FAST PARTICLE FILTERS & THEIR APPLICATIONS TO RECURSIVE IDENTIFICATION & ADAPTIVE CONTROL OF NONLINEAR STOCHASTIC SYSTEMS

by

Yuguo Chen
Tze Leung Lai

August 2004

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
FAST PARTICLE FILTERS & THEIR APPLICATIONS TO RECURSIVE IDENTIFICATION & ADAPTIVE CONTROL OF NONLINEAR STOCHASTIC SYSTEMS

by

Yuguo Chen
Institute of Statistics & Decision Sciences
Duke University

Tze Leung Lai
Department of Statistics
Stanford University

August 2004

This research was supported in part by National Science Foundation grant DMS 0072523

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Fast Particle Filters and Their Applications to
Recursive Identification and Adaptive Control of Nonlinear Stochastic Systems

Yuguo Chen and Tze Leung Lai

Abstract

By proper choice of proposal distributions for importance sampling and of resampling schemes for sequentially updating the importance weights, this paper develops fast particle filters that can be implemented via parallel recursions for on-line identification and adaptive control of nonlinear stochastic systems. Theoretical analysis and simulation studies show the superiority of this new approach over conventional methods for identification and adaptive control of ARX models with occasional parameter jumps. Another nonlinear identification problem considered herein is combined system identification and state estimation in state space models, for which it is shown how particle filters can circumvent the complexities of the problem due to inherent nonlinearities.

Index Terms: Sequential Monte Carlo, importance sampling, resampling, adaptive control, ARX models with parameter jumps, nonlinear state space models.
I. INTRODUCTION

Following the seminal paper of Gordon et al. [1], the past decade has witnessed a surge of interest in sequential Monte Carlo methods and their applications to signal processing and nonlinear filters in hidden Markov models; see the Special Issue on Sequential Monte Carlo Methods in IEEE Trans. Signal Processing, Feb 2002. These methods represent the posterior distributions by a large number of random samples that are sequentially generated over time by using a combination of importance sampling and resampling steps (cf. [2]). Because of the large number of simulated trajectories (“particles”) that are employed in these particle filters, it appears that although these simulation-based methods provide substantial improvements over the extended Kalman filter and other linear approximations to nonlinear systems, they may not provide suitable alternatives to the widely used recursive algorithms for on-line identification and control that are based on linear approximations to nonlinear systems.

In this paper we show that, at least for certain classes of nonlinear systems, it is possible to develop fast particle filters that require a relatively small number (50-100) of simulated trajectories which can be implemented by parallel recursions. The basic idea lies in choosing (i) a statistically and computationally efficient proposal distribution for importance sampling and (ii) a resampling scheme that resamples only when needed (instead of at every stage as in [1]). This is described in Section II where, as an illustration of its statistical performance and computational speed, it is applied to a classical nonlinear filtering problem, namely, estimation of a noisy piecewise constant signal with unknown jump times and magnitude.

It is our earnest hope that the fast particle filters developed in Section II can offer theoretical advances and practical solutions to some long-standing problems concerning identification and adaptive control of nonlinear systems. We apply these particle filters to two such problems. In Section III, after a brief review of the literature on ARX models whose nonlinearities are caused by parameter jumps at unknown times and of unknown magnitude, we use fast particle filters to approximate the Bayes estimates of the time varying parameters in open-loop designs and in feedback control systems. Theoretical analysis and simulation studies show the superiority of the parameter estimates and adaptive control schemes using these particle filters over commonly used procedures for this problem. Section IV considers another long-standing problem in nonlinear system identification, namely, combined system identification and state estimation in state space models. It is shown how particle filters can circumvent the nonlinearities in the problem without using linearization or, in case linearization is applied, can use the importance weights to correct for linearization errors. The concluding remarks in Section V further explain why particle filters,
with suitably chosen proposal distributions and resampling schemes, can resolve the difficulties in system identification due to nonlinearities without incurring unmanageable computational cost for on-line implementation.

II. FAST PARTICLE FILTERS IN HIDDEN MARKOV MODELS

A hidden Markov model (HMM) is a stochastic process \( (x_t, y_t) \) in which (i) \( \{x_t\} \) is an unobservable Markov process with transition probability density function \( f(\cdot|\cdot) \) with respect to some measure \( \nu \) on the state space, and (ii) given \( \{x_t\} \), the observable random variables \( y_t \) are conditionally independent such that \( y_s \) has density function \( g(\cdot|x_s) \) with respect to some measure. The filtering problem for HMM is to find the posterior distribution of the hidden state \( x_n \) given the current and past observations \( y_1, \ldots, y_n \). In particular, the optimal filter with respect to squared error loss is \( E(x_n|y_1, \ldots, y_n) \). In engineering applications, there are often computational constraints for on-line updating of the filter and recursive algorithms are particularly attractive. For infinite state spaces, direct computation of the optimal filters is not feasible except in linear Gaussian state space models, for which Kalman filtering provides explicit recursive filters. Analytic approximations or Monte Carlo methods are therefore used instead. Although Markov chain Monte Carlo has provided a versatile simulation-based tool to calculate the posterior distributions of hidden states in HMMs, it is cumbersome for updating and is too slow for on-line filtering problems.

The past decade has witnessed important developments in sequential Monte Carlo methods, which represent the posterior distributions by a large number \( m \) of random samples that are sequentially generated over time by using a combination of sequential importance sampling and resampling steps. Having such large \( m \) makes them impractical for on-line identification and control applications. We now proceed to show that by choosing appropriate resampling schemes and proposal distributions for importance sampling, we can arrive at good approximations to the optimal filter by using \( m \) (as small as 50) that is manageable for on-line identification and adaptive control.

A. Proposal Distribution for Sequential Importance Sampling

Let \( X_{1:n} = (x_1, \ldots, x_n), Y_{1:n} = (y_1, \ldots, y_n) \), and \( p(\cdot|\cdot), \ p(\cdot) \) denote the conditional and the joint density functions (under the measure \( P \)) of the random variables indicated. Given \( Y_{1:n} \), the conditional distribution of \( X_{1:n} \) is that of an inhomogeneous Markov chain with transition probability density

\[
p_t(x_t|x_{t-1}) := p(x_t|x_{t-1}, Y_{1:n}) \propto f(x_t|x_{t-1}) g(y_t|x_t) \int \prod_{i=t+1}^n f(x_i|x_{i-1}) g(y_i|x_i) d\nu(x_i), \tag{1}
\]
in which the constant of proportionality is the normalizing constant that makes the left hand side of (1) integrate to 1.

It is often difficult to sample directly from this Markov chain for Monte Carlo evaluation of the posterior distribution of \( x_n \) given \( Y_{1:n} \) to estimate the optimal filter \( E\{h(x_n)|Y_{1:n}\} \). Instead we sample from an alternative distribution \( Q \) under which \( X_{1:n} \) is an inhomogeneous Markov chain with transition density

\[
g_{t}(x_t|x_{t-1}) \propto f(x_t|x_{t-1})g(y_t|x_t),
\]

which is tantamount to replacing \( p(x_t|x_{t-1},Y_{1:n}) \) in (1) by \( p(x_t|x_{t-1},Y_{1:t}) \). The optimal filter can be expressed in terms of \( Q \) via

\[
E\{h(x_n)|Y_{1:n}\} = E_{Q}\left\{h(x_n)\prod_{t=1}^{n}(p(x_t|x_{t-1},Y_{1:n})/g_{t}(x_t|x_{t-1}))\right\},
\]

where \( E_{Q} \) denotes expectation under the measure \( Q \). Therefore, instead of drawing \( m \) samples \( X_{1:n}^{(1)}, \ldots, X_{1:n}^{(m)} \) from (1) and using \( m^{-1}\sum_{j=1}^{m}h(x_n^{(j)}) \) to estimate the optimal filter (3), we can draw \( m \) samples from (2) and estimate the optimal filter by

\[
\hat{h}_n = \left\{ \sum_{j=1}^{m}h(x_n^{(j)})w_n^{(j)} \right\} / \sum_{j=1}^{m}w_n^{(j)},
\]

where \( w_n^{(j)} \) are the importance weights given recursively by

\[
w_1^{(j)} = 1, \quad w_t^{(j)} = w_{t-1}^{(j)}p(y_t|x_{t-1}),
\]

noting that \( p(y_t|x_{t-1}) = \int f(x|x_{t-1})g(y_t|x)d\nu(x) \), \( p(y_1|x_0) \) and \( p(Y_{1:n}|x_0) \) are proportionality constants and

\[
\prod_{t=1}^{n}p(x_t|x_{t-1},Y_{1:n}) = p(X_n|x_0,Y_{1:n}) = \left\{ \prod_{t=1}^{n}f(x_t|x_{t-1})g(y_t|x_t) \right\} / p(Y_{1:n}|x_0),
\]

\[
\prod_{t=1}^{n}g_{t}(x_t|x_{t-1}) = \prod_{t=1}^{n}\{f(x_t|x_{t-1})g(y_t|x_t)/p(y_t|x_{t-1})\}.
\]

In the case where \( x_0 \) is specified by an initial distribution \( \pi_0 \), we replace \( x_0 \) above by \( x_0^{(j)} \) drawn from \( \pi_0 \) \( (j = 1, \ldots, m) \).

In situations where the normalizing constant \( \int f(x|x_{t-1})g(y_t|x)d\nu(x) \) in (2) does not have a closed-form expression, although sampling from \( Q \) that is defined via (2) can still be carried out by rejection sampling or other methods, the importance weights (5) do not have explicit formulas and rejection sampling slows down the procedure considerably. A better idea is to choose another
Q which is easier to sample from and has explicit formulas for the importance weights, and which approximates (2) in some sense. One way to do this is to use a finite mixture of Gaussian distributions to approximate (2), with suitably chosen mixing proportions, means and covariance matrices; see Section IV and [3]. Using (2) or more convenient approximations thereof as the proposal distribution for sequential importance sampling provides substantial improvement over the original particle filter of Gordon et al. [1] who simply use \( q_t(x_t|x_{t-1}) = f(x_t|x_{t-1}) \), not adapting to the observed data \( Y_{1:n} \). Whereas the “fully adaptive” distribution (1) is non-recursive (because \( Y_{1:n} \) and \( Y_{1:n+1} \) result in different transition probabilities \( p(x_t|x_{t-1}, Y_{1:n}) \) and \( p(x_t|x_{t-1}, Y_{1:n+1}) \)), the proposal distribution (2) is adaptive and recursive.

B. Periodic Rejuvenation via Resampling

The particle filter of Gordon et al. [1] is often called a “bootstrap filter” because, besides sampling \( x_t^{(j)} \) from \( f(\cdot|x_{t-1}^{(j)}) \) to form \( \hat{X}_{1:t}^{(j)} = (x_1^{(j)}, \ldots, x_{t-1}^{(j)}, x_t^{(j)}) \), it also resamples from \( \{\hat{X}_{1:t}^{(1)}, \ldots, \hat{X}_{1:t}^{(m)}\} \) with probability proportional to the importance weights \( w_t^{(j)} = w_{t-1}^{(j)} g(y_t|\hat{x}_t^{(j)}) \), thereby generating the particles (trajectories) \( X_{1:t}^{(1)}, \ldots, X_{1:t}^{(m)} \). In other words, at every \( t \) there is an importance sampling step followed by a resampling step. We can think of importance sampling as generating a weighted representation \((q(x_1, \ldots, x_t), w_t)\) of \( p(x_1, \ldots, x_t|Y_{1:t}) \) and resampling as transforming the weighted representation to an unweighted approximation of \( p(x_1, \ldots, x_t|Y_{1:t}) \). For the bootstrap filter, since resampling introduces additional variability, resampling at every \( t \) may result in substantial loss in statistical efficiency. In addition, the computational cost of resampling at every \( t \) also accumulates over time.

