Automatica 142 (2022) 110376

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Filtering with degenerate observation noise: A stochastic approximation approach[☆]

Hongjiang Qian[a], Qing Zhang[b], George Yin[a,*]

[a] Department of Mathematics, University of Connecticut, Storrs, CT 06269, United States of America
[b] Department of Mathematics, University of Georgia, Athens, GA 30602, United States of America

ARTICLE INFO

Article history:
Received 9 October 2021
Received in revised form 1 February 2022
Accepted 1 April 2022
Available online xxxx

Keywords:
Filtering
Degenerate observation noise
Stochastic approximation

ABSTRACT

This paper is concerned with a general filtering scheme of a continuous-time dynamic system in which the state is not completely observable. The observation process consists of a function of the state with additive noise. Typically, such noise is assumed to be non-degenerate. In this case, various filtering schemes can be developed. For example, in a linear case, the Kalman–Bucy (KB) filter applies and leads to a recursive filtering equation for the conditional expectation of the state given the observation up to time t. Nevertheless, in applications, only some state variables are directly observable and the rest are not. This gives rise to filtering with degenerate observation noise. In this case, traditional filtering schemes fail. This paper develops a viable scheme to address possible degenerate observation noise. In this paper, we propose a recursive filtering equation in which the gain matrix is a matrix-valued parameter to be determined. We adopt the Monte Carlo training procedure used for deep filtering to determine the best gain matrix. In particular, given the state and observation equations, we generate their Monte Carlo samples. Given a gain matrix, we generate the corresponding state estimation samples, which leads to the error function of the state and its estimation. The problem is to choose the gain matrix to minimize the error function. We develop a stochastic approximation method for such a minimization task; we term the procedure the SA filter, where SA stands for stochastic approximation. We focus on computational experiments and demonstrate the performance of the SA filter and its robustness. We also compare the SA filter with the (extended) Kalman–Bucy filter and the deep filter in both linear and nonlinear models.

1. Introduction

This paper is devoted to a continuous-time filtering problem with possible degenerate observation noise. There are many real-world applications of state estimation and filtering including maneuvered target tracking, speech recognition, telecommunication, and financial engineering. Filtering is concerned with dynamics in which the state variables are not completely observable. The traditional approach is to derive estimators based on observation using the least squares criteria. Under the setup of Gaussian distributions, the corresponding filtering problem is to find the conditional mean of the state given the observation up to time t. The best known filter is the Kalman–Bucy (KB) filter in linear models. We refer the reader to Fleming and Rishel (1975) for details.

Early development in nonlinear filtering can be found in Duncan (1967) focusing on conditional densities for diffusion processes, Mortensen (1968) for the most probable trajectory approach, Kushner (1964) for nonlinear filtering equations, and Zakai (1969) for unnormalized equations.

For recent progress on general filtering, we refer the reader to Frey, Schmidt, and Xu (2018) and Gao and Tembine (2016). In Frey et al. (2018), the authors used Galerkin's approximation to solve a Zakai equation, whereas in Gao and Tembine (2016), the authors considered distributed mean-field filters for traffic networks and developed a scheme decomposing the entire state space into subspaces and performing the distributed filters independently. Their main effort was still on developing approximation methods of infinite-dimensional filtering equations. We note that the difficulty of using the conditional distribution based filtering is the underlying stochastic differential equations are infinite dimensional. Thus the aforementioned methods still have to deal with the inherent 'curse of dimensionality'. This makes the filtering very difficult and challenging for general nonlinear systems.

[☆] The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Wei Xing Zheng under the direction of Editor Torsten Söderström.

[*] Corresponding author.
E-mail addresses: hongjiang.qian@uconn.edu (H. Qian), qz@uga.edu (Q. Zhang), gyin@uconn.edu (G. Yin).

https://doi.org/10.1016/j.automatica.2022.110376
0005-1098/© 2022 Elsevier Ltd. All rights reserved.

In this paper, we study the filtering problem and focus on finite dimensional filters. In particular, we consider the model in which the state X_t is not completely observable. The observation Y_t is given by a function of X_t with additive noise. In the literature, typically or almost always, the observation noise is assumed to be non-degenerate in the sense that the product of the noise coefficient matrix and its transpose is invertible. In this case, various filtering schemes can be developed. For instance, in a linear model, the Kalman–Bucy filter applies and leads to a recursive filtering equation for conditional expectation of X_t given the observation up to time t. Nevertheless, in applications, often some components of X_t are directly observable and the rest are not. To begin, we consider several examples.

Example 1. Consider a stochastic acceleration problem, which is well known in dynamic systems and statistical physics; see Kesten and Papanicolaou (1980/81), Nguyen and Yin (2021), and references therein. For simplicity, we consider a real-valued process representing the stochastic acceleration. Let X_t be the position of a particle on the real line with t being the time. Then $\dot{X}_t$ is the velocity and $\ddot{X}_t$ is the acceleration. Use $a(\cdot) : [0, T] \mapsto \mathbb{R}$ to denote a nonlinear continuous function. Suppose that the particle is subject to stochastic disturbances, which are unavoidable in reality. Denote the source of disturbances or the driving noise by W_t, a Brownian motion with intensity $\sigma(X_t)$. We can then represent the stochastic acceleration using the following differential equation $\ddot{X}_t = a(X_t)dt + \sigma(X_t)\dot{W}_t$, where $\dot{W}_t$ denote the formal time derivative of a standard Brownian motion W_t. Define $X_t^1 = X_t$ and $X_t^2 = \dot{X}_t^1$. Interpreting the above in the usual sense of stochastic calculus, for the vector $X_t = (X_t^1, X_t^2)' \in \mathbb{R}^2$ (with z' denoting the transpose of z) and $\widetilde{W}_t = (0, W_t)$, we have

$$dX_t = \begin{bmatrix} X_t^2 \\ a(X_t^1) \end{bmatrix} dt + \begin{bmatrix} 0 & 0 \\ 0 & \sigma(X_t^1) \end{bmatrix} d\widetilde{W}_t.$$

Assume the velocity X_t^2 is observable. Then, the observation equation is given by $dY_t = [X_t^1, X_t^2]'dt + [1, 0]'dV_t$, where V_t is a Brownian motion independent of $\widetilde{W}_t$. The covariance matrix of the observation noise is given by $[1, 0]'[1, 0]$, which is degenerate. In this example, the covariance matrix of the state equation is also degenerate.

