples of such a procedure can be found in [28]. Again, (A6.1) and the mutual absolute continuity are the only conditions (in addition to (A4.1) and (A4.2)) that need to be verified for Theorem 5.1 to hold for the $\Pi^h(\cdot)$ defined in this example. Keep in mind that we are illustrating only one type of procedure, and even that has many variations. Consider the following procedure.

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 $\delta Y_n$ 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. This "double" sampling explains the complexity of the following

algorithm. Nevertheless, such algorithms are sometimes 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 $\Pi^h(n)$. The details follow in algorithmic form.

**(1)** Let $\hat{X}^{h,l,n}(1)$ $(l \leq v^h)$ denote an "indicator" quantity, which (hopefully) is highly correlated with the "value" of sampling with initial condition $x^{h,l,n}$. [The points for which we get such an "indicator" quantity might also be chosen by some sampling procedure.]

**(2)** Compute the conditional Gaussian density $p(\delta Y_n | X(n) = \hat{X}^{h,l,n}(1))$, and define the "conditional likelihood" of the observation

$$p^{h,l,n} = \frac{p(\delta Y_n | X(n) = \hat{X}^{h,l,n}(1))}{\sum_k \Pi_k^h(n-1) p(\delta Y_n | X(n) = \hat{X}^{h,k,n}(1))}, \;\; l \leq v_h. \tag{6.7}$$

The numerator of $p^{h,l,n}$ up to a normalizing factor is $R(\hat{X}^{h,l,n}(1), \delta Y_n)$. Note that $p^{h,l,n}$ is not a probability. If the numerator in (6.7) is the same for all points, then $p^{h,l,n} = 1$ for all $l$.

**(3)** Sample $m^{h,n} \geq v^h$ times (with replacement) from the set

$$\{x^{h,l,n}, l \leq v^h\}$$

with weights proportional to the $p^{h,l,n}\Pi_l^h(n-1)$. Note that $\sum_l p^{h,l,n}\Pi_l^h(n-1) = 1$. This yields a set which we denote by

$$\{\overline{x}^{h,l,n}, l \leq m^{h,n}\}.$$

It is found in practice that the performance is often better if $m^{h,n}$ is several times $v^h$. This tends to assure a better spread for the support of the conditional distribution.

**(4)** Sample $\{\zeta^{h,k,n}, k \leq m^{h,n}\}$ from $P_\zeta^h$ and compute

$$\overline{X}^{h,l,n}(1) = b_h(\overline{x}^{h,l,n}, \zeta^{h,k,n}), \;\; k \leq m^{h,n}. \tag{6.8}$$

**(5)** If $v^h = m^{h,n}$, then set $\tilde{X}^{h,l,n}(1) = \overline{X}^{h,l,n}(1) = x^{h,l,n+1}$. If $v^h < m^{h,n}$, then resample at random (with replacement) $v^h$ times from $\{\overline{X}^{h,l,n}(1), l \leq m^{h,n}\}$, to get the set $\{\tilde{X}^{h,k,n}(1), k \leq v^h\}$, and set $x^{h,k,n+1} = \tilde{X}^{h,k,n}(1)$.

In this procedure, the measure $M^{h,n}$ was defined by defining $M_0^{h,n}$ via the weight

$$M_{0,k}^{h,n} = p^{h,k,n}\Pi_l^h(n-1) \tag{6.9}$$

that it puts on the initial point $x^{h,l,n}$. Finally, we use the filter defined by (6.3) with

$$L^{h,l,n} = \frac{1}{p^{h,l,n}}. \tag{6.10}$$

The set of likelihood functions clearly satisfy (A6.1) and the corresponding measures satisfy the mutual absolute continuity requirement.

# 7 The Continuous Time Problem

In this section we will study the continuous time analogs of the various random sampling and combined random sampling-integration algorithms studied in Sections 5 and 6. Our basic filtering model will be that in Section 2. To fix ideas, we will begin by indicating the form of the approximating filter for the case where the random samples are mutually independent and identically distributed, analogously to what was done in (5.1). We will then consider a general form of the approximating filter which would cover not only the case of such i.i.d. samples but also various variance reduction schemes and importance sampling algorithms of the type studied in Examples 2, 6 and 7.

## 7.1 Example and motivation.

**On the approximation (2.7), (2.8).** In typical uses of the approximation (2.7), the approximating signal process $\tilde{X}^h(\cdot)$ is a piecewise constant interpolation of a discrete time process. One good example is the Markov chain approximation such as used in Example 5. Most current applications seem to use such Markov chain based approximations, whether they are of the explicit forms discussed in [23] or other forms which satisfy the required local consistency property; e.g., based on approximate solutions to the Fokker-Planck equation over small intervals.