If we forgo resampling altogether, then we have a weighted representation \((q(x_1, \ldots, x_n), w_n)\) of \( p(x_1, \ldots, x_n|Y_{1:n}) \) at stage \( n \). In view of (3), if we use the normalized weights

\[
   w_t = \prod_{i=1}^{t} \frac{p(x_i|x_{i-1}, Y_{1:n})}{q_t(x_i|x_{i-1})},
\]

then \( m^{-1} \sum_{j=1}^{m} h(x_n^{(j)})w_n^{(j)} \) is an unbiased estimate of \( E\{h(x_n)|Y_{1:n}\} \). However, for large \( n \), sequential importance sampling without resampling also has difficulties because of the large variance of \( w_n \) in view of the following.

**Theorem 1:** Suppose the unobserved Markov chain (with transition density \( f(\cdot|\cdot) \)) of the HMM is Harris recurrent with stationary distribution \( \pi \). Conditional on the observations \( y_1, \ldots, y_n \) of the HMM, let \( Q \) be a probability measure under which the states \( x_t \) form a (possibly inhomogeneous) Markov chain with transition densities \( q_t(\cdot|\cdot) \) with respect to \( \nu \). Define \( w_t \) by (6). Then conditional on \( Y_{1:n}, \{w_t, \ t \geq 1\} \) is a martingale with mean 1 under \( Q \). Moreover, for the special cases (i) \( q_t(x_t|x_{t-1}) = f(x_t|x_{t-1}) \) and (ii) \( q_t(x_t|x_{t-1}) = f(x_t|x_{t-1}) g(y_t|x_t) \), \( n^{-1} \log w_n \) converges almost
surely under certain integrability assumptions and $E\{\text{var}_Q(w_n|Y_{1:n})\} \to \infty$ as $n \to \infty$.

The integrability assumptions in Theorem 1 are given in Appendix A together with its proof. A compromise between forgoing resampling altogether and resampling at every stage $t$ is to resample periodically. The motivation for resampling is to make multiple copies of the trajectories with large weights and to prune away those with small weights. The trajectories with small weights contribute little to the final estimate and it is a waste to carry many trajectories with very small weights. In particular, Kong et al. [4] propose to monitor the coefficient of variation (cv) of the importance weights $w_t$, defined by

$$cv^2 = \frac{\text{var}_Q(w_t)}{E_Q^2(w_t)},$$

and to perform resampling if the $cv^2$ of the current weights $w_t$ is greater than or equal to a certain bound. Specifically the procedure can be described as follows, starting with $m$ samples $X_{t-1}^{(1)}, \ldots, X_{t-1}^{(m)}$ having weights $w_{t-1}^{(1)}, \ldots, w_{t-1}^{(m)}$ at time $t-1$.

- a. Draw $\hat{x}_t^{(j)}$ from $q_t(x_t|x_{t-1}^{(j)})$ and update the weight $w_t^{(j)}$, $j = 1, \ldots, m$.

- b. If the $cv^2$ of $\{w_t^{(1)}, \ldots, w_t^{(m)}\}$ exceeds or equals a certain bound, resample from $\{\hat{x}_t^{(1)}, \ldots, \hat{x}_t^{(m)}\}$ with probabilities proportional to $\{w_t^{(1)}, \ldots, w_t^{(m)}\}$ to produce a random sample $\{x_t^{(1)}, \ldots, x_t^{(m)}\}$ with equal weights. Otherwise let $\{x_t^{(1)}, \ldots, x_t^{(m)}\} = \{\hat{x}_t^{(1)}, \ldots, \hat{x}_t^{(m)}\}$ and return to step a.

Strictly speaking, since the weight $w_t^{(j)}$ is associated with the entire path $\hat{X}_{1:t}^{(j)} = (x_1^{(j)}, \ldots, x_{t-1}^{(j)}, \hat{x}_t^{(j)})$, resampling should be performed on $\{\hat{X}_{1:t}^{(1)}, \ldots, \hat{X}_{1:t}^{(m)}\}$. However, because of the Markovian structure, the past observations $x_1, \ldots, x_{s-1}$ can be discarded after generating the current state $x_s$. This explains why $X_{1:t-1}^{(1)}, \ldots, X_{1:t-1}^{(m)}$ are discarded in Step b above. See [3] for further discussion of this resampling procedure and the theory of sequential importance sampling with resampling (SISR). In the second part of the following example, since the SISR filter is defined via certain functions of the Markov chain (instead of the Markov chain itself), resampling has to be performed on the sample of $m$ trajectories.

C. Application: Normal Mean Shift Model

Let $y_1, y_2, \ldots$ be a sequence of independent normal random variables with variance 1 and means $\mu_t$, where the $\mu_t$'s undergo occasional changes. Yao [5] studied the fully Bayesian model of mean shifts. At time $t$, independently of what happens at other times, $\mu_t$ equals $\mu_{t-1}$ (i.e., undergoes no change) with probability $1 - \rho$ and assumes a new value, which is normally distributed with mean 0 and variance $\sigma^2$, with probability $\rho$. Yao developed an algorithm to compute the Bayes estimate $E(\mu_n|Y_{1:n})$ with $O(n^2)$ operations. Note that $\{(\mu_t, y_t), t \geq 1\}$ forms a HMM, with $g(\cdot|\mu)$ being the normal density function with mean $\mu$ and variance 1 and such that the transition probability distribution of $\mu_t$ has (i) a discrete component putting mass $1 - \rho$ at $\mu_{t-1}$ and (ii) an absolutely
continuous component having density function \(\rho \sigma^{-1} \phi(\cdot/\sigma)\), where \(\phi\) denotes the standard normal density function. The proposal distribution (2) (with \(x_t = \mu_t\)) for \(\mu_t|\mu_{t-1}\) is a mixture of a degenerate distribution at \(\mu_{t-1}\) and a normal distribution with mean \(y_t/(1 + \sigma^{-2})\) and variance \(1/(1 + \sigma^{-2})\), with mixing probabilities proportional to \((1 - \rho)\phi(y_t - \mu_{t-1})\) and \(\rho(\sigma^2 + 1)^{-1/2} \phi(y_t/(\sigma^2 + 1)^{1/2})\), respectively. It is, therefore, easy to sample from this proposal distribution. Because of the discrete and absolutely continuous components of the transition probability distribution, the importance weights (see (5)) \(w_t\) are now given recursively by \(w_1 = 1\) and

\[
w_t = w_{t-1}\{\rho(\sigma^2 + 1)^{-1/2} \phi(y_t/(\sigma^2 + 1)^{1/2}) + (1 - \rho)\phi(y_t - \mu_{t-1})\}.
\]

Instead of working with the unobserved Markov chain \(\{\mu_t\}\), it is more efficient to consider the latent variable: \(I_t = 1_{\{\mu_t \neq \mu_{t-1}\}}\), indicating whether \(t\) is a change-point. Let \(I_{1:n} = \{I_1, \ldots, I_n\}\). Following [5], we rewrite the optimal filter as

\[
E(\mu_n|Y_{1:n}) = E\{E[\mu_n|I_{1:n}, Y_{1:n}]|Y_{1:n}\} = E\left\{(n - C_n + 1 + \sigma^{-2})^{-1} \sum_{j = C_n}^n y_j|Y_{1:n}\right\},
\]

where \(C_t = \max\{j \leq t : I_j = 1\}\) (\(\max \emptyset = 1\)), i.e., \(C_t\) is the most recent change-point up to time \(t\). Consider the proposal distribution \(Q\) for which \(I_t|I_{1:t-1}\) has the same distribution as \(P(I_t = I_{1:t-1}, Y_{1:t})\). It is easy to sample \(I_1, \ldots, I_n\) sequentially from \(Q\), under which \(I_t|I_{1:t-1}\) is Bernoulli assuming the values 1 and 0 with probabilities in the proportion

\[
\frac{\rho}{(1 + \sigma^2)^{1/2} \phi\left(\frac{y_t}{(1 + \sigma^2)^{1/2}}\right)} : \frac{1 - \rho}{(1 + v_t)^{1/2} \phi\left(\frac{y_t - v_t \sum_{j = C_{t-1}}^{t-1} y_j}{(1 + v_t)^{1/2}}\right)},
\]

where \(v_t = (t - C_{t-1} + \sigma^{-2})^{-1}\). Letting \(a_t(\rho)\) and \(b_t(\rho)\) denote the two terms in (9), note that \(p(y_t|I_{1:t-1}, Y_{1:t-1}) = a_t(\rho) + b_t(\rho)\). Combining this with

\[
\frac{p(I_{1:n}|Y_{1:n})}{q(I_{1:n})} = \frac{p(y_1)}{p(Y_{1:n})} \prod_{t=2}^n \frac{p(I_{1:t-1}, Y_{1:t})}{p(I_{1:t-1}, Y_{1:t-1})} \propto \prod_{t=2}^n p(y_t|I_{1:t-1}, Y_{1:t-1})
\]

yields the following recursive formula for the importance weights \(w_t\):

\[
w_t = w_{t-1}\{a_t(\rho) + b_t(\rho)\}, \quad t \geq 2; \quad w_1 = 1.
\]

When \(\rho\) is small, change-points occur very infrequently and many sequences \(I_{1:n}^{(j)}\) sampled from \(Q\) may contain no change-points. We can modify \(Q\) by increasing \(\rho\) in (9) to \(\tilde{\rho}\), thereby picking up more change-points, and adjust the importance weights accordingly. Specifically, take \(\tilde{\rho} > \rho\) and choose the proposal distribution \(\tilde{Q}\) for which \(I_t|I_{1:t-1}\) is a Bernoulli random variable with success probability \(a_t(\tilde{\rho})/(a_t(\tilde{\rho}) + b_t(\tilde{\rho}))\). Since

\[
p(I_{1:n}|Y_{1:n})/\tilde{q}(I_{1:n}) = \{p(I_{1:n}|Y_{1:n})/q(I_{1:n})\}\{q(I_{1:n})/\tilde{q}(I_{1:n})\},
\]

6
the importance weights $\tilde{w}_t$ can be determined recursively by

$$
\tilde{w}_t = \begin{cases} 
\tilde{w}_{t-1}\{a_t(\tilde{\rho}) + b_t(\tilde{\rho})\}a_t(\rho)/a_t(\tilde{\rho}) & \text{if } I_t = 1, \\
\tilde{w}_{t-1}\{a_t(\tilde{\rho}) + b_t(\tilde{\rho})\}b_t(\rho)/b_t(\tilde{\rho}) & \text{if } I_t = 0,
\end{cases}
$$

(11)

with $\tilde{w}_1 = (\rho/\tilde{\rho})1_{\{I_1 = 1\}} + 1_{\{I_1 = 0\}}(1 - \rho)/(1 - \tilde{\rho})$, assuming $\mu_0 \sim N(0, \sigma^2)$.