Example 2. Denote the position of a particle in a fluid by X_t with $X_t \in \mathbb{R}^d$, and its velocity by $\xi_t = (d/dt)X_t$. Consider the random movement of the particle in a fluid due to collisions with the molecules of the fluid, whose state is given by $m\frac{d\xi_t}{dt} = -\lambda\xi_t + \dot{W}_t$, where m is the mass, λ is the friction coefficient, and W_t is a standard d-dimensional Brownian motion. However, the velocity is not directly observable, rather we can only observe the position of the particle with noise that is described by a chemical Langevin equation $\ddot{X}_t = b(X_t) + \sigma(X_t)\dot{V}_t$, with $X_0 = x_0 \in \mathbb{R}^d$ and $\dot{X}_0 = \xi_0 \in \mathbb{R}^d$. $b(x) + \sigma(x)\dot{V}_t$ represents a particle in the force field, and V_t is an m-dimensional standard Brownian motion and $\dot{V}_t$ is its formal derivative. Assume that W_t and V_t are independent. Although the state is non-degenerate, the observation, similar to the stochastic acceleration model, is degenerate. For details of a more general model, see Nguyen and Yin (2021). Note that in that reference, an additional small parameter ε is included to represent the strong damping.

Example 3. Building on the research of Ross and Hudson, "compartmental" epidemic models were first introduced by Kermack and McKendrick (1927) in a series of three papers (known as the "trilogy"). Then the study on mathematical models has flourished. Much attention has been devoted to analyzing, predicting the spread, and designing controls of infectious diseases in host populations. The so-called SIR models have received much attention; they have been used in a wide range of applications and also had much influence on the COVID-19 modeling. The SIR models subdivide the population into Susceptible S_t, Infected I_t, and Removed R_t classes leading to

$$\begin{cases} dS_t = \big(\alpha_t - \beta_t S_t I_t - \mu S_t\big)dt + \sigma_{1,t} S_t dW_t \\ dI_t = \big(\beta_t S_t I_t - (\mu_t + \rho_t + \gamma_t) I_t\big)dt + \sigma_{2,t} I_t dW_t \\ dR_t = (\gamma_t I_t - \mu_t R_t)dt + \sigma_{3,t} R_t dW_t. \end{cases} \tag{1.1}$$

Note that all three components are subject to the same Brownian motion perturbation reflecting that the noise comes from the same source. Clearly, this is a degenerate diffusion model. For the meaning of the various parameters, we refer to Dieu, Du, Nguyen, and Yin (2016) for further details. In fact, there has been much effort on studying a more general class of nonlinear stochastic models known as Kolmogorov systems; see the most recent results (Nguyen, Nguyen, & Yin, 2021). Assume that each variable can be observed with the same additive noise. That is, the observation Y_t is given by $dY_t = [S_t, I_t, R_t]'dt + [1, 1, 1]'dV_t$, with V_t being a standard real-valued Brownian motion independent of W_t. It is readily seen that the observation noise is degenerate.

In the above examples, the observation noise is degenerate. Thus, the traditional filtering schemes fail. It is the purpose of this paper to develop a viable filtering scheme to address possible degenerate observation noise.

Recently, a deep neutral network (DNN) based filtering was developed in Wang, Yin, and Zhang (2021). The idea was to generate Monte Carlo samples from the given model and make use of these samples to train a deep neutral network. The observation process from the Monte Carlo is used as the inputs to the DNN and the state is used as the target. A least squares loss function of the target and calculated output is used for the neutral network training in order to obtain the corresponding weight vectors. Then these weight vectors are applied to a separate set of Monte Carlo samples from of an actual dynamic model. The corresponding process is called the deep filter (DF). It is shown in Wang et al. (2021), the deep filter compares favorably with the traditional Kalman filter in linear cases and the extended Kalman filter in nonlinear cases. In addition, the deep filter can deal with models involving random switching processes.

In this paper, to address possible degenerate observation noise, we consider a recursive equation for the state estimation in which the gain is a matrix-valued parameter to be estimated. We adopt the Monte Carlo training procedure in deep filter to determine the best gain matrix. In particular, given the state and observation equations, we generate their Monte Carlo samples. Given a gain matrix, we generate the corresponding state estimation samples. This leads to the error function of the state and its estimation. The goal is to choose the gain matrix to minimize the error function.

We use a stochastic approximation method to search for the optimal gain in connection with state estimations. The algorithms are of stochastic gradient descent type. A comprehensive study of asymptotic properties of stochastic approximation methodologies can be found in Kushner and Yin (2003).

The main contribution of this paper is the development of a novel approach for general SA filtering of systems with possible degenerate observation noise. Such an approach provides a recursive form and is shown to be computationally comparable with the deep filtering methods. The recursive form is desirable from an application point of view in construct to the 'black-box' type filtering of the deep filtering. In addition, the SA filter is robust when the actual observation noise is higher than that of the nominal model.

The rest of the paper is arranged as follows. In the next section, we present the model under consideration and the corresponding SA filter. Section 3 presents the asymptotic results. Numerical experiments are reported in Section 4. Finally, some concluding remarks are given in Section 5.

2. The model

Let $X_t \in \mathbb{R}^{d_1}$ denote the state process satisfying the stochastic differential equation

$$dX_t = f(X_t)dt + \sigma dW_t,\ X_0 = x, \tag{2.2}$$

where $f(x)$ is a function of X_t, σ is a matrix of appropriate dimensions, and W_t is an $\mathbb{R}^{d_2}$ Brownian motion. We consider the case that X_t is not fully observable. Assume a function of X_t is observable with additive noise so that the observation process $Y_t \in \mathbb{R}^{d_3}$ is given by the equation

$$dY_t = h(X_t)dt + \sigma_1 dV_t, \tag{2.3}$$

where h is a function of X_t, σ_1 is a matrix of suitable dimensions, and $V_t \in \mathbb{R}^{d_4}$ is a Brownian motion. When $\sigma_1\sigma_1'$ is invertible, various filtering schemes can be developed. For example, in a linear model, Kalman–Bucy filter renders a recursive equation for the conditional expectation of X_t given the observation up to time t. Nevertheless, in applications, very often, only some components of X_t are observable and the rest are not, which leads to degenerate observation noise with singular $\sigma_1\sigma_1'$. In this case, the existing filtering schemes fails. A main objective of this paper is to develop a viable filtering scheme to address possible degenerate observation noise.