Following the idea and terminology of Example 5, let $X_n^h$ denote the underlying Markov chain, and $\tilde{X}^h(\cdot)$ its piecewise constant interpolation, with interpolation interval $\delta_h$. Then one can use the approximating filter (2.7). Since the approximating process $\tilde{X}^h(\cdot)$ there is piecewise constant, $R(\tilde{X}_{0,t}^h, Y_{0,t})$ equals

$$\exp \sum_{k=0}^{[t/\delta_h]-1} \left[ g'(X_k^h)\left[Y(k\delta_h+\delta_h) - Y(k\delta_h)\right] - \frac{\delta_h}{2}\left|g(X_k^h)\right|^2 \right] \\ \times \exp\left[ g'(X_{[t/\delta_h]}^h)\left[Y(t) - Y([t/\delta_h]\delta_h)\right] - \frac{t-[t/\delta_h]\delta_h}{2}\left|g(X_{[t/\delta_h]}^h)\right|^2 \right]. \tag{7.1}$$

However, in practice observations cannot truly be taken continuously and one would incorporate the observation into the filter at the discrete time instants $n\delta_h$ only. In fact, for such nonlinear problems the notion of continuous updating seems to be a mathematical fiction, although the times between updating might be very small. Thus, one would approximate $R(\tilde{X}^h_{0,t}, Y_{0,t})$ by $R^h(\tilde{X}^h_{0,t}, Y_{0,t})$ which is defined by the following (piecewise constant) expression

$$\exp \sum_{k=0}^{[t/\delta_h]-1} \left[ g'(X^h_k) \left[ Y(k\delta_h + \delta_h) - Y(k\delta_h) \right] - \frac{\delta_h}{2} \left| g(X^h_k) \right|^2 \right]. \tag{7.2}$$

Whatever the form of $\tilde{X}^h(\cdot)$, whether it is obtained explicitly as an interpolation of a discrete time chain $X^h_n$ or not, the samples $\tilde{X}^h(n\delta_h)$ are always a Markov chain. For notational simplicity, we always write $\tilde{X}^h(n\delta_h) = X^h_n$.

**A random sampling algorithm.** The above paragraph argues that it is not a restriction to require the approximating filter process to be piecewise constant. This logic also holds for algorithms based on random sampling. It holds even if some higher order interpolation method (say of the Milstein or other types used in [16]) are used, since even then we would use the interpolation to get a better approximation to the signal process at discrete time instants $n\delta_h$. Thus, in the approximate filters that are considered here, we approximate the conditional distribution at the instants $n\delta_h$, and suppose that the filter is constant on $[n\delta_h, n\delta_h + \delta_h)$.

Since the estimate of the conditional distribution will be updated at each $n\delta_h$, we could try to duplicate the various methods in Examples 1 to 7, with the basic interval being $\delta_h$, and then prove convergence as $\delta_h \to 0$. The resampling at the beginning of each interval in the various examples exploited the updated information to get more sample trajectories from the points which seemed to be more likely, in view of the information in the observations. But random sampling also loses information. There is always a chance that the better points will not be sampled. This chance is reduced as $v^h$ increases. When resampling occurs very frequently, say at each time instant $n\delta_h$, the procedure can degenerate very fast as $\delta_h \to 0$, unless $v^h$ increases fast enough as $\delta_h \to 0$. One can quantify such a statement. But it is also a common observation in simulations, including the ones that we have carried out. The reference [4] resamples at each discrete interval, in a "minimum variance" way, but $v^h$ must grow as $1/\delta_h$. Such a rapid increase in $v^h$ is an inefficient use of the computational resources, especially in view of the fact that the estimates of the conditional distribution do not change much in small intervals.

Owing to the above observations, we take the following "practical" approach. Divide time into subunits of (small, but fixed–they do not go to zero) length $\epsilon$, and suppose that $\epsilon/\delta_h = n_h$ is always an integer. We resimulate the $\tilde{X}^h(\cdot)$ each $\epsilon$ units of time, although the observations are incorporated at the instants $n\delta$.

The general model given below is motivated by the ideas of Examples 3 to 7, and we give an analog of condition (A5.3) (namely, conditions (A7.1), (A7.2)) which covers many cases of interest. But, for specificity, let us first

consider the case where the process $\tilde{X}^h(\cdot)$ satisfies the consistency assumption (A2.1) and the samples taken on $[n\epsilon, n\epsilon+\epsilon)$ are mutually independent and independent of $Y(\cdot)$ given their initial distribution $\Pi^h(n\epsilon)$. Denote these samples by $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$. Define

$$R^h(\tilde{X}_{0,s}, Y_{n\epsilon,n\epsilon+s}) = \exp \sum_{k=0}^{[s/\delta_h]-1} \left[ g'(X_k^h)\left[Y(n\epsilon+k\delta_h+\delta_h) - Y(n\epsilon+k\delta_h)\right] - \frac{\delta_h}{2}\left|g(X_k^h)\right|^2 \right]. \tag{7.3}$$

Thus $R^h(x_{0,s}, Y_{a,b})$ differs from $R(x_{0,s}, Y_{a,b})$ only in that in the former the function $x(\cdot)$ is replaced by the piecewise constant function with value $x(k\delta_h)$ on $[k\delta_h, k\delta_h+\delta_h)$. The basic approximating filter based on random sampling, for $s = q\delta_h < \epsilon$ where $q$ is an integer, is

$$\langle \Pi^h(n\epsilon+s), \phi\rangle = \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(s)) R^h(\tilde{X}^{h,l,n}_{0,s}, Y_{n\epsilon,n\epsilon+s})/v^h}{\sum_{l=1}^{v^h} R^h(\tilde{X}^{h,l,n}_{0,s}, Y_{n\epsilon,n\epsilon+s})/v^h}. \tag{7.4}$$

This is just the "continuous time" analog of (5.1).

