Online Maximum-Likelihood Estimation of the Parameters of Partially Observed Diffusion Processes

Simone Carlo Surace and Jean-Pascal Pfister

***Abstract*—We revisit the problem of estimating the parameters of a partially observed diffusion process, consisting of a hidden state process and an observed process, with a continuous time parameter. The estimation is to be done online, i.e., the parameter estimate should be updated recursively based on the observation filtration. We provide a theoretical analysis of the stochastic gradient ascent algorithm on the incomplete-data log-likelihood. The convergence of the algorithm is proved under suitable conditions regarding the ergodicity of the process consisting of state, filter, and tangent filter. Additionally, our parameter estimation is shown numerically to have the potential of improving suboptimal filters, and can be applied even when the system is not identifiable due to parameter redundancies. Online parameter estimation is a challenging problem that is ubiquitous in fields such as robotics, neuroscience, or finance in order to design adaptive filters and optimal controllers for unknown or changing systems. Despite this, theoretical analysis of convergence is currently lacking for most of these algorithms. This paper sheds new light on the theory of convergence in continuous time.**

***Index Terms*—Maximum likelihood estimation, parameter estimation, filtering theory, gradient methods, stochastic processes.**

I. Introduction

We CONSIDER the following family of partially observed dimensional diffusion process under the probability measure P_θ:

$$dX_t = f(X_t, \theta)dt + g(X_t, \theta)dW_t \tag{1}$$

$$dY_t = h(X_t, \theta)dt + dV_t \tag{2}$$

parameterized by $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is an open subset. The process X_t is called the hidden state or signal process with values in $\mathbb{R}^n$, and Y_t is called the observation process with values in $\mathbb{R}^{n_y}$. In addition, W_t and V_t are independent $\mathbb{R}^{n'}$- and $\mathbb{R}^{n_y}$-valued standard Wiener processes (signal and observation noise). For all $\theta \in \Theta$, we assume the initial conditions $X_0 \sim p_0(\theta)$ to be independent of W_t and V_t, we set $Y_0 = 0$, and we assume that $f(\cdot, \theta), g(\cdot, \theta), h(\cdot, \theta)$ are functions from $\mathbb{R}^n \times \mathbb{R}^p$ to $\mathbb{R}^n$, $\mathbb{R}^{n \times n'}$, and $\mathbb{R}^{n_y}$, respectively, that ensure the existence and uniqueness in probability of strong solutions to (1) and (2) for all $t \geq 0$. Additional regularity conditions for f, g, h in both arguments will be required for the convergence proof.

This setting is familiar in classical filtering theory, where the problem is to find (assuming the knowledge of θ) the conditional distribution of X_t conditioned on the history of observations $\mathcal{F}_t^Y = \sigma\{Y_s, 0 \leq s \leq t\}$. In this paper, we focus on the following parameter estimation problem: assuming that a system with parameter θ_0 generates observations Y_t, we want to estimate θ_0 from $\mathcal{F}_t^Y$ recursively.

We will consider a well-known algorithm for parameter estimation, the so-called stochastic gradient ascent (SGA) on the incomplete-data log-likelihood function. The stochasticity comes from the online estimation from the stream of observations, which provides a noisy estimate of the gradient of the asymptotic log-likelihood. The main open issue we are addressing in this paper is the analysis of the convergence of the parameter estimate.

This paper is structured as follows. In Section II, we describe the method of obtaining recursive parameter estimates. In Section III, we prove the almost sure convergence of the recursive parameter estimates to stationary points of the asymptotic likelihood. In Section IV, we provide a few numerical examples, including cases where the model is not identifiable and the filter is suboptimal. Finally, in Section V, we discuss the theoretical similarities and differences to related methods of recursive parameter estimation.

II. Methods

In this paper, we consider the problem of finding an estimator $\tilde{\theta}_t$ that is $\mathcal{F}_t^Y$-measurable and recursively computable, such as to estimate θ_0 online from the continuous stream of observations. For this task, we propose an approach based on a modification of offline maximum-likelihood estimation, and therefore need to compute the likelihood of the observations (also called incomplete-data likelihood) as a function of the model parameters.