3. Asymptotic results

Let $\widehat{X}_t$ denote an estimate of X_t. We consider the filtering scheme of the form

$$d\widehat{X}_t = f(\widehat{X}_t)dt + R(dY_t - h(\widehat{X}_t)dt),\ \widehat{X}_0 = EX_0, \tag{3.4}$$

where R is a constant gain $(d_1 \times d_2)$ matrix to be determined. Based on Monte Carlo simulations and stochastic approximation methods, our new approach converts the filtering problem to a stochastic optimization problem to estimate the best gain R. Specifically, for each fixed R, we generate Monte Carlo samples of X_t, Y_t and $\widehat{X}_t$ from (2.2), (2.3), and (3.4). Using these sample paths, we can define an error function of R

$$J(R) = E\int_0^T |X_t - \widehat{X}_t|^2 dt, \tag{3.5}$$

for a given T. However, when we carry out the simulation, we are taking samples. Thus in lieu of $J(R)$, we are using a noisy sampled error $\widetilde{J}(R,\zeta)$ where ζ denotes the noise appeared in the sample. Then we use a stochastic approximation approach to search for R that minimizes the error. In particular, let R_n denote the approximation sequence given by the recursive equation

$$R_{n+1} = R_n - \varepsilon_n D\widetilde{J}(R_n, \zeta_n), \tag{3.6}$$

where the finite difference

$$D\widetilde{J}(R_n, \zeta_n) = \frac{1}{2\delta_n}\left(\widetilde{J}(R_n + \delta_n, \zeta_n^+) - \widetilde{J}(R_n - \delta_n, \zeta_n^-)\right),$$

approximates the gradient of $J(R)$ with $\zeta_n^\pm$ denotes two different noise processes, and $\varepsilon_n > 0$, $\delta_n > 0$ satisfying $\sum \varepsilon_n = \infty$, $\sum \delta_n = \infty$, and $\varepsilon_n/\delta_n \to 0$ as $n \to \infty$. The ε_n is the stepsize for the stochastic approximation, whereas δ_n is the stepsize for finite difference; see Kushner and Yin (2003) for different choices of these stepsize.

Remark 4. Note that (3.6) is a stochastic approximation algorithm for approximating a matrix-valued parameter R. A proper way of writing the recursive formula is to use a vectorization of R by piling up the columns of R. That is, $\widehat{R} = \text{vec}(R^1, \ldots, R^{d_2}) \in \mathbb{R}^{(d_1 d_2)\times 1}$ (a $(d_1 d_2)$ column vector), where R^i is the ith column of R. Next, define $\widetilde{J}_i^\pm = \widetilde{J}(r_{\ell,i} \pm \delta_\ell e_i, \zeta_\ell^\pm)$ and the central finite difference $[\widetilde{J}_i^+ - \widetilde{J}_i^-]/(2\delta_\ell)$ and so on, where e_i is the standard unit vector. Then we can write the recursion for $\widehat{R}_\ell$ either componentwise or in a vector form. However, for notational simplicity, we decided to use (3.6) instead. We name such a state estimation procedure *stochastic approximation based filter* or in short *SA filter*. In what follows, we carry out numerical experiments to examine the performance of our stochastic approximation based filtering scheme. Before proceeding further, we state a convergence theorem.

Theorem 5. *Consider (3.6). Suppose that*

- $\varepsilon_n \to 0$, $\delta_n \to 0$, *and* $\varepsilon_n/\delta_n \to 0$ *as* $n \to \infty$, $\sum_n \varepsilon_n = \infty$, $\sum_n \delta_n = \infty$, *but* $\sum_n (\varepsilon_n/\delta_n)^2 < \infty$;
- $\{\zeta_n\}$ *is a stationary* ϕ*-mixing sequence such that for each* R, $E\widetilde{J}(R,\zeta_n) = J(R)$ *and that for each* R, $E|\widetilde{J}(R,\zeta_n)|^2 < \infty$;
- *the differential equation*

$$\dot{R} = -\nabla J(R),\ R(0) = R_0 \tag{3.7}$$

has a unique solution for each initial condition R_0;

- *the differential equation (3.7) has a unique stationary point* R^*, *which is stable in the sense of Lyapunov.*

Then the sequence $\{R_n\}$ *converges weakly (also with probability one) to* R^* *as* $n \to \infty$.

Proof. The proof of the above theorem is essentially contained in the book of Kushner and Yin (2003). We only make some brief comments. The main idea is the use of the so-called ordinary differential equation approach, which relies on the interplay of the discrete-time iterations and continuous-time ordinary differential equation (3.7). Define $t_n = \sum_{j=0}^{n-1} \varepsilon_j$, $m(t) = \max\{n : t_n \le t\}$, and the piecewise constant interpolation $R^0(t) = R_n$ for $t \in [t_n, t_{n+1})$ and the shift sequence $R^n(t) = R^0(t + t_n)$. Note that $m(t)$ is just a "look back" map indicating what is the discrete iteration number associated with the continuous time t. The limit ordinary differential equation is obtained first, and then the convergence to the stationary point R^* is derived using stability of the ordinary differential equation. A crucial step is averaging. The rational is that R_n varies relatively slowly compared to the noise in $\{D\widetilde{J}(R,\zeta_n)\}$. Choose $\Delta_n > 0$ with $\Delta_n \to 0$ as $n \to \infty$ such that $\lim_n \sup\{\varepsilon_j : j \ge n\}/\Delta_n \to 0$. For each n choose an increasing sequence $m_{n,1} < m_{n,2} < \cdots$ (with $m_{n,1} = n$) such that $\sum_{j=m_{n,l}}^{m_{n,l+1}-1} \varepsilon_j/\Delta_n = 1 + o(1)$ with $o(1) \to 0$ as $n \to \infty$ so that $(t_{m_{n,l+1}} - t_{m_{n,l}})/\Delta_n \to 1$ as $n \to \infty$ uniformly in l. For a fixed R, the noise is averaged out in that

$$\frac{1}{\Delta_n}\sum_{j=m_{n,l}}^{m_{n,l+1}-1} \varepsilon_j D\widetilde{J}(R,\zeta_j) \to \nabla J(R) \ \text{ as } \ n \to \infty. \tag{3.8}$$

Note that in the above, we need the convergence to hold in the sense of in probability, which follows from the fact that $\{D\widetilde{J}(R,\zeta_n)\}$ is a stationary mixing sequence. In fact the limit also holds in the sense of with probability one because mixing implies ergodicity. Then the ordinary differential equation (3.7) can be obtained. The convergence of $\{R_n\}$ follows from the argument in Kushner and Yin (2003, Chapter 8) (see also Kushner & Yin, 2003, Chapter 6); the details are omitted, however.

Remark 6. The conditions we used are not the weakest but it is good enough for our purposes. The use of stationary mixing sequence instead of independent and identically distributed random variables enables one to use correlated random seeds in simulation so as to do something like variance reduction. In lieu of stationary mixing, we can simply require (3.8) holding. The results above are based on a sequence of decreasing stepsizes $\{\varepsilon_n\}$. A constant stepsize algorithm can be used with certain modifications. In such a case, the pertinent notion of convergence is in the sense of weak convergence.