As stated earlier, we will suppose that $\Pi^h(\cdot)$ is constant on the intervals $[q\delta_h, q\delta_h+\delta_h)$. Alternatively, if desired, we can interpolate and one natural interpolation would use the form (7.4), but with the term

$$g'(X^h_{[s/\delta_h]})\left[Y(n\epsilon+s) - Y(n\epsilon+[s/\delta_h]\delta_h))\right] - \frac{s-[s/\delta_h]\delta_h}{2}\left|g(X^h_{[s/\delta_h]})\right|^2 \tag{7.5}$$

added to the sum in (7.3). The treatment of both forms is nearly identical.

As in the discrete time case, we are interested in random sampling and random sampling-integration algorithms which are more general than (7.4), with (conditional) i.i.d. samples. We would like to allow the possibility of incorporating variance reduction methods as in Example 2, or perhaps the sampling can be guided using some importance sampling scheme as in Examples 6 and 7. We might even be interested in algorithms which are part integration and part random sampling, say of the form discussed in Examples 4 and 5. In view of this, we work with the following general form of the approximate filter, which is our continuous time analog of the discrete time filter defined via (5.10). The chosen general structure is motivated by the same considerations which led to (5.10) and (A5.11), the desire to include many types of approximations of interest under one roof, with a general assumption which can be verified in particular cases of interest, analogously to what was done for the discrete time case. The conditions (A7.1) and (A7.2) hold for the independent samples case under (A2.1).

Let $\delta_h, \epsilon, n_h$ be as above. Define the approximating filter $\Pi^h(\cdot)$ as follows. Having defined $\Pi^h(t)$ for $0 \le t \le n\epsilon$, let $P^{h,n}_{\Pi^h(n\epsilon)}$ is conditionally independent of $\{Y_t - Y_{n\epsilon}; t \ge n\epsilon\}$ given $\Pi^h(n\epsilon)$. Define, for $1 \le j \le n_h$,

$$\langle \Pi^h(n\epsilon+j\delta_h), \phi\rangle = \frac{\int \phi(x(j\delta_h)) R^h(x_{0,j\delta_h}, Y_{n\epsilon,n\epsilon+j\delta_h}) dP^{h,n}_{\Pi^h(n\epsilon)}(x(\cdot))}{\int R^h(x_{0,j\delta_h}, Y_{n\epsilon,n\epsilon+j\delta_h}) dP^{h,n}_{\Pi^h(n\epsilon)}(x(\cdot))}. \tag{7.6}$$

For points in $[n\epsilon, (n+1)\epsilon)$ not of the form $n\epsilon + j\delta_h$, the filter is defined via the piecewise constant and right continuous interpolation. In the independent sample case (7.4), $P^{h,n}_{\Pi^h(n\epsilon)}$ is the occupation measure defined analogously to the the way it was defined above (5.9) for the discrete time case. We assume that the family $\{P^{h,n}_{\Pi^h(n\epsilon)}\}$ satisfies (A7.1) and (A7.2) below. Given the very general structure allowed for the filter, some condition such as (A7.2) is needed for the continuous time problem in order to avoid simulated processes that are "wild." Under (A2.1), Condition (A7.2) holds for the i.i.d case illustrated in (7.4), since there each $\tilde{X}^{h,l,n}(\cdot)$ is a replica of the $\tilde{X}^h(\cdot)$ of (A2.1) with initial conditions in the compact set $G$.

**A7.1.** For every bounded and continuous real valued function $\Phi(\cdot)$ of $x(\cdot)$ on the interval $[0, \epsilon]$ and which depends on $x(\cdot)$ only at a finite number of points

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

We will impose another condition on the $P^{h,n}_{\Pi^h(n\epsilon)}$. For motivation, consider the case of independent samples discussed above. For $\mu > 0, \delta > 0,$, define the set of paths

$$C^\delta_\mu = \left\{x(\cdot) : \sup_{s\le\delta, t+s\le\epsilon, t\le\epsilon} |x(t+s) - x(t)| \ge \mu\right\}.$$

Then

$$\lim_{\delta\to 0} \limsup_h \sup_n E P^{h,n}_{\Pi^h(n\epsilon)}(C^\delta_\mu) = 0 \quad \text{for each } \mu > 0. \tag{7.7}$$

The limit (7.7) continues to hold for the model of Example 5 for our current case where the observations are incorporated at each time $n\delta_h$. It holds if the samplings of the initial conditions $\tilde{X}^{h,l,n}(0)$ are determined by importance sampling as in Example 7. It also holds for many variance reduction methods. For example, stratified sampling of the initial condition or of the intermediate noises. The expression (7.7) simply states that for small $h$, the sampled paths don't change much in the mean over small intervals, uniformly in $n$. We will impose this reasonable property as a requirement by assuming:

**A7.2.** The condition (7.7) holds.

The following lemma will be used in some of the tightness arguments used below.