Manuscript received September 19, 2017; accepted September 21, 2018. Date of publication November 9, 2018; date of current version June 26, 2019. This work was supported by the Swiss National Science Foundation under Grant PP00P3_179060 and Grant PP00P3_150637. This paper was presented in part at the 2016 Time Series Workshop at Neural Information Processing Systems Conference, Barcelona, Spain, December 2016. Recommended by Associate Editor G. Pillonetto.
(Corresponding author: Simone Carlo Surace.)

The authors are with the Institute of Neuroinformatics and Neuroscience Center Zürich, University of Zürich and ETH Zurich, Zürich 8057, Switzerland, and Department of Physiology, University of Bern, Bern 3012, Switzerland (e-mail: surace@ini.uzh.ch; jpfister@ini.uzh.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2018.2880404

It is a fundamental theorem of filtering theory[1] that the *innovation process* I_t, defined by

$$I_t = Y_t - \int_0^t \hat{h}_s(\theta) ds, \quad \hat{h}_s(\theta) = \mathbb{E}_\theta \left[h(X_s, \theta) \middle| \mathcal{F}_s^Y \right] \tag{3}$$

is a $(P_\theta, \mathcal{F}_t^Y)$-Brownian motion. By applying Girsanov's theorem, we can change to a measure $\tilde{P}$ under which Y_t is a $(\tilde{P}, \mathcal{F}_t^Y)$-Brownian motion and thus (statistically) independent of both the hidden state X_t and the parameter θ. The change of measure has a Radon–Nikodym derivative

$$\frac{dP_\theta}{d\tilde{P}}\Big|_{\mathcal{F}_t^Y} = \exp \left[\int_0^t \hat{h}_s(\theta) \cdot dY_s - \frac{1}{2} \int_0^t \|\hat{h}_s(\theta)\|^2 ds \right] \tag{4}$$

where $\cdot$ denotes the Euclidean scalar product.

Since the reference measure $\tilde{P}$, restricted on $\mathcal{F}_t^Y$, does not depend on θ, we can express the incomplete-data log-likelihood function in terms of the optimal filter as

$$\mathcal{L}_t(\theta) = \log \frac{dP_\theta}{d\tilde{P}}\Big|_{\mathcal{F}_t^Y} = \int_0^t \hat{h}_s(\theta) \cdot dY_s - \frac{1}{2} \int_0^t \|\hat{h}_s(\theta)\|^2 ds. \tag{5}$$

A. Offline Algorithm

We start by describing an offline method for parameter estimation using the log-likelihood function in (5), which serves as a basis for the online method.

If we were interested in offline learning, our goal would be to maximize the value of $\mathcal{L}_t(\theta)$ for fixed t. There is a number of methods to solve this optimization problem. Among these, a simple iterative method is the gradient ascent, where an estimate $\tilde{\theta}_k$ at iteration k is updated according to

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + \gamma_k \partial_\theta^\top \mathcal{L}_t(\theta)\Big|_{\theta = \tilde{\theta}_k} \tag{6}$$

where $\gamma_k > 0$ is called the *learning rate*, and $\partial_\theta^\top$ denotes the (Euclidean) gradient operator with respect to the parameter θ. At each iteration, the derivative of the likelihood function has to be recomputed. From (5), we obtain

$$\partial_\theta \mathcal{L}_t(\theta) = \int_0^t \left(dY_s - \hat{h}_s(\theta) ds \right)^\top \hat{h}_s^\theta(\theta) \tag{7}$$

where $\cdot^\top$ denotes the matrix transpose and the last factor of the integrand, denoted by

$$\hat{h}_s^\theta(\theta) \doteq \partial_\theta \hat{h}_s(\theta) \tag{8}$$

takes values in the matrices of size $n_y \times p$ and is called the *filter derivative* of h with respect to θ.[2]

[1]For a detailed exposition of the mathematical background [such as Girsanov's theorem, changes of measure, or the filtering equation (9)], we suggest a look at the standard literature on filtering theory, e.g., [1].