4. Numerical experiments

In this section, we study several representative examples and examine the performance of the SA filter given in (3.4)–(3.6). We compare the SA filter with the Kalman–Bucy filter given in Fleming and Rishel (1975) and the deep filter developed in Wang et al. (2021). By and large, the KB filter is a mean–variance optimal filter. However, when the observation noise is degenerate, the KB filter cannot be applied. The SA filter is in a recursive form with gain parameter matrix to be estimated. The main advantage of the SA filter is its recursive structure and its capability of treating systems with possible degenerate observation noise. The deep filter is a deep neural network based filter and is less structured and uses 'black box' type approach thanks to the hidden layers of the underlying neural network. In particular, we consider the filtering over the interval $[0, T]$ and set $T = 1$. We take discretization step size $h = 0.002$ which corresponds to the partition of $[0, T]$ into $N = 500$ subintervals and let $\delta_n = \delta = 0.5$ when computing $\widetilde{DJ}(R, \zeta)$. We also take the window size $n_0 = 50$ for deep filter and take the initial value R_0 as the matrix of ones with appropriate dimension. In the figures and tables, we use the SA, KB, and DF to denote SA filter, KB filter, and deep filter, respectively.

4.1. A linear model (1-D)

First we consider a one-dimensional linear model with

$$\begin{cases} dX_t = 0.5X_t dt + dW_t, \\ dY_t = X_t dt + \sigma_1 dV_t, \end{cases} \tag{4.9}$$

where W_t and V_t are independent standard Brownian motions. For training purposes, we generate $M = 1000$ Monte Carlo samples. For the SA filter, we take $\varepsilon_n = \varepsilon = 0.1$ and the number of maximum iteration $\widetilde{M} = 5000$ times. To compare with the deep filter, we consider a fully connected deep neural network with seven layers including input and output layers. The number of neurons for each layer are n_0, 64, 32, 16, 8, 8 and 1, respectively. In addition, we take the learning rate lr= 0.01, batch size 64, and epoch 50. The sigmoid activation function is used when performing optimization procedure with stochastic gradient descent (SGD). Finally, the performance of each filter is measured with a separate set of $M = 1000$ Monte Carlo samples. For any two processes ξ^1 and ξ^2 (e.g., $\xi^1 = \{X_{nh}\}$ and $\xi^2 = \{\widehat{X}_{nh}\}$), we measure their difference by

$$\|\xi^1 - \xi^2\| = \frac{\sum_{n=n_0}^{N} \sum_{m=1}^{M} |\xi_n^1(\omega_m) - \xi_n^2(\omega_m)|}{M(N - n_0 + 1)\Xi}, \quad \text{with}$$

$$\Xi = \frac{\sum_{n=n_0}^{N} \sum_{m=1}^{M} (|\xi_n^1(\omega_m)| + |\xi_n^2(\omega_m)|)}{M(N - n_0 + 1)}.$$

In Table 1, it shows the relative errors of the corresponding state and the SA filter, KB filter, and the deep filter, respectively. The optimal value of R^* for the SA filter is also provided. As can be seen from Table 1, as the observation noise $\sigma_1 \neq 0$, all filters perform well. The SA filter performs better than the deep filter, and the KB filter outperforms both the SA filter and deep filter. Nevertheless, when $\sigma_1 = 0$, the KB filter fails while both the SA filter and deep filter perform well. In addition, the optimal R^* appears to decrease w.r.t. σ_1. This suggests that larger filtering gain is needed when the observation noise gets smaller.

In Figs. 1, 2, and 3, a sample path of state and the corresponding SA, KB, and deep filters are plotted immediately below with $\sigma_1 = 0$, 0.1, 0.5, 1, 1.5, and 2.

Next, we examine the robustness of the SA filter. We consider two models: the nominal model (NM) used to train the system

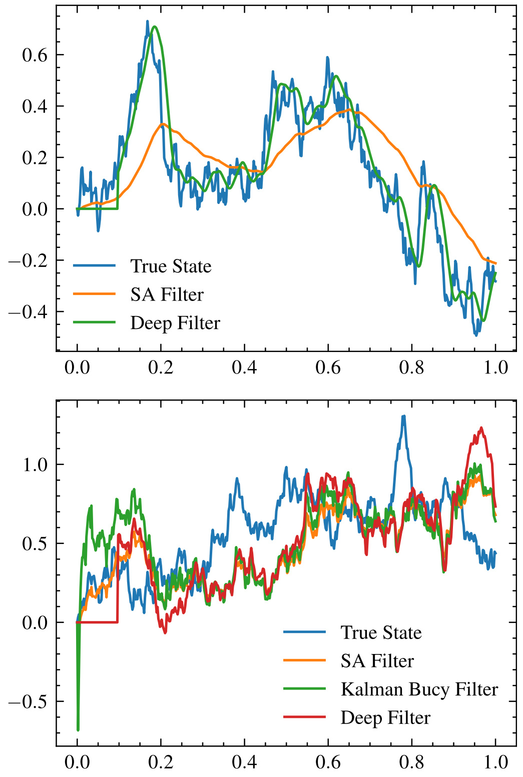

Fig. 1. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 0.0$ and $\sigma_1 = 0.1$, resp.

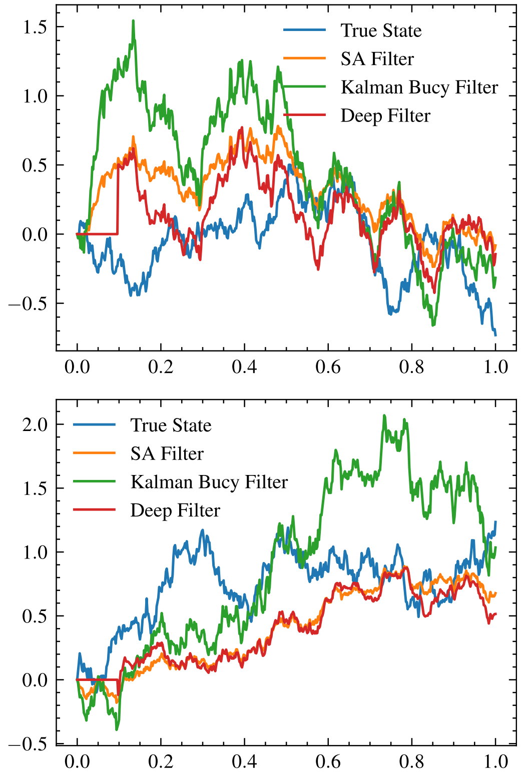

Fig. 2. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 0.5$ and $\sigma_1 = 1.0$, resp.

Table 1
Relative errors of the state and the SA, KB, and deep filters (1-D: linear model).

σ_1	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.1565	0.2112	0.4793	0.6658	0.7156	0.7233
KB		0.2014	0.4418	0.5614	0.6249	0.6272
DF	0.0678	0.2192	0.4954	0.6872	0.7622	0.8069
R^*	8.9688	7.6544	1.4832	0.4691	0.2831	0.1813

Table 2
Relative error dependence on σ_1^{AM}.