Table I studies how the $cv^2$ bound for resampling affects performance, using the sum of squared error criterion $\text{SSE} = E\{\sum_{t=1}^n(\hat{\mu}_t - \mu_t)^2\}$ to evaluate the performance of a procedure. For $\rho = 0.001$, 100 sequences of observations, each of length $n = 10000$, were generated. We applied SISR ($m = 50$) with $\tilde{\rho} = \rho$ to each sequence with different $cv^2$ bounds. As pointed out in the paragraph preceding this example, resampling is performed at time $t$ with the entire vector $I_{1:t}$ (instead of $I_t$) so that we can keep track of the most recent change-point in each sequence. Table I displays the average number of resamplings (Resampling #) used for each $cv^2$ bound, together with the SSE and its standard error (s.e.) based on 100 simulation runs. It shows the best value of SSE around 188 when we choose 1 as the $cv^2$ bound, involving an average of 51 resamplings. Table II compares the performance of the SISR estimate $\hat{\mu}_t$ with that of the Bayes estimate $\hat{\mu}_t^B$, which can be computed by Yao's [5] algorithm when $n$ is not too large. We used $m = 25$ simulated trajectories in SISR. Data were generated for three different scenarios: $\rho = 0.001$, 0.003, 0.01. For each scenario, 200 sequences of observations, each of length $n = 10000$, were generated. For each sequence, the ratio $\sum_{t=1}^n(\hat{\mu}_t - \mu_t)^2 / \sum_{t=1}^n(\hat{\mu}_t^B - \mu_t)^2$ was calculated as a measure of the relative performance of the particular approximation to the Bayes estimate. Table II shows that this ratio is close to 1 for all different scenarios, indicating that the SISR estimate performs almost as well as the Bayes estimate. Computing the exact Bayes estimate, however, is about 800–1000 times slower than SISR for these scenarios.

INSERT TABLES I AND II ABOUT HERE

We have also computed the SSE of the SISR filter based on $\mu_n$ and have found over 50% reduction in SSE by working with $I_n$ instead of $\mu_n$. In addition, we have studied how SISR performs when different $\tilde{\rho}$'s are used in the sampling distribution by simulating data from the same setting as Table I, but with the $cv^2$ bound fixed at 1. Our results show that for $\rho(= 0.001) < \tilde{\rho} < 100\rho$, the SSE is always smaller than that of $\tilde{\rho} = \rho$, with the smallest SSE at $\tilde{\rho} = 5\rho$, which shows the benefits of tilting.
III. IDENTIFICATION AND CONTROL OF ARX MODELS WITH PARAMETER JUMPS

As noted by Ljung and Gunnarson [6], a parameterized description of a dynamic system that is convenient for identification is to specify the model’s prediction of the output $y_t$ as a function of the parameter vector $\theta$ and past inputs and outputs $u_s$ and $y_s$, respectively, for $s < t$. When the function is linear in $\theta$, this yields the regression model $y_t = \theta^T \varphi_t + \epsilon_t$, which includes as a special case the ARX model (autoregressive model with exogenous inputs) that is widely used in control and signal processing. Here the regressor vector is

$$\varphi_t = (-y_{t-1}, \ldots, -y_{t-k}, u_{t-1}, \ldots, u_{t-h})^T,$$  \hspace{1cm} (12)

consisting of lagged inputs and outputs.

Whereas a comprehensive methodology has been developed for identification and control of ARX systems with time-invariant parameters (see e.g. [7]-[11]), the case of time-varying parameters in system identification and adaptive control still awaits definitive treatment despite a number of major advances during the past decade ([12]-[14]). In this section we show how particle filters can be used to resolve some of the long-standing difficulties due to the nonlinear interactions between the dynamics of the regressor vector (12) and of the parameter changes in the model $y_t = \theta_t^T \varphi_t + \epsilon_t$. Unlike continually fluctuating parameters modeled by a random walk in [12]-[14], we consider here the parameter jump model similar to that in Section IIC or that in Eq. (21)-(22) of [6]. As reviewed in [6, p. 11], an obvious way to handle parameter jumps is to apply carefully designed on-line change detection algorithms to segment the data. Another approach, called AFMM (adaptive forgetting through multiple models), is to use Bayesian updating formulas to calculate the posterior probability of each member in a family of models locating the change-points. To keep a fixed number of such models at every stage, the model with the lowest posterior probability is deleted while that with the highest posterior probability gives birth to a new model by allowing for a possible change-point from it. The fast particle filters in Section II enable us to develop a much more precise implementation of the Bayesian approach than AFMM, with little increase in computational cost, and to come up with more efficient adaptive control schemes.

Letting $\theta_t = (a_{1,t}, \ldots, a_{k,t}, b_{1,t}, \ldots, b_{h,t})^T$, we can write the ARX model

$$y_t + a_{1,t} y_{t-1} + \ldots + a_{k,t} y_{t-k} = b_{1,t} u_{t-1} + \ldots + b_{h,t} u_{t-h} + \epsilon_t$$  \hspace{1cm} (13)

in the regression form $y_t = \theta_t^T \varphi_t + \epsilon_t$, $t \geq t_0 \overset{\triangle}{=} \max(k, h) + 1$. As in Section IIC, suppose that the change-times of $\theta_t$ form a discrete renewal process with parameter $\rho$, or equivalently, that $I_t \overset{\triangle}{=} 1_{\{t \neq t_{t-1}\}}$ are independent Bernoulli random variables with $P(I_t = 1) = \rho$ for $t \geq t_0 + 1$, \hspace{1cm} (14)
assuming \( I_{t_0} = 1 \). At a change-point, \( \theta_t \) takes a new value which is assumed to have the multivariate normal distribution with mean \( \mu \) and covariance matrix \( V \). Assume also that the \( \epsilon_t \) are independent normal with mean 0 and variance \( \nu_t \), which is chosen to be 1 in the following for simplicity.

Let \( C_t \) be the most recent change-time up to time \( t \), as in Section IIC. The conditional distribution of \( \theta_t \) given \( C_t, y_t \) and \( \varphi_s, s \leq t \), is normal with mean \( \mu_{C_t,t} \) and covariance matrix \( V_{C_t,t} \), where for \( t_0 < s \leq t \),

\[
V_{s,t} = \left( V^{-1} + \sum_{j=s}^{t} \varphi_j \varphi_j^T \right)^{-1}, \quad \mu_{s,t} = V_{s,t} \left( V^{-1} \mu + \sum_{j=s}^{t} y_j \varphi_j \right),
\]

which can be computed by standard recursions that follow from the matrix inversion lemma:

\[
V_{t,t} = V - V \varphi_t \varphi_t^T V / (1 + \varphi_t^T V \varphi_t),
\]

\[
V_{s,t} = V_{s,t-1} - V_{s,t-1} \varphi_t \varphi_t^T V_{s,t-1} / (1 + \varphi_t^T V_{s,t-1} \varphi_t) \text{ if } s < t,
\]

\[
\mu_{s,t} = \mu_{s,t-1} + V_{s,t-1} \varphi_t (y_t - \varphi_t \mu_{s,t-1}) / (1 + \varphi_t^T V_{s,t-1} \varphi_t).
\]

Therefore, analogous to (8), the optimal filter is given by

\[
E(\theta_n | y_n, \varphi_t, t \leq n) = E\{E(\theta_n | C_n, y_n, \varphi_t, t \leq n) | y_n, \varphi_t, t \leq n\}
= E\{\mu_{C_n,n} | y_n, \varphi_t, t \leq n\}.
\]

We can compute (15) by simulating \( m \) trajectories \( I^{(j)}_{1:n} \) \( (j = 1, \ldots, m) \) via sequential importance sampling with resampling. The proposal distribution \( Q \) is similar to that in Section IIC. Analogous to (9), the conditional distribution of \( I_t \) given \( I_{1:t-1} \) is Bernoulli assuming the values 1 and 0 with probabilities in the proportion

\[
\frac{\rho}{(1 + \varphi_t^T V \varphi_t)^{1/2}} \Phi \left( \frac{y_t}{(1 + \varphi_t^T V \varphi_t)^{1/2}} \right) \quad \frac{1 - \rho}{(1 + \varphi_t^T V_{C_t-1,t-1} \varphi_t)^{1/2}} \Phi \left( \frac{y_t - \mu_{C_t-1,t-1} \varphi_t}{(1 + \varphi_t^T V_{C_t-1,t-1} \varphi_t)^{1/2}} \right).
\]

Letting \( a_t(\rho) \) and \( b_t(\rho) \) denote the two terms in (16), we can define the importance weights \( w_t \) recursively by (10). Resampling is performed when the squared coefficient of variation of the importance weights \( w_t^{(j)} \) exceeds some threshold, which we can choose as 1 that usually works quite well. When \( \rho \) is small, we can modify \( Q \) by increasing \( \rho \) in (16) to \( \rho \) and adjusting the importance weights accordingly, as described in Section IIC.

So far we have assumed \( \rho \) to be fully specified in the change-point model. When the frequency of change-points is unknown, we can put a beta distribution on \( \rho \), namely, the Beta\( c, d \) distribution with density function proportional to \( \rho^{c-1}(1 - \rho)^{d-1}, \) \((0 < \rho < 1)\), which has mean \( c / (c + d) \) and variance \( cd / \{(c + d)^2(c + d + 1)\}\), where \( c \) and \( d \) are positive integers. We can use the
same arguments as before to derive an analogous proposal distribution \( Q \) from which \( I_1, \ldots, I_t \) are sampled sequentially. In particular, it can be shown that under \( Q \), \( I_t | I_{1:t-1} \) is Bernoulli assuming the values 1 and 0 with probabilities in the proportion

\[
\frac{n_{t-1,1} + c}{(t - 1 + c + d)(1 + \varphi_t^T V \varphi_t)^{1/2}} \phi \left( \frac{y_t}{(1 + \varphi_t^T V \varphi_t)^{1/2}} \right) : \frac{n_{t-1,0} + d}{(t - 1 + c + d)(1 + \varphi_t^T V \varphi_t)^{1/2}} \phi \left( \frac{y_t - \mu_{t-1,t-1}^T \varphi_t}{(1 + \varphi_t^T V \varphi_t)^{1/2}} \right)
\]

(17)

where \( n_{t-1,i} \) is the number of \( i \)'s in \( \{I_1, I_2, \ldots, I_{t-1}\} \) for \( i = 0, 1 \). Note how closely (17) resembles (16). Accordingly, letting \( a_t \) and \( b_t \) denote the two terms in (17), the importance weights \( w_t \) are given recursively by \( w_t = w_{t-1}(a_t + b_t) \), \( t \geq 1 \), with \( w_0 = 1 \).

### 4. A. Identification in Open-Loop Designs

Consider an open-loop design in which \( \{u_t\} \) is independent of \( \{\epsilon_t\} \) and is a stationary ergodic sequence such that

\[
n^{-1} \sum_{t=h+1}^{n} u_t u_t^T \text{ converges almost surely to a positive definite matrix as } n \to \infty,
\]

(18)

where \( u_t = (u_{t-1}, \ldots, u_{t-h})^T \). In the time-invariant case (i.e., \( \theta_t \equiv \theta_0 \)), \( \mu_{t+h,1} \) given by (14) is a consistent estimate of \( \theta_0 \) for such open-loop designs even when the system is unstable, as can be shown by a modification of the argument in [15] for general (not necessarily stationary) AR models. In control applications, one wants a stable system, but the normal prior distribution \( N(\mu, V) \) assumed in (14) assigns positive probability to unstable parameter configurations. Let \( A \) be a stability region such that

\[
\inf_{|z| \leq 1} |1 + \alpha_1 z + \ldots + \alpha_k z^k| > 0 \text{ if } (\alpha_1, \ldots, \alpha_k)^T \in A.
\]

(19)

We shall therefore replace the multivariate normal distribution \( N(\mu, V) \) on \( \theta \) by a truncated normal distribution \( T_A N(\mu, V) \); a random vector \( \theta = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_h)^T \) is said to have this truncated normal distribution if \( P\{\theta \in C\} = P\{W \in C| (w_1, \ldots, w_k)^T \in A\} \) for all \( C \subset A \times \mathbb{R}^k \), where \( W = (w_1, \ldots, w_{k+h})^T \) has the \( N(\mu, V) \) distribution. Although the constraint set \( A \) is important in generating a stable open-loop system, it has little effect on the performance of the particle filter which involves (14) - (17) that ignores the stability constraint in the data generating mechanism, and which we can still use for on-line estimation of the parameters undergoing occasional jumps.