**Lemma.** [18, Theorem 2.7b]. *Let $\{Z_n(\cdot), n\}$ be a family of processes with paths in the Skorohod space $D[S_0; 0, \infty)$, where $S_0$ is a complete and separable metric space with metric $\gamma(\cdot)$. For each $\delta > 0$ and each $t$ in a dense set, let there be a compact set $S_{\delta,t} \subset S_0$ such that*

$$\sup_n P\{Z_n(t) \notin S_{\delta,t}\} \le \delta.$$

*Let $\mathcal{F}_t^n$ denote the minimal $\sigma-$algebra which measures $\{Z_n(u), u \leq t\}$, and $\mathcal{T}_n(T)$ the set of $\mathcal{F}_t^n$-stopping times which are less than $T > 0$. Suppose that for each $T$*

$$\lim_{\delta \to 0} \limsup_n \sup_{\tau \in \mathcal{T}_n(T)} E\left[\gamma\left(Z_n(\tau+\delta), Z_n(\tau)\right) \wedge 1\right] = 0.$$

*Then $\{Z^n(\cdot)\}$ is tight.*

**Theorem 7.1.** *Let $(X(\cdot), Y(\cdot))$ be as in Section 2. Assume* (A3.1), (A7.1) *and* (A7.2). *Then the conclusions of Theorem* 3.1 *hold for the approximate filter $\Pi^h(\cdot)$ defined as above.*

**Proof.** Many of the details of the proof are the same as in the proof of Theorem 3.1, and in the way that its was used in Theorem 5.1 and its successors for the discrete time case. The key differences are in the proof of tightness of $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ for any sequences $h_k \to 0, T_k \to \infty$, and the proof of the representation (3.8) and we concentrate on these points. In the discrete time theorems, the tightness of $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ was essentially obvious due to the compactness of $G$. There was no issue of "path properties," in showing the tightness due to the discrete time parameter. In the current continuous time case, we need to deal with the path properties. Owing to the use of the weak topology, it is enough to prove the tightness of the set

$$\left\{\langle \Pi^{h_k}(t_k + \cdot), \phi\rangle; h_k, T_k\right\}$$

for each bounded and continuous real valued function $\phi(\cdot)$ on $G$.

In proving the tightness of the above family, we use the criterion in the lemma. The main step is establishing that for each $\phi(\cdot)$ as above

$$\lim_{\delta \to 0} \limsup_{h \to 0} \sup_t \sup_{\tau \in \mathcal{T}^{h,t}(\rho)} E\left|\langle \Pi^h(t+\tau+\delta), \phi\rangle - \langle \Pi^h(t+\tau), \phi\rangle\right|_1^2 = 0, \tag{7.8}$$

where $\mathcal{T}^{h,t}(\rho)$ denotes the set of stopping times bounded by $\rho$, for the process $\Pi^h(t+\cdot)$. We can assume without loss of generality that $\delta$ in the above expression is less than $\epsilon$. Thus $t+\tau$ and $t+\tau+\delta$ are either in the same interval of the form $[j\epsilon, (j+1)\epsilon)$ or they are in adjacent such intervals. This observation along with an application of a triangle inequality shows that, in order to prove (7.8) it suffices to prove that

$$\limsup_{h \to 0} \sup_j \sup_{0 \leq K_1 \leq K_2 \leq n_h, |K_1 - K_2|\delta_h \leq \delta} E\left|\langle \Pi^h(j\epsilon + K_1\delta_h), \phi\rangle - \langle \Pi^h(j\epsilon + K_2\delta_h), \phi\rangle\right|_1^2 \tag{7.9}$$

converges to 0 as $\delta \to 0$. Note that $j\epsilon + K_1\delta_h$ and $j\epsilon + K_2\delta_h$ are both in the interval $[j\epsilon, j\epsilon + \epsilon]$.