σ_1^{AM}	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.4653	0.4067	0.5041	0.5717	0.6178	0.6756
KB	0.3222	0.3248	0.4655	0.6042	0.6753	0.7387
DF	0.5287	0.5218	0.5309	0.5804	0.6120	0.6551

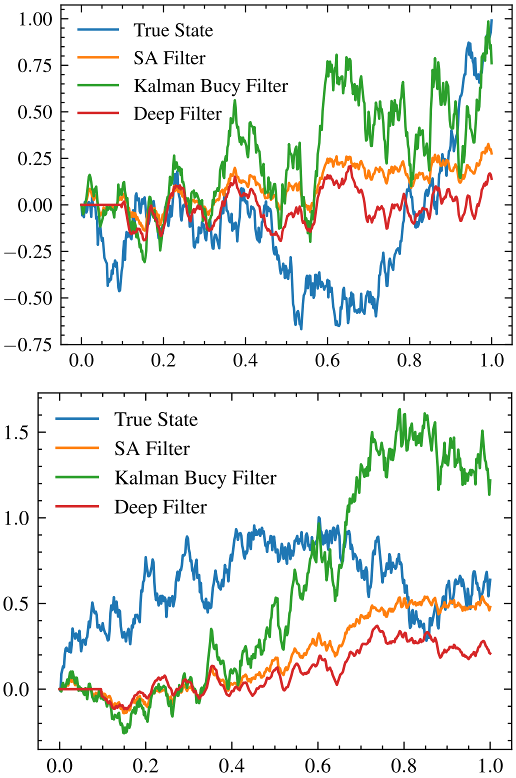

Fig. 3. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 1.5$ and $\sigma_1 = 2.0$, resp.

and the actual model (AM) used for testing. Here we consider the NM and AM with different observation noise:

$$(NM):\begin{cases} dX_t = 0.5X_t dt + dW_t, \\ dY_t = X_t dt + \sigma_1^{NM} dV_t, \end{cases}$$
$$(AM):\begin{cases} dX_t = 0.5X_t dt + dW_t, \\ dY_t = X_t dt + \sigma_1^{AM} dV_t. \end{cases} \tag{4.10}$$

We fix $\sigma_1^{NM} = 0.5$ and vary the value of σ_1^{AM}. When testing robustness, the coefficients of the NM in (4.10) are used to feed the KB filter, while the actual states and the corresponding observation driving the KB filter are generated by the AM. The same data usage applies to both the SA filter and deep filter. That is, the NM is used to train the neural network, while the AM yields the actual 'physical' process. The corresponding errors for the SA filter, the Kalman–Bucy filter, and the deep filter are given in Table 2. It is clear from Table 2 when $\sigma_1^{AM} \leq \sigma_1^{NM} = 0.5$, the KB filter performs better. However, when $\sigma_1^{AM} > \sigma_1^{NM} = 0.5$, both the SA filter and the deep filter provide better results. This suggests that both the SA filter and the deep filter are more robust when the actual observation noise is large.

Table 3
Relative errors of the state and the SA, KB, and deep filters (1-D: nonlinear).

σ_1	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.1659	0.2185	0.4879	0.6171	0.6105	0.6649
EKB		0.2080	0.5295	0.6078	0.5976	0.6248
DF	0.0752	0.2298	0.5485	0.7377	0.7706	0.8304
R^*	8.7331	7.4385	1.6909	0.4558	0.3134	0.2056

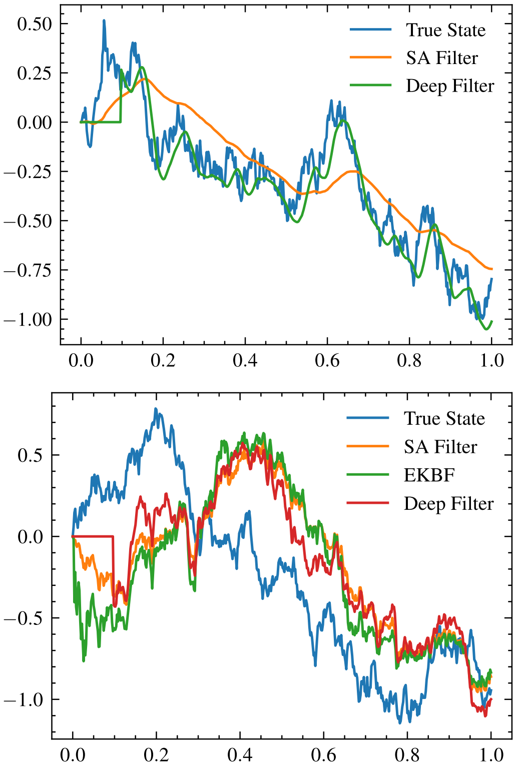

Fig. 4. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 0.0$ and $\sigma_1 = 0.1$, resp.

4.2. A nonlinear model (1-D)

Next, we consider the following one-dimensional non-linear model

$$\begin{cases} dX_t = \sin(5X_t)dt + dW_t, \\ dY_t = X_t dt + \sigma_1 dV_t, \end{cases} \tag{4.11}$$

where W_t and V_t are independent standard Brownian motions. For deep filter, we use the learning rate lr= 0.01, the batch size 64, and epoch 50. Furthermore, we take $\varepsilon_n = \varepsilon = 0.1$ in the SA filter. In this case, the KB filter needs to be replaced by extended KB filter. Table 3 displays the errors of different algorithms and the corresponding optimal R^* of the SA filter. A sample path of the state and that of the corresponding EKB filter, the SA filter and the deep filter are given in Figs. 4, 5, and 6, respectively, for different σ_1. As in the linear model, the EKB filter fails when $\sigma_1 = 0$. When $\sigma_1 \neq 0$, all three filters perform similarly. The deep filter falls behind when σ_1 gets larger.

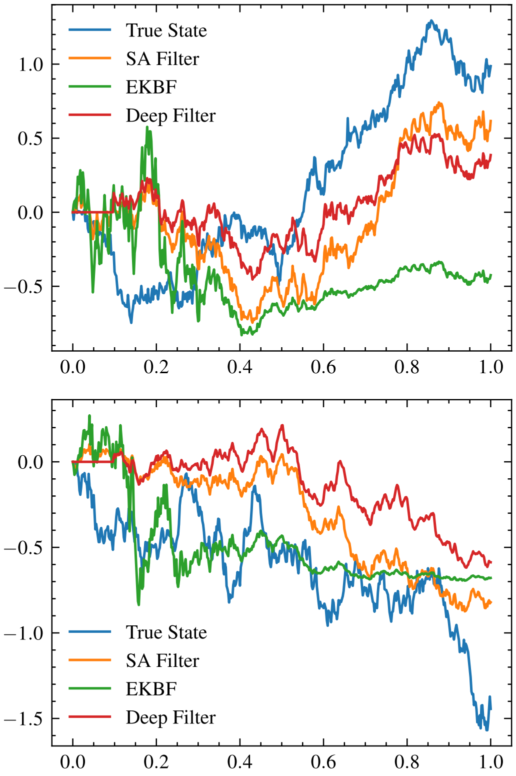

Fig. 5. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 0.5$ and $\sigma_1 = 1.0$, resp.