The following theorem makes use of the convergence theory of Markov chains (cf. [16]) to derive the
asymptotic theory of Bayes estimates of $\theta_t$ in the ARX model (13) when the inputs are Markovian. Let $\hat{\theta}_t^{(B)}$ denote the Bayes estimate of $\theta_t$ given by (14)-(15) that ignores, for computational and analytic tractability, the stability constraint (19) on the prior distribution $\theta_t$.

**Theorem 2**: In the ARX model (13), suppose that $\{u_t\}$ is an irreducible (with respect to some irreducibility measure) and aperiodic Markov chain such that $\{u_t\}$ is bounded and satisfies (18), and that $\theta_t = I_t W_t + (1 - I_t) \theta_{t-1}$, in which $I_t$ are independent Bernoulli variables with $P(I_t = 1) = \rho = 1 - P(I_t = 0)$ and $W_t$ has the truncated normal $T_N(\mu, V)$ distribution, where $A$ satisfies the stability condition (19). Then $(\theta_t, \varphi_{t+1})$ is a Markov chain with stationary distribution $\Pi$ and converges weakly to $\Pi$ exponentially fast as $t \to \infty$. Moreover, $\hat{\theta}_t^{(B)}$ has a limiting distribution to which it converges weakly as $t \to \infty$, at an exponential rate.

The proof of Theorem 2 is given in Appendix B. In view of Theorem 2, we can make use of the stationary distribution $\Pi$ to evaluate the limiting mean squared error $E \| \hat{\theta}_t^{(B)} - \theta_t \|^2$ as $t \to \infty$. We can also make use of $\Pi$ to evaluate the limiting mean squared errors of other estimates of $\theta_t$ proposed in the literature. In particular, the following two modifications of the usual least squares estimate have been commonly used to handle occasional jumps in $\theta_t$ (cf. [17, pp. 140, 161]):

(a) Sliding window: The least squares estimate is applied only to data in the immediate past, i.e., to the data set $\{(y_s, \varphi_s) : t - k < s \leq t\}$, where $k$ is the window size.

(b) Forgetting factor: A weighted least squares estimate is used, with weight $p^{t-s}$ for $(y_s, \varphi_s)$, i.e., the estimate at time $t$ minimizes $\sum_{s=1}^t p^{t-s}(y_s - \theta^T \varphi_s)^2$, where $0 < p < 1$ is the “forgetting factor” to discount past observations.

Theorem 2 can be applied to show that the sliding window estimate $\hat{\theta}_t^{(W)}$ and the forgetting factor least squares estimate $\hat{\theta}_t^{(F)}$ have limiting distributions as $t \to \infty$; see Remark in Appendix B, which also compares these asymptotic distributions with that of $\hat{\theta}_t^{(B)}$ that is demonstrated to be superior. The following simulation study comparing $\hat{\theta}_t^{(B)}$ with $\hat{\theta}_t^{(W)}$ and $\hat{\theta}_t^{(F)}$ again shows the superiority of $\hat{\theta}_t^{(B)}$, in agreement with the asymptotic theory.

**Example 1**: Consider the ARX model (13) with $k = 2$, $h = 1$, $\rho = 0.001$, $V =$ identity matrix, and assume the stability region $A$ in (19) is of the form $\{(a_{1,t}, a_{2,t}) : |a_{1,t}| + |a_{2,t}| < 1\}$. The inputs $u_t$ are assumed to be independent standard normal random variables. To compare the performance of $\hat{\theta}_t^{(B)}$ with $\hat{\theta}_t^{(W)}$ and $\hat{\theta}_t^{(F)}$, we simulated 100 sequences of observations, each of length $n = 3000$, from the model. We implemented $\hat{\theta}_t^{(B)}$ by its SISR approximation, with $m = 50$ sequentially generated trajectories. We used $\bar{\rho} = 5\rho$ to form the proposal distribution and took $\text{cv}^2$ bound $= 1$ for the resampling scheme. In addition, we also considered the rule SISR* (with $m = 50$) that does not assume $\rho$ to be known but uses a beta prior distribution $\text{Beta}(1, 999)$ with mean 0.001 and
standard deviation 0.001. We evaluated the estimates \( \hat{\theta}_t \ (= \text{SISR}, \text{SISR}^*, \hat{\theta}_t^{(W)}, \hat{\theta}_t^{(F)}) \) by the two performance measures

\[
\text{SSE}_1 = E\left\{ \sum_{t=m}^{n} (\hat{\theta}_t - \theta_t)^T \varphi_t \right\}, \quad \text{SSE}_2 = E\left\{ \sum_{t=m}^{n} ||\hat{\theta}_t - \theta_t||^2 \right\}.
\]  

(20)

The second measure considers how well \( \hat{\theta}_t \) estimates \( \theta_t \), whereas the first measure evaluates how well \( \hat{\theta}_t^T \varphi_t \) estimates the minimum variance predictor \( \theta_t^T \varphi_t \) of \( y_{t+1} \). The results are reported in Table III where we consider 1-500, 501-1000, 1001-2000 and 2001-3000 for the range from \( m \) to \( n \) in (20).

**INSERT TABLE III ABOUT HERE**

**B. Application of Fast Particle Filters to Adaptive Control**

We now consider the control problem of choosing the inputs \( u_t \) in the ARX model (13) so that the outputs \( y_{t+1} \) are as close as possible (in \( L^2 \)) to \( y^*_t \), some reference signal such that \( \{y^*_n\} \) and \( \{\theta_n, \epsilon_n\} \) are independent. In the case of known \( \theta_{t+1} \), the optimal input is defined by \( \theta_{t+1}^T \varphi_{t+1} = y^*_t \).

When \( \theta_{t+1} \) is unknown, the certainty equivalence rule replaces \( \theta_{t+1} \) in the optimal input by an estimate \( \hat{\theta}_{t+1|t} \) based on the observations up to time \( t \) so that \( u_t \) is given by \( \hat{\theta}_{t+1|t}^T \varphi_{t+1} = y^*_t \). We have assumed above that \( b_1, t+1 \) (or its estimate \( \hat{\theta}_{1,t+1|t} \) in the certainty equivalence rule) is nonzero.

Letting \( \gamma_{t+1} = (a_1, t+1, \ldots, a_k, t+1, b_2, t+1, \ldots, b_h, t+1)^T \),

\[
\gamma_{t+1|t} = (\hat{a}_1, t+1|t, \ldots, \hat{a}_k, t+1|t, \hat{b}_2, t+1|t, \ldots, \hat{b}_h, t+1|t)^T, \quad \psi_t = (-\gamma_{t+1}, \ldots, -\gamma_{t-k+1}, u_{t-1}, \ldots, u_{t-h+1})^T,
\]

we modify the certainty equivalence rule by

\[
u_t = \begin{cases} (y^*_t - \gamma_{t+1|t}^T \psi_t) / \gamma_{1,t+1|t} & \text{if } |\gamma_{1,t+1|t}| \geq \delta_t, \\ y^*_t + \omega_t & \text{if } |\gamma_{1,t+1|t}| < \delta_t, \end{cases}
\]  

(21)

where \( \delta_t \) is some small prespecified number and \( \omega_t \) is extraneous noise used to enhance the information content of the reference signal (including the case \( \omega_t \equiv 0 \) if the reference signal is already persistently exciting), cf. [18].

A drawback of certainty equivalence rules is that they do not adjust for uncertainties in the estimates \( \hat{\theta}_{t+1|t} \) that are used to replace \( \theta_{t+1} \) in the optimal rule \( \theta_{t+1}^T \varphi_{t+1} = y^*_t \). To develop an alternative approximation to the optimal control in the case of unknown \( \theta_{t+1} \), we focus on the one-step ahead error \( y_{t+1} - y_{t+1}^* \) without making use of dynamic programming to determine how the current control \( u_t \) impacts on the information content of the estimates \( \hat{\theta}_{t+1|t} \) for the future.
errors \( y_{s+1} - y^*_t \) \((s > t)\). Noting that

\[
\arg \min_u \mathbb{E}\{(b_1, t+1)_u + \gamma_t^T \psi_t - y^*_t)^2 | y_t, y_{t-1}, u_{t-1}, \ldots, y_1, u_1\} = \frac{\{y^*_t E_t(b_1, t+1) - E_t(b_1, t+1) \gamma_t^T \psi_t\}}{E_t(b_1, t+1)}
\]

(22)

where \( E_t \) denotes conditional expectation given \( \{y_t, y_{t-1}, u_{t-1}, \ldots, y_1, u_1\} \), we define the following variant of (21) that incorporates uncertainty adjustments due to unknown parameters into the optimal rule \( \theta_t^T \eta_{t+1} = y^*_t \) assuming known \( \theta_t \):

\[
u_t = \begin{cases} 
\{y^*_t E_t(b_1, t+1) - E_t(b_1, t+1) \gamma_t^T \psi_t\}/E_t(b_1, t+1) & \text{if } \sqrt{E_t(b_1, t+1)} \geq \delta_t, \\
y^*_t + \omega_t & \text{if } \sqrt{E_t(b_1, t+1)} < \delta_t.
\end{cases}
\]

(23)

To implement the above adaptive control strategies, we need to compute the one-step ahead predictors \( E_t(b_1, t+1) \), \( E_t(b_1, t+1) \gamma_t^T \) and \( E_t(b_1, t+1) \gamma_t^T \). Note that

\[
\mathbb{P}(\theta_{t+1} | y_t, y_{t-1}, u_{t-1}, \ldots, y_1, u_1) = (1 - \rho)\mathbb{P}(\theta_t | y_t, y_{t-1}, u_{t-1}, \ldots, y_1, u_1) + \rho \mathbb{P}(\theta_{t+1}).
\]

(24)

The first term on the right hand side of (24) can be approximated by fast particle filters, whereas the second term corresponds to a change-point at time \( t + 1 \). Note that replacing \( E_t(b_1, t+1), E_t(b_1, t+1), E_t(b_1, t+1) \gamma_t^T \) by \( (\hat{b}_{1, t+1}, \hat{b}_{1, t+1} \gamma_t^T) \) in (23) reduces it to the certainty equivalence rule (21), which simply uses the estimates \( (\hat{b}_{1, t+1}, \hat{b}_{1, t+1} \gamma_t^T) \) to substitute for \( (b_1, t+1, \gamma_t^T) \) in the optimal control assuming known \( \theta_{t+1} \). The rule (23) introduces uncertainty adjustments for the unknown \( \theta_{t+1} \) by considering the expected one-step ahead control error \( E_t(y_{t+1} - y^*_t)^2 \) that leads to (22), and by introducing extraneous white noise to enhance the information content of the control for future estimates of \( \theta_t \) whenever (22) has a small denominator that may lead to a large (and numerically unstable) control action. The choice of \( \delta_t \) depends on whether \( \theta_t \) is known to belong to some stability region. If the \( \theta_t \) are restricted to a stability region satisfying (19), then white noise perturbations do not destabilize the system while improving the experimental design for future estimates of \( \theta_t \). On the other hand, without such a priori stability assumption on the \( \theta_t \), small \( \delta_t \) should be used in (23) because the perturbations \( \omega_t \) can have an explosive effect. The following two examples illustrate this point and show that the adaptive control rule (23) can also perform well without the stability constraint (19). Both examples consider the case \( y_t^* \equiv 0 \).