Now we bound the above expectation by the sum of the following three terms. The first two terms are, for $i = 1, 2$,

$$\limsup_{h \to 0} \sup_j \sup_{0 \leq K_i\delta_h \leq \epsilon} F_1(j, h, K_i\delta_h), \tag{7.10}$$

where

$$F_1(j,h,K_i\delta_h) = E\left|\langle \Pi^h(j\epsilon + K_i\delta_h),\phi\rangle - \frac{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_i\delta_h}}\phi(\tilde{X}(K_i\delta_h))R^h(\tilde{X}_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})}{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_i\delta_h}}R^h(\tilde{X}_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})}\right|_1^2.$$

The third term is

$$\limsup_{h\to 0}\sup_j \sup_{0\le K_1\delta_h\le K_2\delta_h\le\epsilon,|K_1-K_2|\delta_h\le\delta} F_2(j,h,K_1\delta_h,K_1\delta_h), \tag{7.11}$$

where

$$F_2(j,h,K_1\delta_h,K_1\delta_h) = E\left|\frac{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_1\delta_h}}\phi(\tilde{X}(K_1\delta_h))R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h})}{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_1\delta_h}}R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h})} - \frac{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_2\delta_h}}\phi(\tilde{X}(K_2\delta_h))R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})}{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_2\delta_h}}R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})}\right|_1^2.$$

In dealing with (7.11), owing to the properties of the $|\cdot|_1^2$ metric defined by (4.14), we need only work with the differences of the numerators and denominators separately. Then, (7.11) is easily dealt with using the continuity property of $\tilde{X}(\cdot)$. In particular, it follows from the fact that for any bounded and continuous function $\phi(\cdot)$

$$\lim_{\delta\to 0}\sup_{|K_1-K_2|\delta_h\le\delta}\sup_\pi E\Big|E_{\pi,Y_{j\epsilon,j\epsilon+K_1\delta_h}}\phi(\tilde{X}(K_1\delta_h)R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h}) - E_{\pi,Y_{j\epsilon,j\epsilon+K_2\delta_h}}\phi(\tilde{X}(K_2\delta_h))R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})\Big|_1^2 = 0, \tag{7.12}$$

where $K_i\delta_h \le \epsilon$.

Now we consider (7.10). By the definition of $\Pi^h(j\epsilon)$ in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$ in (7.6), the first term inside the bars in (7.10), equals

$$\frac{\int \phi(x(K_i\delta_h))R^h(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}{\int R^h(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}. \tag{7.13}$$

Again, we need only work with the differences of the numerators of the right hand term inside the bars in (7.10) and that in (7.13), for arbitrary bounded and continuous $\phi(\cdot)$.

The proof that the limit of (7.10) as $\delta \to 0$ is zero will use an approximation method. For small $\Delta > 0$, with $\epsilon$ an integral multiple of $\Delta$, define $R^\Delta(x_{0,s},Y_{a,a+s})$ by

$$\exp\{\sum_{i:i\Delta<s} g'(x(i\Delta))\left[Y(a+i\Delta+\Delta)-Y(a+i\Delta)\right] - \frac{\Delta}{2}\sum_{i:i\Delta<s}|g(x(i\Delta))|^2.\}$$

For each $h$, define

$$A^{h,\Delta} = \sup_{K_i\delta_h \le \epsilon} \sup_j \sup_\pi E_\pi \left| R^h(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) - R^\Delta(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) \right|_1^2.$$

For each $\rho > 0$, there is $\Delta_0 > 0$ such that for $\Delta < \Delta_0$ and small $h > 0$ we have $A^{h,\Delta} \le \rho$. Define

$$B^{h,\Delta} = \sup_{K_i\delta_h \le \epsilon} \sup_j E \int \left| R^h(x_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) - R^\Delta(x_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) \right|_1^2 dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot)). \tag{7.14}$$

By (A7.2), for each $\rho > 0$ there is $\Delta_0 > 0$ such that for $\Delta < \Delta_0$, and small $h > 0$, $B^{h,\Delta} \le \rho$. This assertion is proved as follows. Define $x^\Delta(t) = x(i\Delta)$ for $t \in [i\Delta, i\Delta + \Delta)$. To prove the assertion, it is sufficient to show that

$$\limsup_h \sup_j \sup_{K_i \le n_h} \int \left| \sum_{l=0}^{K_i} \left[ g(x(l\delta_h)) - g(x^\Delta(l\delta_h)) \right]' \left[ Y(l\delta_h + d_h) - Y(l\delta_h) \right] \right|^2 d\left[ EP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot)) \right] \tag{7.15}$$

is arbitrarily small if $\Delta$ is small enough. By (A7.2), we can suppose that the difference of the $g-$terms in the bracket in (7.15) is as small as we wish and this implies the assertion.