Table 4
Relative error dependence on σ_1^{AM}.

σ_1^{AM}	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.3556	0.3767	0.4615	0.5600	0.6561	0.6948
EKB	0.3153	0.3475	0.5107	0.5925	0.6905	0.7299
DF	0.5297	0.5384	0.5172	0.5556	0.6275	0.6383

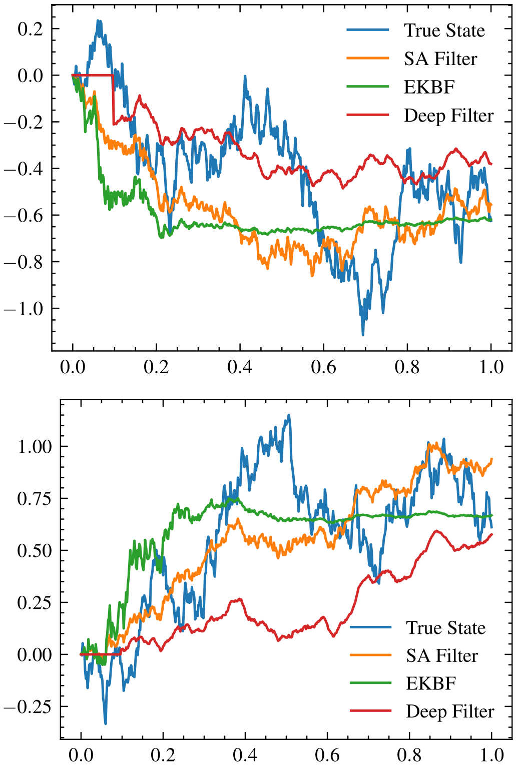

Fig. 6. A sample path of state and the corresponding SA, KB, and deep filters with $\sigma_1 = 1.5$ and $\sigma_1 = 2.0$, resp.

Table 5
Relative errors of the state and the corresponding filters (2-D: linear model).

s_1	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.3796	0.3843	0.5129	0.6228	0.6644	0.6758
KB		0.3371	0.4904	0.5480	0.5942	0.6247
DF	0.2980	0.3491	0.5444	0.6404	0.6861	0.7269

Next, as in the previous subsection, we examine the robustness of each filters. We consider the models:

$$\begin{aligned}
&(NM): \begin{cases} dX_t = \sin(5X_t)dt + dW_t, \\ dY_t = X_t dt + \sigma_1^{NM} dV_t, \end{cases} \\
&(AM): \begin{cases} dX_t = \sin(5X_t)dt + dW_t, \\ dY_t = X_t dt + \sigma_1^{AM} dV_t. \end{cases}
\end{aligned} \tag{4.12}$$

The relative errors for three algorithms are exhibited in Table 4. For smaller σ_1^{AM}, the EKB turns to outperform both the SA filter and deep filter. When σ_1^{AM} is large, both the SA filter and deep filter exhibit more robustness performance.

4.3. A linear model (2-D)

In this section, we consider a two-dimensional model given as follows:

$$\begin{cases} dX_t = AX_t dt + \sigma dW_t, \\ dY_t = HX_t dt + \sigma_1 dV_t, \end{cases} \tag{4.13}$$

where W_t and V_t are independent standard Brownian motions in $\mathbb{R}^2$. We take

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} s_1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Here s_1 is a scalar parameter. Clearly, when $s_1 = 0$, the observation noise is degenerate. In this example, we generate $M = 500$ Monte Carlo samples to train the deep filter and to search for the optimal parameter matrix R^* in the SA filter. For deep filter, we choose the forward neural network with layers $(2n_0, 128, 64, 32, 16, 8, 8, 2)$ and use the stochastic gradient descent algorithm with learning rate lr= 0.01. We also take the batch size to be 64 and epoch to be 200. For the SA filter, we iterate $\widetilde{M} = 500$ times. The relative errors for these filters are listed in Table 5.

Sample paths of the states and their corresponding SA, KB, and deep filters are given in Figs. 7–12. As shown in this table, the KB filter fails when $s_1 = 0$. However, when $s_1 \neq 0$, The KB filter outperforms SA filter and deep filter. Recall that for linear models with non-degenerate observation noise, the KB is optimal. In addition, the SA filter performs a little better than the deep filter in this example. Finally, we test the robustness of each filters by examining the error dependence on σ_1^{AM} with fixed $s_1^{NM} = 0.5$.

$$\begin{aligned}
&(NM): \begin{cases} dX_t = AX_t dt + dW_t, \\ dY_t = HX_t dt + \sigma_1^{NM} dV_t, \end{cases} \\
&(AM): \begin{cases} dX_t = AX_t dt + dW_t, \\ dY_t = HX_t dt + \sigma_1^{AM} dV_t, \end{cases}
\end{aligned} \tag{4.14}$$

where

$$\sigma_1^{NM} = \begin{bmatrix} s_1^{NM} & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_1^{AM} = \begin{bmatrix} s_1^{AM} & 0 \\ 0 & 1 \end{bmatrix}$$

As shown in Table 6, when $s_1^{AM} \leq s_1^{NM} = 0.5$, the KB filter performs better. On the other hand, when $s_1^{AM} > s_1^{NM}$, both the SA

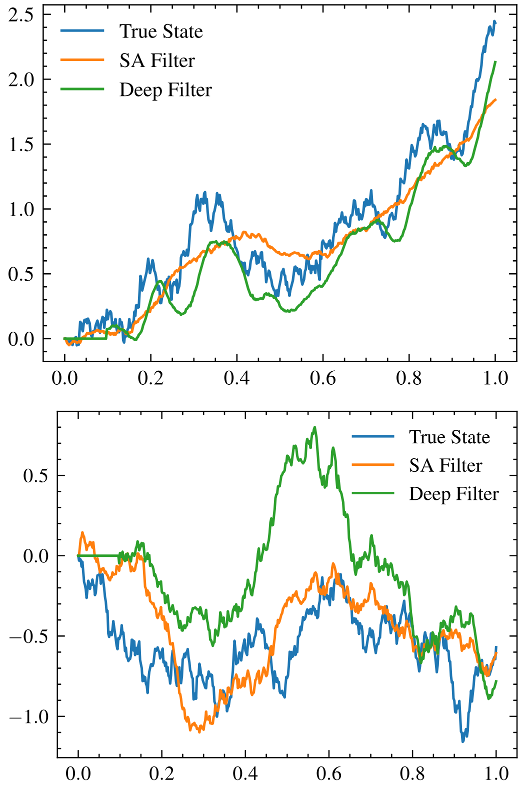

Fig. 7. A sample path of X_t^1 and the corresponding SA, KB, and deep filters (top), and that of X_t^2 and the corresponding SA, KB, and DF filters (bottom) with $s_1 = 0.0$.