**Example 2:** Table IV reports the results of a simulation study for the ARX model of Example 1 but with the inputs \( u_t \) determined by the certainty equivalence rule (21) and the uncertainty-adjusted certainty equivalence rule (23), in which \( \delta_t = 1/4 \) and the \( \omega_t \) are i.i.d. \( N(0, 1/4) \) random variables. We used two performance criteria to evaluate a procedure: \( \text{SSE}_1 \) and \( \text{SSE}_3 = \)
\[ E\{\sum_{t=m}^{n} (y_t - \epsilon_t)^2\} \]. Besides \( \rho = 0.001 \) as in Example 1, Table IV also considers the case \( \rho = 0.01 \) giving more frequent change-points. It shows that the certainty equivalence rules based on \( \hat{\theta}_t^{(W)} \) or \( \hat{\theta}_t^{(F)} \) perform much worse than those based on \( \hat{\theta}_t^{(B)} \) implemented by the SISR approximation, and that the modification (23) of the certainty equivalence (21) based on \( \hat{\theta}_t^{(B)} \) outperforms (21). We have also considered the effect of using weaker perturbing signals \( \omega_t \) that have smaller variances and lower thresholds \( \delta_t \) in the rules (21) and (23) based on \( \hat{\theta}_t^{(B)} \) and have found somewhat worse performance than those shown in Table IV.

**INSERT TABLE IV ABOUT HERE**

**Example 3:** Without the stability constraint (19), the threshold \( \delta_t = 1/4 \) and the variance 1/4 for \( \text{var}(\omega_t) \) used in the SISR-based rules (21) and (23) fail to stabilize the system. This is shown in a simulation study that is reported in Table IV for the same setting as that in Example 2 but with \( \alpha_{2,t} = 0 \) assumed to be known and with a larger truncation set \( A = \{|a_{1,t}t| \leq 1.1\} \). The entries \( " > 10^{308}" \) indicate that they exceed the computer’s limit of very large numbers. The simulation study also considers using \( \delta_t = 0.04 = \text{var}(\omega_t) \) to weaken the extraneous perturbations and the results are reported in Table V for \( A = \{|a_{1,t}t| \leq \alpha\} \) with \( \alpha = 1.1, 1.3, 1.5 \), showing that the SISR-based rule with this choice of \( \delta_t \) and \( \text{var}(\omega_t) \) can still stabilize the system.

**INSERT TABLE V ABOUT HERE**

We next provide an asymptotic theory to explain the simulation results in Examples 2 and 3 showing the stability and efficiency of the uncertainty-adjusted certainty equivalence rule (23). To begin with, consider the case in which \( h = 1 \) and \( b_{1,t} \equiv \beta \) with known value of \( \beta \) undergoing no change. In this case, \( \theta_t = (a_{1,t}, \ldots, a_{k,t})^T \) and the ARX model (13) can be written in the regression form \( \tilde{y}_t = \theta^T \varphi_t + \epsilon_t \), with \( \varphi_t = (-y_{t-1}, \ldots, -y_{t-k})^T \) and \( \tilde{y}_t = y_t - \beta u_{t-1} \). Assume also that \( y_{*t} \) are independent bounded random variables having a common distribution with mean 0 and variance \( \sigma^2 > 0 \). If the control is effective in bringing \( y_s \) near \( y_{*t} \), then \( \varphi_s \) should be persistently exciting. Moreover, the information matrix \( \sum_{s=1}^{t} \varphi_s \varphi_s^T \) does not involve the inputs \( u_s \) which are absorbed into \( \tilde{y}_{s+1} \) (observed response variable in the regression model). Therefore choosing the input \( u_t \) to minimize \( E_t(\beta u + \theta_{t+1}^T \varphi_{t+1} - y_{*t+1})^2 \) should be asymptotically optimal in this case, leading to the rule

\[
{u}_t = \{y_{*t+1} - (E_t \theta_{t+1}^T) \varphi_{t+1}\}/ \beta. \tag{25}
\]

Whether this rule can stabilize the system depends on how well \( E_t \theta_{t+1} \) estimates the unobservable and time-varying \( \theta_{t+1} \). When \( \rho \) is small, the mean squared error \( E_t||\theta_{t+1} - E_t \theta_{t+1}||^2 \) is also expected
to be small. On the other hand, if jumps occur frequently, then the system may not be stabilizable
and \( E_t||\theta_{t+1} - E_t\theta_{t+1}||^2 \) may not be small enough for \( E_t\theta_{t+1} \) to provide an adequate approximation
for \( \theta_{t+1} \). What is the threshold for \( E_t||\theta_{t+1} - E_t\theta_{t+1}||^2 \) below which the rule \( (25) \) yields a stable
system? Making use of the stability theory of Markov chains, we prove in Appendix B the following
theorem which not only gives an answer to this question but also provides a similar result for the
rule \( (23) \) in the case of general \( h \) and unknown \( b_{j,t} \) that undergoes concurrent jumps with the \( a_{i,t} \).
The theorem is applicable to any Markovian model of parameter dynamics (not necessarily the
jump model in the preceding discussion) such that \( \{\theta_t\} \) is independent of \( \{\epsilon_t\} \); the conditional
expectation \( E_t \) is with respect to the assumed model. Let \( \|A\| = \sup_{\|\varphi\|=1} \|A\varphi\| \) denote the norm
of a square matrix \( A \).

**Theorem 3:** (i) Suppose that in the ARX model \( (13) \), \( h = 1 \) and \( b_{1,t} \equiv \beta \), with known \( \beta \neq 0 \).
Letting \( \theta_t = (a_{1,t}, \ldots, a_{k,t})^T \) and \( \varphi_t = (y_{t-1}, \ldots, y_{t-k})^T \), suppose that \( \{\theta_t\} \) is a Markov chain on
\( \Theta \subset \mathbb{R}^k \) with stationary distribution \( \pi \) and that \( u_t \) is given by \( (25) \). Suppose the reference signal
\( \{y_t^*\} \) is an irreducible and aperiodic Markov chain and is bounded. Letting

\[
A_{t+1} = \begin{pmatrix}
-\theta_{t+1}^T & + & E_t\theta_{t+1}^T \\
1 & 0 & \cdots & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & 1 & 0
\end{pmatrix},
\]

suppose that \( \limsup_{t \to \infty} E_t\|A_{t+1}\| < 1 \) with probability 1. Then \( (\theta_t, \varphi_{t+1}) \) is an ergodic Markov
chain on \( \Theta \times \mathbb{R}^k \) and \( \varphi_t \) converges weakly to its stationary distribution geometrically fast.

(ii) For the general ARX model \( (13) \), define \( \varphi_t \) by \( (12) \) and let \( \theta_t = (a_{1,t}, \ldots, a_{k,t}, b_{1,t}, \ldots, b_{h,t})^T \).
Assume that \( \{\theta_t\} \) is a Markov chain on \( \Theta \subset \mathbb{R}^{k+h} \) with stationary distribution \( \pi \). Suppose that
\( \liminf_{t \to \infty} E_t b_{1,t}^2 \geq \delta \) with probability 1 for some \( \delta > 0 \) and that \( u_t \) is given by \( (23) \) in which \( \delta_t = \delta \).
Letting

\[
\alpha_{t+1} = (a_{1,t+1}, \ldots, a_{k,t+1})^T, \quad \beta_{t+1} = (b_{2,t+1}, \ldots, b_{h,t+1})^T,
\]

define the \((k + h - 1) \times (k + h - 1)\) matrix
Suppose that $\limsup_{t \to \infty} E_t \|B_{t+1}\| < 1$ with probability 1. Then $(\theta_t, \varphi_t)$ is an ergodic Markov chain on $\Theta \times R^{k+h}$ and $\varphi_t$ converges to its stationary distribution geometrically fast.

C. Extensions to Hammerstein and Nonlinear ARX Systems

The particle filter described in (14) - (17) can be applied to estimate the piecewise constant $\theta_t$ in the general stochastic regression model

$$y_t = \theta_t^T \varphi_t + \epsilon_t,$$

in which $\varphi_t$ is a vector-valued function of past lagged outputs and inputs and the change-points of $\theta_t$ form a discrete renewal process with parameter $\rho$, with $\theta_t$ taking a new value from the $N(\mu, V)$ (or $T_d N(\mu, V)$) distribution at each change-point. The ARX model (13) is only a special case of (26) with $\varphi_t$ given by (12). Another important special case is the Hammerstein system that has a static nonlinearity on the input side, replacing $u_t$ in (13) (and therefore (12) accordingly) by some nonlinear transformation $f(u_t)$. When $f$ is unknown, it is usually approximated by a polynomial $f(u) = c_1 u + \cdots + c_l u^l$ (cf. [19]). To identify the Hammerstein system, we express it in the form of (26) with

$$\varphi_t = (-y_{t-1}, \ldots, -y_{t-k}; u_{t-1}, \ldots, u_{t-k}; \ldots, u_{t-1}^l, \ldots, u_{t-k}^l)^T.$$

Instead of using polynomial to approximate $f$, we can use other basis functions (e.g., splines), yielding the representation $f(u) = \sum_{i=1}^l c_i \psi_i(u)$. Moreover, we can allow nonlinear interactions among the lagged outputs by making use basis function approximations, and thereby express nonlinear ARX models with occasionally changing parameters in the form of (26) with

$$\varphi_t = (\tilde{\psi}_1(y_{t-1}, \ldots, y_{t-k}), \ldots, \tilde{\psi}_m(y_{t-1}, \ldots, y_{t-k}); \psi_1(u_{t-1}), \ldots, \psi_l(u_{t-1}), \ldots, \psi_1(u_{t-k}), \ldots, \psi_l(u_{t-k}))^T.$$
Theorems 2 and 3, which are based on the stability theory of Markov chains, can be extended to these nonlinear ARX models that are linear in the parameters \( \theta_t \) which may undergo occasional changes.

IV. COMBINED SYSTEM IDENTIFICATION AND STATE ESTIMATION

In this section we make use of particle filters to estimate recursively both the parameter vector \( \theta \) and the unobserved state \( x_n \) in the state space model

\[
 x_{t+1} = \mu_t(x_t, \theta) + \epsilon_{t+1}, \quad y_t = h_t(x_t, \theta) + \eta_t, \tag{27}
\]

in which \( y_1, y_2, \ldots \) are the observations, \( \epsilon_t \) and \( \eta_t \) are unobserved disturbances that are independent normal with zero means and covariance matrices \( V_\epsilon \) and \( V_\eta \), and \( \mu_t \) and \( h_t \) are known functions that may involve observed inputs up to time \( t \). Even in the case where \( \mu_t \) and \( h_t \) are linear functions of \( x_t \) so that the Kalman filter can be used to estimate \( x_t \) when \( \theta \) is known, the joint estimation of \( (\theta, x_t) \) when \( \theta \) is also unknown is a well known nonlinear problem and has no closed-form solution; see e.g. [20], [21]. We shall assume that \( \mu_t \) and \( h_t \) are linear functions of \( \theta \) (but can be nonlinear in \( x_t \)), as is the case considered in Section IIIC. Moreover, as in Section III, we assume a normal (or truncated normal) prior distribution on \( \theta \).