Thus, to show that the limit of (7.10) is zero as $\delta \to 0$, it is sufficient to show that

$$\begin{aligned} &\lim_{\Delta\to 0} \limsup_{h\to 0} \sup_j \sup_{0 \le K_i\delta_h \le \epsilon} \\ &\quad E \Big| \int \phi(x(K_i\delta_h)) R^\Delta(x_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot)) \\ &\qquad - E_{\Pi^h(j\epsilon), Y_{j\epsilon,j\epsilon+K_i\delta_h}} \phi(\tilde{X}(K_i\delta_h)) R^\Delta(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h}) \Big|_1^2 = 0 \end{aligned} \tag{7.16}$$

By (A7.2), is is sufficient to show (7.16) if the $K_i\delta_h$ in the $\phi(x(K_i\delta_h))$ and $\phi(\tilde{X}(K_i\delta_h))$ are replaced by the closest integral multiple of $\Delta$ for small $\Delta$. Now, all the $K_i\delta_h$ in (7.16) are integral multiples of $\Delta$ for some fixed $\Delta$. Hence, the $\sup_{0\le K_i\delta_h\le\epsilon}$ in (7.16) is redundant and can be dropped. We would like to use (A7.1) at this point. But (A7.1) holds only for each function $\Phi(\cdot)$. In (7.16), $\Delta$ is fixed, and we can suppose, without loss of generality that $\epsilon = k_0\Delta$ for some integer $k_0$. Given any small $\rho > 0$, there is a bounded set $B_\rho$ such that the values of $\{Y(j\epsilon + t) - Y(j\epsilon), t \le \epsilon\}$ are confined to $B_\rho$ with at least probability $1 - \rho$ for all $j$. Because of this, and the continuity of $R^\Delta(\cdot)$, we need only verify (7.16) for some finite set of values of the $Y-$variables. Due to this and to the fact that $k_0 < \infty$, we need only evaluate (7.16) for each $Y(\cdot)$ and $K_i\delta_h \le \epsilon$ being some arbitrary multiple of $\Delta$. Then (A7.1) can be used, and guarantees (7.16). This completes the proof of tightness of $\{Q^{h_k,T_k}(\cdot), k \ge 1\}$.

We now prove the representation in (3.8). The general scheme used in Theorem 5.1 for this characterization will be followed. Let $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot))$ denote the (canonical paths of the signal process, the conditional probability process, the observation process and the observation noise process). They are connected by $y(t) = \int_0^t g'(x(s))ds + b(t)$. For arbitrary bounded and continuous $\phi(\cdot)$, arbitrary $\psi(\cdot)$ and times $t, s$, define the function $A(\psi(\cdot); t, s))$ by

$$A(\psi(\cdot); t, s) = \langle \pi(t+s), \phi \rangle - \frac{E_{\{\pi(t), y_{t,t+s}\}} \left[ \phi(\tilde{X}_{0,s}) R(\tilde{X}_{0,s}, y_{t,t+s}) \right]}{E_{\{\pi(t), y_{t,t+s}\}} R(\tilde{X}_{0,s}, y_{t,t+s})}. \quad (7.17)$$

Recall the definition of $\Psi^\omega(\cdot)$ from Theorem 3.1. We will also use other notations from Section 3. The aim of the proof of Theorem 3.2 in [2], which is our Theorem 3.1, was to show that, for almost all $\omega$ and all $t, s$,

$$A(\Psi^\omega(\cdot); t, s) = 0, \text{ with probability 1} \quad (7.18)$$

which implies (3.8). In fact it suffices to consider $s \leq \epsilon$. Hereafter we will consider only such values of $s$ without any further comment.

The statement in (7.18) will be proved by showing that

$$0 = E \int Q^\omega(d\psi) \left[ A(\psi(\cdot); t, s) \right]_1^2, \quad (7.19)$$

The prelimit form of the right side of (7.19) is

$$E \int Q^{h,T}(d\psi) \left[ A(\psi(\cdot); t, s) \right]_1^2,$$

which, by the definition of $Q^{h,T}(\cdot)$, equals

$$\frac{1}{T} \int_0^T E \left[ A(\Psi^h(\cdot); u+t, s) \right]_1^2 du, \quad (7.20)$$

where

$$A(\Psi^h(\cdot); t, s) = \langle \Pi^h(t+s), \phi \rangle - \frac{E_{\{\Pi^h(t), Y_{t,t+s}\}} \left[ \phi(\tilde{X}(s)) R(\tilde{X}_{0,s}, Y_{t,t+s} \right]}{E_{\{\Pi^h(t), Y_{t,t+s}\}} R(\tilde{X}_{0,s}, Y_{t,t+s})}. \quad (7.21)$$

In order to show (7.19), it suffices to show that

$$\lim_h \sup_t E[A(\Psi^h(\cdot); t, s)]_1^2 \to 0. \quad (7.22)$$

Furthermore, in view of the properties of the $\tilde{X}(\cdot)$ process and the tightness of the set $\{\Pi^h(t); h, t\}$, to prove (7.22) it is clearly sufficient to show that

$$\sup_t E \left| \langle \Pi^h(t+s), \phi \rangle - \frac{E_{\{\Pi^h(t), Y_{t,t+s}\}} \left[ \phi(\tilde{X}(s)) R^h(\tilde{X}_{0,s}, Y_{t,t+s}) \right]}{E_{\{\Pi^h(t), Y_{t,t+s}\}} R^h(\tilde{X}_{0,s}, Y_{t,t+s})} \right|_1^2 \to 0 \quad (7.23)$$

as $h \to 0$.