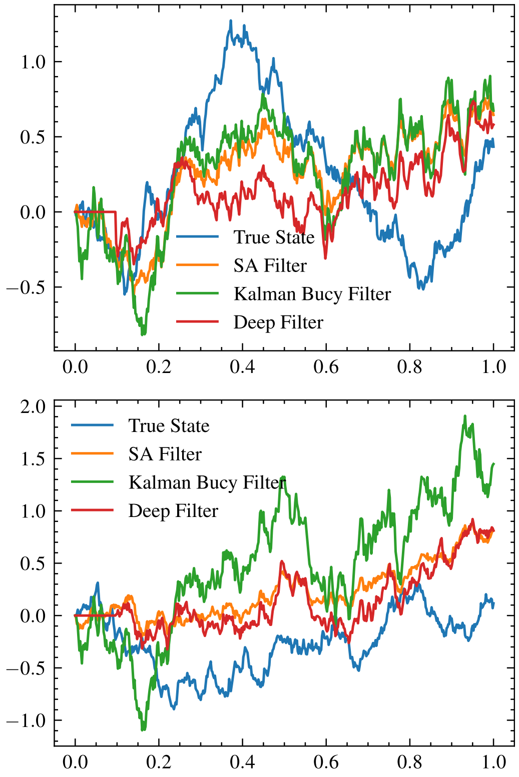

Fig. 9. A sample path of X_t^1 and the corresponding SA, KB, and DF filters (top) and that of X_t^2 and the corresponding SA, KB, and deep filters (bottom) with $s_1 = 0.5$.

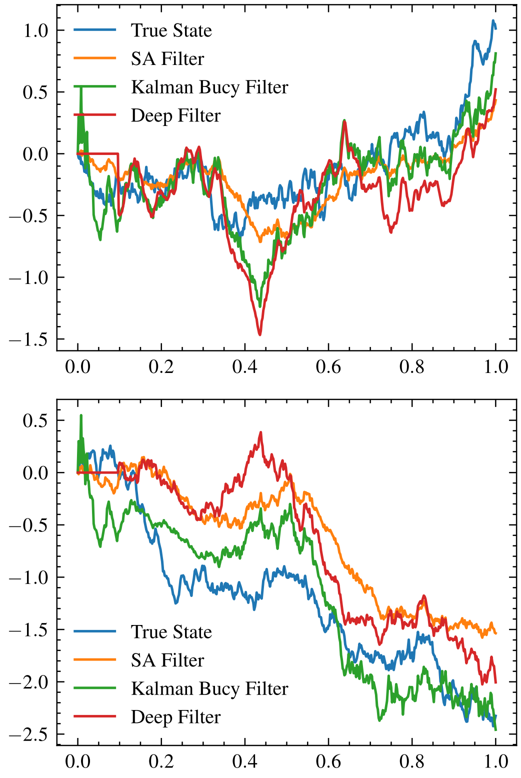

Fig. 8. A sample path of X_t^1 and the corresponding SA, KB, and DF filters (top) and that of X_t^2 and the corresponding SA, KB, and deep filters (bottom) with $s_1 = 0.1$.

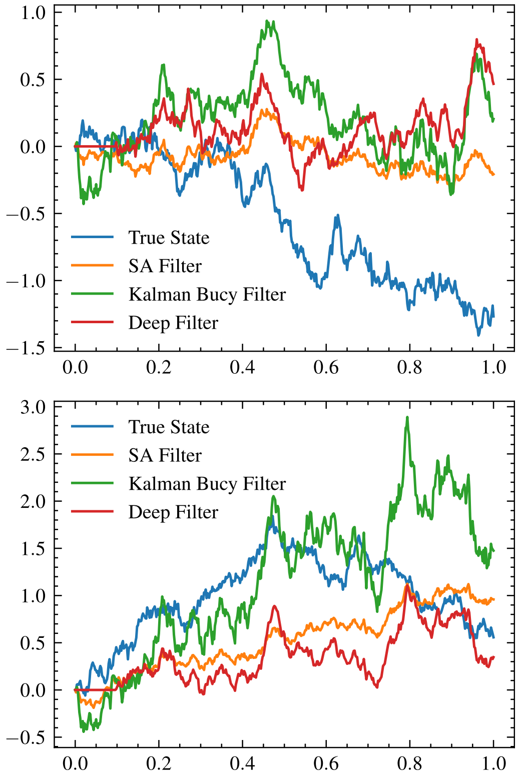

Fig. 10. A sample path of X_t^1 and the corresponding SA, KB, and deep filters (top) and that of X_t^2 and the corresponding SA, KB, and deep filters (bottom) with $s_1 = 1.0$.

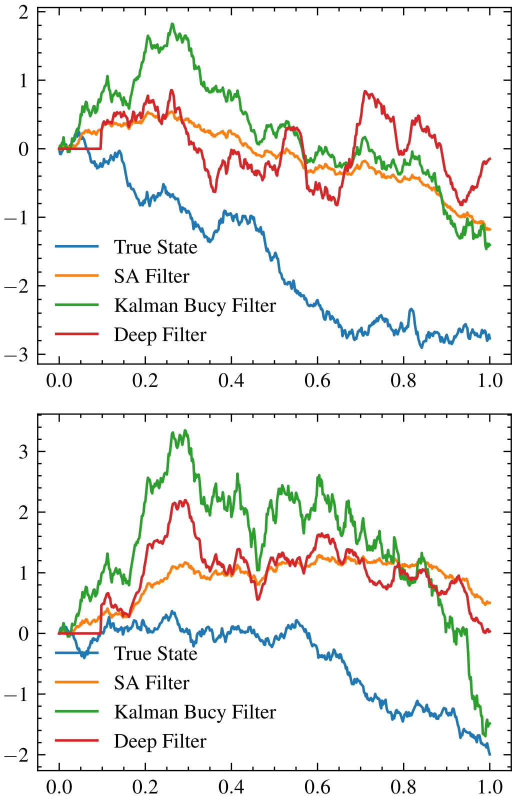

Fig. 11. A sample path of X_t^1 and the corresponding SA, KB, and DF filters (top) and that of X_t^2 and the corresponding SA, KB, and deep filters (bottom) with $s_1 = 1.5$.