A. Case of Linear Observation Equation

We begin by considering the case in which \( y_t = A_1 \theta + (A_2 + A_3 \theta^T A_4)x_t + \eta_t \) for some specified matrices \( A_i \) with suitably chosen orders. The prior distribution of \( \theta \) is multivariate normal, with mean \( \bar{\theta} \) and covariance matrix \( V_\theta \). Being linear in \( \theta \), the function \( \mu_t(x, \theta) \) in (27) can be expressed as \( M_t(x)\theta \) for some matrix \( M_t(x) \). We can make use of the procedure in Section IIA to generate sequentially \( m \) trajectories \( (\theta^{(j)}), X^{(j)}_{1:t−1}) \), with weights \( w^{(j)}_n \), as follows. First generate \( m \) independent \( \theta^{(1)}, \ldots, \theta^{(m)} \) from the \( N(\bar{\theta}, V_\theta) \) distribution. Having generated \( (\theta^{(j)}, X^{(j)}_{1:t−1}) \), the proposal distribution (2) used to generate \( x^{(j)}_t \) has density function proportional to

\[
 \phi_{V_\epsilon}(x - M_{t-1}(x_{t-1})\theta^{(j)})\phi_{V_\eta}(y_t - A_1\theta^{T(j)} - (A_2 + A_3\theta^{T(j)} A_4)x), \tag{28}
\]

where \( \phi_V \) denotes the density function of the \( N(0, V) \) distribution. Simple algebra shows that (28) is proportional to the density function of a multivariate normal distribution with mean

\[
 m^{(j)}_t = \{V_\epsilon^{-1} + (A_2 + A_3\theta^{T(j)} A_4)^T V_\eta^{-1} (A_2 + A_3\theta^{T(j)} A_4)\}^{-1} \\
 \times \{V_\epsilon^{-1} M_{t-1}(x_{t-1})\theta^{(j)} + (A_2 + A_3\theta^{T(j)} A_4)^T V_\eta^{-1}(y_t - A_1\theta^{(j)})\}
\]

17
and covariance matrix

\[ V_t^{(j)} = \{ V_{\varepsilon}^{-1} + (A_2 + A_3 \theta_t^{(j)} A_4)^T V_{\eta}^{-1} (A_2 + A_3 \theta_t^{(j)} A_4) \}^{-1}. \]

Noting that the conditional density \( p(\cdot | \theta, x_{t-1}) \) of \( y_t \) given \((\theta, x_{t-1})\) is normal with mean \( A_1 \theta + (A_2 + A_3 \theta T A_4) M_{t-1}(x_{t-1}) \theta \) and covariance matrix \( V_{\eta} + (A_2 + A_3 \theta A_4) V_{\varepsilon} (A_2 + A_3 \theta A_4)^T \), the weights \( w_t^{(j)} \) in (5) can be defined recursively in terms of the normal densities \( p(y_t | \theta_t^{(j)}, x_{t-1}^{(j)}) \). We also use periodic resampling as in Section IIB to keep the coefficients of variation of the importance weights within prespecified bounds.

Whereas the particle filter in Section II yields an approximation to the posterior distribution of \( x_n \) given \( Y_{1:n} \), evaluating the posterior distribution of \( \theta \) given \( Y_{1:n} \) is actually a smoothing (rather than filtering) problem. However, in view of the assumed linearity in \( \theta \) and the normal prior distribution on \( \theta \), we can retrieve the posterior distribution of \( \theta \) given \( Y_{1:n} \) from the \( x_n^{(j)} \) thus generated. Specifically, since \( P(\theta \in A | Y_{1:n}) = E[P(\theta \in A | X_{1:n}, Y_{1:n}) | Y_{1:n}] \), it follows that

\[ P(\theta \in A | Y_{1:n}) = \sum_{j=1}^m w_j^{(n)} P(\theta \in A | X_{1:n}^{(j)}, Y_{1:n}) / \sum_{j=1}^m w_j^{(n)}, \]

Moreover, \( P(\theta \in A | X_{1:n}^{(j)}, Y_{1:n}) \) has an explicit formula since the posterior distribution of \( \theta \) given \( (X_{1:n}, Y_{1:n}) \) is normal with mean

\[ m(X_{1:n}, Y_{1:n}) = V^T(X_{1:n}, Y_{1:n}) \{ \sum_{t=1}^n [(A_1 + A_3 x_t^T A_4)^T V_{\eta}^{-1} (y_t - A_2 x_t) + M_{t-1}^T (x_{t-1}) V_{\varepsilon}^{-1} x_t] + V_{\theta}^{-1} \theta \} \]

and covariance matrix

\[ V(X_{1:n}, Y_{1:n}) = \{ \sum_{t=1}^n [(A_1 + A_3 x_t^T A_4)^T V_{\eta}^{-1} (A_1 + A_3 x_t^T A_4) + M_{t-1}^T (x_{t-1}) V_{\varepsilon}^{-1} M_{t-1} (x_{t-1})] + V_{\theta}^{-1} \}^{-1}. \]

B. Modification and Enhancement at Resampling

Instead of the Markov-based resampling procedure described by Steps a, b in Section IIB, suppose that resampling is performed at stage \( t \) on \( \{ \hat{X}_{1:t}^{(1)}, \ldots, \hat{X}_{1:t}^{(m)} \} \), as in the last paragraph of Section IIC. Resampling then yields \( \{ X_{1:t}^{(1)}, \ldots, X_{1:t}^{(m)} \} \) with weights \( w_t^{(j)} \) (for \( X_{1:t}^{(j)} \) reset to 1. The empirical distribution of \( \{ x_t^{(1)}, \ldots, x_t^{(m)} \} \) is an approximation to the posterior distribution of \( x_t \) given \( Y_{1:t} \). Since the posterior distribution of \( \theta \) given \( (X_{1:t}, Y_{1:t}) \) is normal with \( m(X_{1:t}, Y_{1:t}) \) and covariance matrix \( V(X_{1:t}, Y_{1:t}) \), we can approximate it by the empirical distribution of \( \{ \theta_t^{(1)}, \ldots, \theta_t^{(m)} \} \), in which \( \theta_t^{(j)} \) is generated from \( N(m(X_{1:t}^{(j)}, Y_{1:t}), V(X_{1:t}^{(j)}, Y_{1:t})) \). Thus, instead of adhering to \( \theta_t^{(1)}, \ldots, \theta_t^{(m)} \) generated at the beginning from the prior distribution of \( \theta \), we regenerate them from their posterior distributions at the times when resampling is carried out.
Example 4: Consider as in Section 4.2 of [21] the model

\[ x_{t+1} = ax_t + u_t + \epsilon_{t+1}, \quad y_t = x_t + \eta_t, \]  

(29)

where \( \epsilon_{t+1} \) and \( \eta_t \) are standard normal random variables, \( x_1 = -5 \), and \( a = 0.9 \). Four different inputs \( u_t \) are used: (a) \( u_t \sim N(0,1) \); (b) \( P\{u_t = 2\} = P\{u_t = -2\} = 1/2 \); (c) \( u_t = \frac{16}{\pi^2} \sin \frac{\pi t}{k} \) and (d) \( u_t = 4\sqrt{10} \) if \( t = 10i + 1 \) for some nonnegative integer \( i \) (\( u_t = 0 \) otherwise). The problem is to estimate sequentially the autoregressive parameter \( a \) and the unobservable states \( x_t \) based on \( Y_{1:t} \). After simulating a trajectory \( (x_t, y_t) \), \( 1 \leq t \leq 1000 \), from this model, we apply the SISR filter, with \( m = 50 \), \( \text{cv}^2 \) bound = 20, and the enhanced resampling scheme described above to estimate \( x_t \) sequentially. The estimate of \( a \) at time \( t \) is computed by a weighted average of the posterior means of \( \theta \) given \( (X_{1:t}, Y_{1:t}) \), with weights \( w_j(t) \) as described above. For each of the four input types, Figure 1 plots the states \( x_t \) and their particle filters \( \hat{x}_t \). We have also simulated 100 such trajectories and compute the root mean squared errors \( \text{RMS}(x) = \{E[\sum_{t=1}^{n}(\hat{x}_t - x_t)^2/n]\}^{1/2} \) and \( \text{RMS}(a) = \{E(\hat{a}_n - a_n)^2\}^{1/2} \) at different time points \( n \leq 1000 \). The results are given in Table VI. Both Figure 1 and Table VI show that the above particle filter provides good estimates of \( x_t \) and \( a \).

C. More General Observation Equations

When \( h_t(x_t, \theta) \) in (27) is nonlinear in \( x_t \), the proposal distribution defined by (2) is non-Gaussian and the normalizing constant \( p(y_t|x_{t-1}) = \int f(x|x_{t-1})g(y_t|x)dx \) does not have a closed-form expression for convenient updates of the importance weights (5). Chen et al. [3] propose to use a finite mixture of Gaussian distributions to approximate (2) in this case. Their basic idea is to modify Gaussian sum approximations in the nonlinear filtering literature [22], [23] and use them to derive Gaussian mixture proposal distributions. To begin with, consider the special case of a single Gaussian distribution that is obtained via (2) by replacing \( h_t(x, \theta) \) by the linear approximation

\[ h_t(x_{\hat{t}|t-1}, \theta) + (\frac{\partial}{\partial x} h_t(x, \theta)) \bigg|_{x=x_{\hat{t}|t-1}} (x - x_{\hat{t}|t-1}), \]  

(30)

in which \( x_{\hat{t}|t-1} \) denotes the estimate of \( x_t \) based on \( Y_{1:t-1} \). Note that the extended Kalman filter also uses this linearization, and that the particle filter only uses such linearization to arrive at a tractable proposal distribution. Unlike the extended Kalman filter whose performance depends heavily on the adequacy of the linear approximation (30), the importance weights used in the particle filter can correct for the inaccuracies in the linear approximation (30) in view of (3) and therefore can still perform well even when the linear approximation is inadequate.
Instead of linearizing \( h_t \) at \( x_{t|t-1} \), one can linearize it at several other values when the conditional distribution of \( x_t \) given \( Y_{1:t-1} \) is multi-modal. Another possibility is to ignore the noise \( \eta_t \) in (27) and solve for \( y_t = h_t(x_t, \theta) \), taking a normal distribution centered at a solution as a component of the mixture. Details on the choice of the mixing proportions and covariance matrices defining the Gaussian mixture proposal distribution and on applications of these particle filters to signal processing are given in [3].

V. CONCLUSION

We have shown that recent advances in sequential Monte Carlo methods can provide relatively fast and accurate approximations to optimal, but intractable, filters in HMMs for general state spaces. In particular, using as few as 50 simulated trajectories in these sequential Monte Carlo filters already gives good approximations, making them implementable on-line for fast updating of the state estimates in nonlinear filtering applications, and achieving a good balance between statistical efficiency and computational complexity.

We have considered in particular two such nonlinear filtering applications. One is related to identification and control of ARX models in which the regression parameters are piecewise constant, with the unknown change-times forming a discrete renewal process. We make use of fast particle filters to compute Bayes estimates of the time-varying regression parameters, and apply the stability theory of Markov chains to develop an asymptotic theory of the parameter estimates in open-loop designs and also in feedback systems. We introduce an uncertainty-adjusted version of the certainty equivalence rule using these Bayes estimates and show its superiority over other rules in the literature in our asymptotic analysis and simulation studies.