Since $s < \epsilon$, the points $t$ and $t+s$ are either in the same subinterval of the form $[j\epsilon, (j+1)\epsilon]$ or are in adjacent intervals of this form. We consider below the case of adjacent intervals. The arguments required for the same interval case are simpler versions of the former case and thus are omitted. Let now $t \in [j\epsilon + i\delta_h, j\epsilon + i\delta_h + \delta_h)$ and $t+s \in [(j+1)\epsilon + i'\delta_h, (j+1)\epsilon + i'\delta_h + \delta_h)$. Showing (7.23) is equivalent to proving that, for each $s$

$$E\left|\langle \Pi^h(t+s), \phi\rangle - \frac{E_{\{\Pi^h(j\epsilon+i\delta_h), Y_{j\epsilon+i\delta_h, t+s}\}}\left[\phi(\tilde{X}(\alpha\delta_h))R^h(\tilde{X}_{0,\alpha\delta_h}, Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h(j\epsilon+i\delta_h), Y_{j\epsilon+i\delta_h, t+s}\}}R^h(\tilde{X}_{0,\alpha\delta_h}, Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})}\right|_1^2 \tag{7.24}$$

converges to 0 as $h \to 0$, uniformly in $t$, where $\alpha = n_h + i' - i$. Thus, $|\alpha\delta_h - s| \leq \delta_h$.

The expectation in (7.24) can be bounded above by the sum of

$$E\left|\langle \Pi^h(t+s), \phi\rangle - \frac{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon, t+s}\}}\left[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon, t+s}\}}R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})}\right|_1^2 \tag{7.25}$$

and

$$E\left|\frac{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon, t+s}\}}\left[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon, t+s}\}}R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})} - \frac{E_{\{\Pi^h(j\epsilon+i\delta_h), Y_{j\epsilon+i\delta_h, t+s}\}}\left[\phi(\tilde{X}(\alpha\delta_h))R^h(\tilde{X}_{0,\alpha\delta_h}, Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h(j\epsilon+i\delta_h), Y_{j\epsilon+i\delta_h, t+s}\}}R^h(\tilde{X}_{0,\alpha\delta_h}, Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})}\right|_1^2. \tag{7.26}$$

Using the definition from (7.6) of $\Pi^h(t+s)$ in terms of $P^{h,j+1}_{\Pi^h((j+1)\epsilon)}$ in (7.25) and working with numerators and denominators separately, as we may, it follows that showing the convergence to zero of the $\sup_t$ of (7.25) as $h \to 0$ to zero is equivalent to showing the same for

$$E\left|\int \phi(x(i'\delta_h))R^h(x_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})dP^{h,j+1}_{\Pi^h((j+1)\epsilon)}(x(\cdot)) - E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon, t+s}\}}\left[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})\right]\right|_1^2 \tag{7.27}$$

Let $\Delta > 0$ be small. Now, repeat the logic which led to (7.16). By (A7.2) and the continuity properties of $\tilde{X}(\cdot)$, it is sufficient to prove the result if both of the $R^h(\cdot)$ in (7.27) are replaced by $R^\Delta(\cdot)$, and the $i'\delta_h$ in $\tilde{X}(i'\delta_h)$ and $x(i'\delta_h)$ are replaced by the nearest integral integral multiple of $\Delta$. Then (A7.1) yields the desired convergence. This takes care of (7.25).

We now turn to (7.26). This time we do not work with the numerators and denominators separately. For motivation, note that if $\Pi^h(\cdot)$ were the true

conditional distribution for the discrete time signal process $X(n\delta_h)$, then (7.26) is identically zero. Let $\Delta > 0$ be small and $\epsilon$ an integral multiple of $\Delta$. Owing to the properties of $\tilde{X}(\cdot)$ and the tightness of the set $\{\Pi^h(t+\cdot); h, t\}$ it is sufficient to show that the $\lim_h \sup_t$ of (7.26) is zero if $R^h(\cdot)$ were replaced by $R^\Delta(\cdot)$ and the $i\delta_h$ and $i'd_h$ were integral multiples of $\Delta$. Thus we can write $t = j\epsilon + k_1\Delta$ and $t+s = (j+1)\Delta + k_2\Delta$, where $k_i\Delta \le \epsilon$. Using the fact that the $k_i$ have only finitely many values, it is sufficient to show that

$$\lim_h \sup_j E \left| \frac{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\}} \left[\phi(\tilde{X}(k_2\Delta)) R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta})\right]}{E_{\{\Pi^h((j+1)\epsilon), Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\}} \left[R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta})\right]} \right.$$
$$\left. - \frac{E_{\{\Pi^h(j\epsilon+k_1\Delta), Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\}} \left[\phi(\tilde{X}(\epsilon - k_1\Delta + k_2\Delta)) R^\Delta(\tilde{X}_{0,\epsilon-k_1\Delta+k_2\Delta}, Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta})\right]}{E_{\{\Pi^h(j\epsilon+k_1\Delta), Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\}} \left[R^\Delta(\tilde{X}_{0,\epsilon-k_1\Delta+k_2\Delta}, Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta})\right]} \right|_1^2 = 0. \tag{7.28}$$