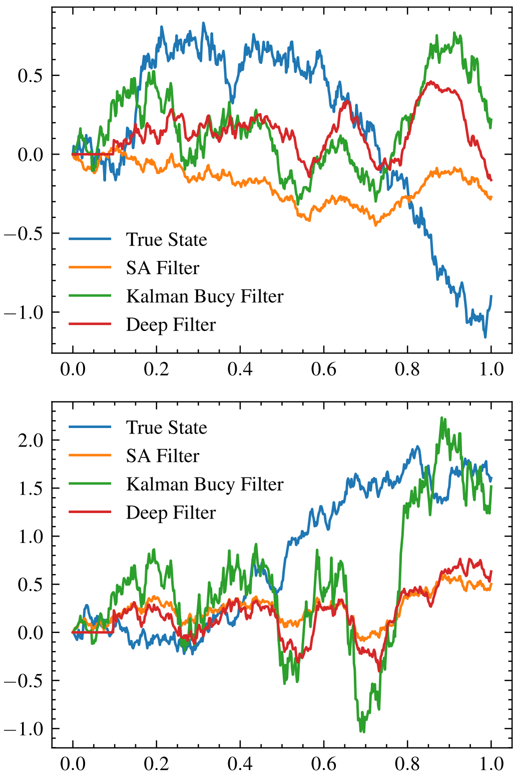

Fig. 12. A sample path of X_t^1 and the corresponding SA, KB, and DF filters (top) and that of X_t^2 and the corresponding SA, KB, and deep filters (bottom) with $s_1 = 2.0$.

Table 6
Relative error dependence of s_1^{AM}.

s_1^{AM}	0.0	0.1	0.5	1.0	1.5	2.0
SA	0.4877	0.4858	0.5223	0.5651	0.6337	0.6669
KB	0.4348	0.4338	0.4861	0.5695	0.6559	0.6968
DF	0.5290	0.5326	0.5431	0.5772	0.6246	0.6508

filter and the deep filter outperforms that KB. This demonstrates the robustness of both the SA filter and deep filter when the actual noise is larger than the nominal noise.

5. Concluding remarks

This work has been devoted to studying a novel filtering method for systems whose observations are degenerate. For such systems, the standard techniques in the literature do not work due to the degeneracy. The proposed methods are based on stochastic approximation methods. A number of examples, including linear and nonlinear systems are considered. The computational results are promising. The suggested methods open up new windows for further investigation. In particular, it would be interesting to consider the SA filtering under a discrete-time system of the form $\widehat{X}_{n+1} = F(\widehat{X}_n) + R(Y_n - H(\widehat{X}_n))$, for suitable functions F and H, to develop the corresponding SA algorithms, and to study related convergence properties and numerical performance.

Acknowledgment

This research was supported in part by the National Science Foundation under grant DMS-2204240.

References

Dieu, N. T., Du, N. H., Nguyen, D. H., & Yin, G. (2016). Classification of asymptotic behavior in a stochastic SIR model. *SIAM Journal on Applied Dynamical Systems*, *15*, 1062–1084.

Duncan, T. E. (1967). *Probability Densities for Diffusion Processes with Applications To Nonlinear Filtering Theory and Detection Theory* (Ph.D. Diss.), Stanford Univ..

Fleming, W. H., & Rishel, R. W. (1975). *Deterministic and Stochastic Optimal Control*. New York: Springer-Verlag.

Frey, R., Schmidt, T., & Xu, L. (2018). On Galerkin approximations for the Zakai equation with diffusive and point process observations. https://arxiv.org/pdf/1303.0975.pdf.

Gao, J., & Tembine, H. (2016). Distributed mean-field type filters for big data assimilation. In *Proceedings of the 18th IEEE International Conference on High Performance Computing and Communications*, (pp. 1446–1453).

Kermack, W. O., & McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics–I, II, III. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *115*, 700–721, 138 (1932) 55–83, 141 (1933) 94–122.

Kesten, H., & Papanicolaou, G. C. (1980/81). A limit theorem for stochastic acceleration. *Communications in Mathematical Physics*, *78*, 19–63.

Kushner, H. J. (1964). On the differential equations satisfied by conditional probability densities of Markov processes, with applications. *SIAM Journal on Control and Optimization Series A*, *2*, 106–119.

Kushner, H. J., & Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications* (2nd Ed.). New York: Springer-Verlag.

Mortensen, R. E. (1968). Maximum-likelihood recursive nonlinear filtering. *Journal of Optimization Theory and Applications*, *2*, 386–394.

Nguyen, D., Nguyen, N., & Yin, G. (2021). Stochastic functional Kolmogorov equations I: Persistence. *Stochastic Processes and their Applications*, *142*, 319–364.

Nguyen, N., & Yin, G. (2021). Large deviations principles for langevin equations in random environment and applications. *Journal of Mathematical Physics*, *62*, Article 083301.

Wang, L. Y., Yin, G., & Zhang, Q. (2021). Deep filtering. *Communications in Information and Systems*, *21*, 651–667.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *11*, 230–243.

Hongjiang Qian received the B.S degree in mathematics from Huazhong University of Science and Technology, Wuhan, P.R. China, in 2018. He is currently pursuing his Ph.D. degree at the University of Connecticut, Storrs, CT. He spent two years studying in the department of Mathematics at Wayne State University. His current research interests include stochastic approximation and stochastic systems theory and applications.

Qing Zhang is Professor of Mathematics at the University of Georgia. He received his Ph.D. in Applied Mathematics from Brown University. He specializes in stochastic systems and control, filtering, and applications in finance. He has published five monographs on production planning and two-time scale Markovian systems and applications and over 200 research papers. He co-edited six books and was Associate Editor of Automatica, IEEE Transactions on Automatic control, and SIAM Journal on Control and Optimization. He is currently Corresponding Editor of SIAM Journal on Control and Optimization. He also served on a number of international conference organizing committees including Co-Chair of the organizing committee for the SIAM Conference on Control and Applications in 2017.

George Yin received the B.S. degree in mathematics from the University of Delaware in 1983, and the M.S. degree in electrical engineering and the Ph.D. degree in applied mathematics from Brown University in 1987. He joined the Department of Mathematics, Wayne State University in 1987, and became Professor in 1996 and University Distinguished Professor in 2017. He moved to the University of Connecticut in 2020. His research interests include stochastic processes, stochastic systems theory and applications. Dr. Yin was the Chair of the SIAM Activity Group on Control and Systems Theory, and served on the Board of Directors of the American Automatic Control Council. He is the Editor-in-Chief of *SIAM Journal on Control and Optimization*, was a Senior Editor of *IEEE Control Systems Letters*, and is an Associate Editor of *ESAIM: Control, Optimization and Calculus of Variations*, *Applied Mathematics and Optimization* and many other journals. He was an Associate Editor of *Automatica* 2005–2011 and *IEEE Transactions on Automatic Control* 1994–1998. He is a Fellow of IFAC, a Fellow of IEEE, and a Fellow of SIAM.