The combined system identification and state estimation problem considered in Section IV further illustrates the effectiveness of particle filters in dealing with nonlinear identification. It provides a computationally tractable alternative to the extended Kalman filter and its variants (e.g. Ljung's [20] modification) that have been applied to this problem with varying success. After simulating the unobserved states \( x_t \), the parameter \( \theta \) can be conveniently estimated by closed-form Bayesian updating formulas for linear Gaussian systems. Even for nonlinear state dynamics in (27), the proposal distribution (2) is still Gaussian and is easy to sample from. We need linearization to find a tractable proposal distribution only when the observation equation is nonlinear. Moreover, even if the linearization used in that case is inadequate, the particle filter thus constructed can still provide a good approximation to the actual filtering density via the importance weights.
APPENDIX A: PROOF OF THEOREM 1

Proof. To begin with, note that \((X_{1:n}, Y_{1:n})\) is generated by the measure \(P\), and that given \(Y_{1:n}\) the measure \(Q\) is used to generate the Markov chain \(\{x_t^*, n \geq t \geq 0\}\) for Monte Carlo filtering, so \(\{x_t^*\}\) and \(\{x_t\}\) are conditionally independent given \(y_0, y_1, \ldots\). Let \(P^*\) denote the product measure induced by \(\{(x_t, y_t, x_t^*): t \geq 0\}\). Clearly the normalized weights \(w_t\) defined by (6) have mean 1 and form a martingale under \(Q\). Since \(\{x_t, t \geq 0\}\) is positive Harris, it suffices for the proof of \(n^{-1} \log w_n \to c < 0\) a.s. \([P^*]\) (i.e., almost surely under \(P^*\)) to show that the desired convergence holds under the assumption that \(X_0\) has the stationary distribution (cf. [16], p. 415), for which \(\{(x_t, y_t), t \geq 0\}\) is a stationary ergodic sequence. A generalization of the Shannon-McMillan-Breiman theorem for stationary sequences then yields

\[
n^{-1} \log p(Y_{1:n}) \to E\{\log p(y_0|y_{-1}, y_{-2}, \ldots)\} \text{ a.s. } [P], \tag{31}
\]

in which we have used stationarity to extend \(\{y_t, t \geq 0\}\) to the stationary sequence \(\{y_t, t = 0, \pm 1, \ldots\}\); see Section 4 of [24]. (Although [24] assumes the state space of \(x_t\) to be finite, what is actually used in that part of the paper is that \(\{y_t, t = 0, \pm 1, \ldots\}\) is a stationary sequence such that \((y_m, \ldots, y_n)\) has density function \(p(y_m, \ldots, y_n)\) with respect to some product measure for \(m \leq n\).)

For the case \(q_t(x_t|x_{t-1}) = f(x_t|x_{t-1})\), \(u_n = \left(\prod_{i=1}^{n} g(y_i|x_i^*)\right)/p(Y_{1:n})\) and therefore

\[
n^{-1} \log w_n = n^{-1} \sum_{i=1}^{n} \log g(y_i|x_i^*) - n^{-1} \sum_{i=1}^{n} \log p(y_i|Y_{1:i-1}). \tag{32}
\]

By the strong law for positive Harris chains (Theorem 17.1.7 in [16]),

\[
n^{-1} \sum_{i=1}^{n} \log g(y_i|x_i^*) \to E^*\{\log g(y_1|x_1^*)\} \text{ a.s. } [P^*], \tag{33}
\]

noting that \(\{x_t^*\}\) has the same stationary distribution as \(\{x_t\}\) in this case. By (31)-(33), \(n^{-1} \log w_n\) converges a.s. \([P^*]\) to a nonrandom constant.

For the case \(q_t(x_t|x_{t-1}) = f(x_t|x_{t-1})g(y_t|x_t)/p(y_t|x_{t-1})\), \(\log w_n = \sum_{i=1}^{n} \log p(y_i|x_{t-1}^*)\). Making use of stationarity and the ergodic theorem, it can be shown that

\[
n^{-1} \log w_n \to E^*\{\log p(y_1|x_0^*)\} \text{ a.s. } [P^*], \tag{34}
\]

in which \(x_0^*\) is initialized given the observations \(y_0, y_{-1}, \ldots\).

Let \(Z_n = \log(w_n/w_{n-1})\). Since \(\{w_t\}\) is a martingale under \(Q\) and since \(Z_n\) is nondegenerate, Jensen’s inequality yields \(E_Q Z_n < 0\), implying that \(E^* Z_n < 0\). Moreover, \(Z_n\) is asymptotically
stationary, so the nonrandom limit of $n^{-1} \log w_n$ established in (34) is negative. Hence $w_n \to 0$ a.s. $[P^*]$ while $E^*(w_n) = E(E_Q(w_n | Y_{1:n})) = 1$, from which it follows that $E(\text{var}_Q(w_n | Y_{1:n})) \to \infty$.

Remark: The preceding proof has assumed finiteness of certain expectations to ensure (31), (33) and (34). Specifically, the integrability assumptions of Theorem 1 used in the proof are:

\begin{align*}
\sup_{n \geq 1} E_\pi \{ |\log p(y_{n+1} | Y_{1:n})| \} &< \infty, \\
\int E_\pi \{ |\log g(y_1 | x)\} d\pi(x) &< \infty, \\
\sup_{n \geq 1} E_\pi \{ E_Q[|\log p(y_{n+1} | x_n^*)| | Y_{1:n})] \} &< \infty,
\end{align*}

in which $E_\pi$ denotes expectation under which $x_0$ is initialized at the stationary distribution $\pi$.

**APPENDIX B: PROOF OF THEOREMS 2 AND 3**

**Proof of Theorem 2:** We shall make repeated use of the drift condition (V4), Lemma 15.2.8 and Theorem 16.0.1 of [16, pp. 367, 370, 383]. First since $\{u_t\}$ is bounded, it clearly satisfies (V4) (with $V(u) = |u| \vee 1$). Moreover, $E(\|\theta_t\| | \theta_{t-1}) \leq (1 - \rho)\|\theta_{t-1}\| + \rho E|W_t|$ and therefore $\{\theta_t\}$ also satisfies (V4) by Lemma 15.2.8 of [16]. Let $y_t = (y_t, \ldots, y_{t-k+1})^T$. Since

$$y_t = -a_1, y_{t-1} - \cdots - a_{k,t} y_{t-k} + \{b_1, t u_{t-1} + \cdots + b_{k,t} u_{t-k} + \epsilon_t$$

and since $\{\theta_t\}$ is a jump process independent of $\{\epsilon_t\}$ and satisfies the stability assumption (19), it then follows that $E(\|y_t\| | y_{t-1}, \theta_{t-1}) \leq \lambda \|y_{t-1}\| + B + E|\epsilon_t|$ for some $0 < \lambda < 1$ and $B > 0$. Hence by Lemma 15.2.8 of [16], $\{y_t\}$ also satisfies (V4) with $V(y) = \|y\| \vee 1$. This shows that $(\theta_t, \varphi_{t+1})$ satisfies (V4) with $V(\theta, \varphi) = \|\theta\| \vee \|y\| \vee \|u\| \vee 1$, writing $\varphi = (y^T, u^T)^T$. Since $(\theta_t, \varphi_{t+1})$ is irreducible and aperiodic, it then follows from Theorem 16.0.1 of [16] that $(\theta_t, \varphi_{t+1})$ has a stationary distribution $\pi$ and converges weakly to $\Pi$ exponentially fast as $t \to \infty$. We next consider $\delta_t^{(B)}$, which is the mean of the posterior distribution $\nu_t$ of $\theta_t$ given $(y_s, \varphi_s), s \leq t$. (We ignore the constraint (19) in computing the posterior distribution, as mentioned earlier.) Let $\mathcal{F}$ be the set of measures $\nu$ on $\mathbb{R}^{k+h}$ such that $\int \|\theta\| d\nu(\theta) < \infty$. By making use of (V4) and a time-shift technique involving the stationarity of $(\theta_t, \varphi_{t+1})$ similar to that used to derive (31), it is shown in [25] that $\nu_t$ converges weakly, with an exponential rate of convergence, in the metric space $(\mathcal{F}, \omega)$, where $\omega$ is the metric $\omega(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|_\omega$ associated with the weighted variation norm $\|\mu\|_\omega = \int \|\theta\| d|\mu|$ of signed measures $\mu$, in which $|\mu|$ denotes the total variation of $\mu$. Hence the desired conclusion on $\delta_t^{(B)}$ follows.
Remark: Since \( \hat{\theta}_t^{(W)} = \mu_{t-k+1,t} \), where \( \mu_{s,t} \) is defined in (14), it follows from Theorem 2 that \( \hat{\theta}_t^{(W)} - \theta_t \) has a limiting distribution as \( t \to \infty \). Whereas \( \hat{\theta}_t^{(B)} \) uses a single window of size \( k \), \( \hat{\theta}_t^{(B)} \) is a weighted average of \( \mu_{s,t} \) for \( s \leq t \), with the weights given by the posterior distribution of the most recent change-time \( C_t \) up to time \( t \). If \( C_t \) were known, then the optimal estimate of \( \theta_t \) would be \( \mu_{C_t,t} \). In ignorance of \( C_t \), \( \hat{\theta}_t^{(B)} \) approximates \( \mu_{C_t,t} \) by a weighted average of \( \mu_{s,t} \), whose weights can be used to locate \( C_t \). It is clearly more efficient than \( \mu_{t-k+1,t} \) that uses a single window of size \( k \) which does not involve \( \rho \). As \( \rho \to 0 \), change-points occur infrequently and the results of [26] suggest that \( C_t \) can be tracked with a detection delay of \( O(\log \rho^{-1}) \). This suggests that

\[
E_{\Pi} \left\{ \sum_{i=C_t}^{t} \| \hat{\theta}_i^{(B)} - \theta_i \|^2 \right\} = E_{\Pi} \left\{ O(\log \rho^{-1}) + O(\sum_{i=1}^{t-C_t} i^{-1}) \right\},
\]

where the first term on the right hand side corresponds to the squared bias caused by the detection delay and the second term corresponds to the variances of least squares estimates based on samples of size \( i \) (\( 1 \leq i \leq t - C_t \)), noting that \( E_{\Pi} \varphi_i \varphi_i^T \) is positive definite. Moreover, under the stationary distribution \( \Pi \), \( t - C_t \) has an approximately geometric distribution with mean \( \rho^{-1} \) for sufficiently small \( \rho \). It then follows from (38) that \( E_{\Pi} \| \hat{\theta}_i^{(B)} - \theta_i \|^2 = O(\rho \log \rho^{-1}) \) as \( \rho \to 0 \). On the other hand, \( E_{\Pi} \| \hat{\theta}_i^{(W)} - \theta_i \|^2 \) does not converge to 0 as \( \rho \to 0 \) since the window size \( k \) is fixed and is unrelated to the expected duration of \( \rho^{-1} \) for \( t - C_t \). Even if we choose \( k \sim \rho^{-1} \), because \( t - C_t \) is approximately normal with mean \( \rho^{-1} \) and standard deviation \( \rho^{-1/2} \) under \( \Pi \) when \( \rho \) is small, the squared bias of \( \hat{\theta}_i^{(W)} \) exceeds \( c\sqrt{\rho} \) for some \( c > 0 \) and therefore \( E_{\Pi} \| \hat{\theta}_i^{(W)} - \theta_i \|^2 \geq c\sqrt{\rho} \) as \( \rho \to 0 \).