The difficulty in treating this term is that the initial times are different, being $(j+1)\epsilon$ in the first term and $j\epsilon + i\delta_h$ in the second. Because of this, we need to represent both initial measures in terms of the same quantity, namely in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$, and the details will now be given. Define the function

$$\Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x\right) =$$
$$E\left[\phi(\tilde{X}(k_2\Delta)) R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}) \middle| \tilde{X}(0) = x, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\right]. \tag{7.29}$$

If $\phi(\cdot)$ is equal to the constant function with value unity, we simply write 1 for $\phi$ in (7.29). Then, using the definition (7.6) of $\Pi^h((j+1)\epsilon)$ in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$, the numerator of the first term inside the bars in (7.28) can be rewritten as

$$\frac{\int \Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon}) dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}{\int R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon}) dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}, \tag{7.30}$$

The denominator of the left hand term inside the bars in (7.28) has the same representation, but with 1 replacing $\phi$. Thus, that left hand term can be written as

$$\frac{\int \Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon}) dP^{h,j}_{\Pi^h(j\epsilon)(}(x(\cdot))}{\int \Theta\left(1, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon}) dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}. \tag{7.31}$$

By (A7.1), without changing the limits in (7.28), this fraction can be replaced by

$$\frac{E_{\{\Pi^h(j\epsilon), Y_{j\epsilon,(j+1\epsilon)+k_2\Delta}\}} \Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})}{E_{\{\Pi^h(j\epsilon), Y_{j\epsilon,(j+1\epsilon)+k_2\Delta}\}} \Theta\left(1, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})}. \tag{7.32}$$

In turn, using the Markov property of $\tilde{X}(\cdot)$, the definition of $\Theta(\cdot)$ as a conditional expectation, and the fact that $R^{\Delta}(\cdot)$ is the exponential of a sum, this equals

$$\frac{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1)\epsilon)+k_2\Delta}\}}\phi(\tilde{X}(\epsilon+k_2\Delta))R^{\Delta}(\tilde{X}_{0,\epsilon+k_2\Delta},Y_{j\epsilon,(j+1)\epsilon+k_2\Delta})}{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1)\epsilon)+k_2\Delta}\}}R^{\Delta}(\tilde{X}_{0,\epsilon+k_2\Delta},Y_{j\epsilon,(j+1)\epsilon+k_2\Delta})} \tag{7.33}$$

Now, we turn our attention to the second term inside the bars in (7.28). This is treated in essentially the same way as was the first term. Consider the numerator of that term. The expectation, conditioned on

$$\left\{\tilde{X}(\epsilon-k_1\Delta)=x,\tilde{X}_{0,\epsilon-k_1\Delta},Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\right\},$$

is just

$$\Theta\left(\phi,k_2\Delta,Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta},x\right)R^{\Delta}(X_{0,\epsilon-k_1\Delta},Y_{j\epsilon+k_1\Delta,(j+1)\epsilon}).$$

We proceeding as we did above with the first term. Using the definition (7.6) yields an expression analogous to (7.29). Then applying first (A7.1) and then the Markov property of $\tilde{X}(\cdot)$ to that expression yields that we can replace the second term in (7.28) by (7.33) as well without changing the limit. We omit the details, which are nearly the same as for the first term. Thus, the term in the bars in (7.28) can be replaced by zero without changing the limit.

The proof of (7.22) is now completed. ■

# References


[1] P. Billingsley. *Convergence of Probability Measures.* John Wiley, New York, 1968.

[2] A. Budhiraja and H.J. Kushner. Approximation and limit results for non-linear filters over an infinite time interval. Lefschetz Center for Dynamical Systems Report. Submitted to SIAM J. on Control and Optmization, 1997.

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

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

[5] 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.

[6] P. Dupuis and H. Ishii. On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics Stochastics Rep.*, 35:31–62, 1991.

[7] R. Elliot. *Stochastic Calculus and Applications.* Springer-Verlag, Berlin and New York, 1982.

[8] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

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

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

[11] 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.

[12] 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.

[13] G. Kallianpur H. Fujisaki and H. Kunita. Stochastic differential equations for the nonlinear filtering problem. *Osaka Math. J.*, 9:19–40, 1972.

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

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

[16] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin and New York, 1992.

[17] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of a Markov process,. *J. Multivariate Anal.*, 1:365–393, 1971.

[18] T.G. Kurtz. *Approximation of Population Processes*, volume 36 of *CBMS-NSF Regional Conf. Series in Appl. Math.* SIAM, Philadelphia, 1981.

[19] H.J. Kushner. Dynamical equations for nonlinear filtering. *J. Differential Equations*, 3:179–190, 1967.

[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. s *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] R. Liptser and A.N. Shiryaev. *Statistics of Random Processes*. Springer-Verlag, Berlin and New York, 1977.

[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.