The forgetting factor least squares estimate \( \hat{\theta}_t^{(F)} \), which uses a forgetting factor \( p \) instead of the fixed window size \( k \), behaves like the sliding window estimate \( \hat{\theta}_t^{(W)} \). In particular, as \( \rho \to 0 \), \( E_{\Pi} \| \hat{\theta}_i^{(F)} - \theta_i \|^2 \) does not converge to 0 for fixed \( p \). If we choose \( p \sim \rho \), then \( E_{\Pi} \| \hat{\theta}_i^{(F)} - \theta_i \|^2 \geq c\sqrt{\rho} \) as \( \rho \to 0 \) because of the asymptotic normality of \( t - C_t \) with standard deviation \( \rho^{-1/2} \) under \( \Pi \) as \( \rho \to 0 \). The order \( O(\rho \log \rho^{-1}) \) of \( E_{\Pi} \| \hat{\theta}_i^{(B)} - \theta_i \|^2 \) is considerably smaller than the \( \sqrt{\rho} \) order for the mean squared errors of \( \hat{\theta}_i^{(W)} \) and \( \hat{\theta}_i^{(B)} \) (with window size or forgetting factor chosen in an optimal way depending on \( \rho \)).

Proof of Theorem 3 (i): From (13) and (25), it follows that

\[
y_{t+1} = -a_{t+1}y_t - \cdots - a_{k,t+1}y_{t-k+1} + y^*_t + (E_t \theta_{t+1}^T)\varphi_{t+1} + \epsilon_{t+1} = (\theta_{t+1}^T - E_t \theta_{t+1}^T)\varphi_{t+1} + y^*_t + \epsilon_{t+1},
\]

recalling that \( \varphi_{t+1} = (-y_t, \ldots, -y_{t-k+1})^T \). Therefore \( \varphi_{t+2} = A_{t+1}\varphi_{t+1} - (y^*_{t+1} + \epsilon_{t+1}, 0, \ldots, 0)^T \). Since \( \lim_{t \to \infty} E_t \| A_{t+1} \| < 1 \) with probability 1, we can apply Theorem 16.0.1 of [16] as in the proof of Theorem 2 to derive the desired conclusion.
Proof of Theorem 3 (ii): From (13) and (23), it follows that analogous to (39),

\[
y_{t+1} = \{\alpha_{t+1}^T - b_{1,t+1}E_t(b_{1,t+1}\alpha_{t+1}^T)/E_t(b_{1,t+1}^2)\}(y_t, \ldots, y_{t-k+1})^T
\]

\[
+ \beta_{t+1}^T - b_{1,t+1}E_t(b_{1,t+1}\beta_{t+1}^T)/E_t(b_{1,t+1}^2)\}(u_{t-1}, \ldots, u_{t-h+1})^T
\]

\[
+ b_{1,t+1}E_t(b_{1,t+1})y_{t+1}^*/E_t(b_{1,t+1}^2) + \epsilon_{t+1}.
\]

In view of the definition of \( B_{t+1} \) and \( \psi_t \), this implies that

\[
\psi_{t+1} = B_{t+1}\psi_t - (b_{1,t+1}E_t(b_{1,t+1})y_{t+1}^*/E_t(b_{1,t+1}^2) + \epsilon_{t+1}, 0, \ldots, 0)^T.
\]

Since \((E_t b_{1,t+1})^2/E_t(b_{1,t+1}^2) \leq 1\) and \(y_{t+1}^*\) is bounded, the proof is similar to that of Theorem 3(i).

References


TABLE I
EFFECT OF CV² BOUND ON PERFORMANCE OF SISR FOR MEAN SHIFT MODEL

<table>
<thead>
<tr>
<th>cv² bound</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resampling #</td>
<td>999</td>
<td>85</td>
<td>51</td>
<td>32</td>
<td>22</td>
<td>14</td>
<td>10</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>SSE</td>
<td>240.38</td>
<td>205.24</td>
<td>188.06</td>
<td>193.24</td>
<td>196.19</td>
<td>206.01</td>
<td>217.51</td>
<td>208.75</td>
<td>603.94</td>
</tr>
<tr>
<td>s.e.</td>
<td>9.65</td>
<td>6.53</td>
<td>6.73</td>
<td>6.31</td>
<td>6.53</td>
<td>6.69</td>
<td>8.27</td>
<td>7.28</td>
<td>28.11</td>
</tr>
</tbody>
</table>

Table II
COMPARISON OF SSE FOR SISR AND EXACT BAYES ESTIMATE

<table>
<thead>
<tr>
<th>ρ</th>
<th>0.001</th>
<th>0.003</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISR</td>
<td>165.30</td>
<td>358.55</td>
<td>802.41</td>
</tr>
<tr>
<td>Bayes</td>
<td>148.73</td>
<td>341.96</td>
<td>775.11</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.11</td>
<td>1.05</td>
<td>1.04</td>
</tr>
</tbody>
</table>

TABLE III
COMPARISON OF PERFORMANCE OF SISR, SISR*, SLIDING WINDOW (SW) WITH LENGTH k, AND RECURSIVE LEAST SQUARES (RLS) WITH FORGETTING FACTOR p

<table>
<thead>
<tr>
<th>Estimator</th>
<th>1-500</th>
<th>501-1000</th>
<th>1001-2000</th>
<th>2001-3000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSE₁</td>
<td>SSE₂</td>
<td>SSE₁</td>
<td>SSE₂</td>
</tr>
<tr>
<td>SW, k = 100</td>
<td>65.8</td>
<td>52.7</td>
<td>74.2</td>
<td>53.2</td>
</tr>
<tr>
<td>SW, k = 50</td>
<td>71.6</td>
<td>49.3</td>
<td>58.8</td>
<td>40.9</td>
</tr>
<tr>
<td>RLS, p = .99</td>
<td>82.3</td>
<td>63.6</td>
<td>103.8</td>
<td>80.1</td>
</tr>
<tr>
<td>RLS, p = .95</td>
<td>65.5</td>
<td>53.4</td>
<td>49.1</td>
<td>33.9</td>
</tr>
<tr>
<td>SISR</td>
<td>25.5</td>
<td>21.6</td>
<td>12.5</td>
<td>10.6</td>
</tr>
<tr>
<td>SISR*</td>
<td>27.5</td>
<td>24.9</td>
<td>20.7</td>
<td>18.1</td>
</tr>
</tbody>
</table>

27
### TABLE IV
**COMPARISON OF RULE (23) WITH CERTAINTY EQUIVALENCE RULE (21) BASED ON SLIDING WINDOW (SW) ESTIMATES (WITH WINDOW SIZE \( k \)) OR RECURSIVE LEAST SQUARES (RLS) ESTIMATES (WITH FORGETTING FACTOR \( p \)) AND WITH THAT BASED ON SISR**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>1-500</th>
<th>501-1000</th>
<th>1001-2000</th>
<th>2001-3000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSE(_1)</td>
<td>SSE(_3)</td>
<td>SSE(_1)</td>
<td>SSE(_3)</td>
</tr>
<tr>
<td>SW, ( k = 100 )</td>
<td>150.9</td>
<td>9876.3</td>
<td>129.7</td>
<td>7.6 \times 10^4</td>
</tr>
<tr>
<td>SW, ( k = 50 )</td>
<td>36.8</td>
<td>7.2 \times 10^4</td>
<td>47.1</td>
<td>2.3 \times 10^6</td>
</tr>
<tr>
<td>RLS, ( p = .99 )</td>
<td>27.7</td>
<td>176.4</td>
<td>1191.0</td>
<td>367.0</td>
</tr>
<tr>
<td>RLS, ( p = .95 )</td>
<td>17.4 \times 10^4</td>
<td>4.1 \times 10^4</td>
<td>1.6 \times 10^7</td>
<td>6.1 \times 10^7</td>
</tr>
<tr>
<td>SISR: (21)</td>
<td>19.9</td>
<td>148.4</td>
<td>9.1</td>
<td>152.2</td>
</tr>
<tr>
<td>SISR: (23)</td>
<td>18.4</td>
<td>129.2</td>
<td>9.8</td>
<td>50.5</td>
</tr>
</tbody>
</table>

\( (a) \) \( \rho = 0.001 \)

<table>
<thead>
<tr>
<th></th>
<th>1-500</th>
<th>501-1000</th>
<th>1001-2000</th>
<th>2001-3000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSE(_1)</td>
<td>SSE(_3)</td>
<td>SSE(_1)</td>
<td>SSE(_3)</td>
</tr>
<tr>
<td>SW, ( k = 100 )</td>
<td>171.4</td>
<td>8594.6</td>
<td>252.5</td>
<td>8.4 \times 10^4</td>
</tr>
<tr>
<td>SW, ( k = 50 )</td>
<td>139.0</td>
<td>1.7 \times 10^4</td>
<td>186.1</td>
<td>1.9 \times 10^4</td>
</tr>
<tr>
<td>RLS, ( p = .99 )</td>
<td>285.5</td>
<td>2278.3</td>
<td>114.9</td>
<td>300.0</td>
</tr>
<tr>
<td>RLS, ( p = .95 )</td>
<td>826.3</td>
<td>2221.4</td>
<td>1201.9</td>
<td>1.6 \times 10^4</td>
</tr>
<tr>
<td>SISR: (21)</td>
<td>57.3</td>
<td>237.2</td>
<td>54.4</td>
<td>269.2</td>
</tr>
<tr>
<td>SISR: (23)</td>
<td>50.5</td>
<td>195.9</td>
<td>44.8</td>
<td>272.0</td>
</tr>
</tbody>
</table>

\( (b) \) \( \rho = 0.01 \)

### TABLE V
**PERFORMANCE OF (23) FOR UNSTABLE SYSTEMS**

<p>| ( |a_{1,t}| \leq 1.1 ) | 1-500 | 501-1000 | 1001-2000 | 2001-3000 |
|-------------------|-------|----------|-----------|-----------|
|                   | SSE(_1) | SSE(_3) | SSE(_1) | SSE(_3) | SSE(_1) | SSE(_3) | SSE(<em>1) | SSE(<em>3) |
| ( (a) ) ( \delta_t = 1/4, \omega_t \sim N(0,1/4) ) | ( \geq 10^{308} ) | 2.0 \times 10^{11} | &gt; 10^{308} | 1.8 \times 10^{21} | &gt; 10^{308} | 2.7 \times 10^{38} | &gt; 10^{308} | 1.5 \times 10^{107} |
| ( (b) ) ( \delta_t = 0.04, \omega_t \sim N(0,0.04) ) | 11.7 | 67.2 | 8.3 | 102.8 | 17.9 | 103.7 | 14.0 | 147.2 |
| ( |a</em>{1,t}| \leq 1.1 ) | 11.0 | 76.8 | 8.4 | 77.5 | 13.8 | 158.4 | 13.6 | 273.2 |
| ( |a</em>{1,t}| \leq 1.3 ) | 11.3 | 111.0 | 7.9 | 127.8 | 17.5 | 248.1 | 14.9 | 188.6 |</p>
<table>
<thead>
<tr>
<th>Input Type</th>
<th>n=200</th>
<th>n=400</th>
<th>n=600</th>
<th>n=800</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_t$</td>
<td>$a$</td>
<td>$x_t$</td>
<td>$a$</td>
<td>$x_t$</td>
</tr>
<tr>
<td>a</td>
<td>.808</td>
<td>.026</td>
<td>.810</td>
<td>.020</td>
<td>.809</td>
</tr>
<tr>
<td>b</td>
<td>.843</td>
<td>.016</td>
<td>.835</td>
<td>.012</td>
<td>.833</td>
</tr>
<tr>
<td>c</td>
<td>.852</td>
<td>.014</td>
<td>.852</td>
<td>.011</td>
<td>.849</td>
</tr>
<tr>
<td>d</td>
<td>.837</td>
<td>.006</td>
<td>.834</td>
<td>.004</td>
<td>.834</td>
</tr>
</tbody>
</table>

**TABLE VI**

ROOT MEAN SQUARED ERRORS IN ESTIMATING $x_t$ AND $a$
Figure 1: True vs. estimated signals for Example 4. True signal: solid lines. Estimates: dots.