[2]Here and in the sequel, we use the convention that the gradient operator adds a covariant dimension to the tensor field it acts on. For example, $\partial_\theta \mathcal{L}_t(\theta)$ takes values that are covectors (row vectors), and the gradient of $\hat{h}_t(\theta)$, which has values in $\mathbb{R}^{n_y}$, w.r.t. θ, is a $(n_y \times p)$-matrix ($\mathbb{R}^{n_y} \otimes \mathbb{R}^{p*}$-tensor, where * denotes a dual space)-valued process, which we denote by $\hat{h}_t^\theta(\theta)$.

In principle, computing the quantities $\hat{h}_t(\theta)$ requires the solution of the Kushner–Stratonovich filtering equation

$$d\hat{\varphi}_t = \widehat{(\mathcal{A}\varphi)}_t dt + \left(\widehat{(h\varphi)}_t - \hat{h}_t \hat{\varphi}_t \right) \cdot \left(dY_t - \hat{h}_t dt \right) \tag{9}$$

for arbitrary integrable $\varphi : \mathbb{R}^n \to \mathbb{R}$, where $\mathcal{A}$ is the generator of the process X_t. However, exact solutions are rarely available. In the following, we assume that (9) admits a finite-dimensional recursive solution or a finite-dimensional recursive approximation. This means that there is an $\mathcal{F}_t^Y$-adapted process $M_t(\theta)$ with values in $\mathbb{R}^m$ and a mapping $\psi_h : \mathbb{R}^m \times \Theta \to \mathbb{R}^{n_y}$ such that either $\hat{h}_t(\theta) = \psi_h(M_t(\theta), \theta)$ (in the case of an exact solution), or such that the equation holds approximately, i.e., with some bounds (preferably uniform in time) on

$$\mathrm{Var}\|\hat{h}_t(\theta) - \psi_h(M_t(\theta), \theta)\|.$$

For example, in the linear-Gaussian case and if X_0 has a Gaussian distribution, the optimal filter can be represented in terms of a Gaussian distribution with mean μ_t and variance P_t, i.e., $m = 2$, $M_t = (\mu_t, P_t)$, and for $h_\theta(x) = \theta x$, we have $\psi_h(M_t(\theta), \theta) = \theta \mu_t$. Apart from the linear-Gaussian case [2] just mentioned, finite-dimensional (exact) recursive solutions only exist for a small class of systems, namely the Beneš class and its extensions [3]–[8]. Meanwhile, finite-dimensional recursive *approximations* are available for a large class of systems, but the appropriate choice of approximation is a complex topic in its own right and will not be explored here. We merely mention a few standard approximation schemes: extended and unscented Kalman filters [9], [10], projection or assumed-density filters (ADFs) [11], [12], particle filters [13], and particle filters without weights [14]–[16].

Given a finite-dimensional representation of the filter, a corresponding representation of the filter derivative may be formally defined by differentiation with respect to θ

$$\hat{h}_t^\theta(\theta) \simeq \partial_\theta \psi_h(M_t(\theta), \theta) + \partial_M \psi_h(M_t(\theta), \theta) M_t^\theta(\theta) \tag{10}$$

where ∂_M denotes the gradient w.r.t. the first argument of ψ_h and $M_t^\theta(\theta)$ denotes the $(m \times p)$-matrix-valued derivative of the process $M_t(\theta)$. For the system in (1) and (2) and for a large class of exact and approximate filters, $M_t(\theta)$ solves a stochastic differential equation (SDE) of the form

$$\begin{aligned} dM_t(\theta) = \mathcal{R}(\theta, M_t(\theta))dt + \mathcal{S}(\theta, M_t(\theta))dY_t \\ + \mathcal{T}(\theta, M_t(\theta))dB_t \end{aligned} \tag{11}$$

where $\mathcal{R}, \mathcal{S}$, and $\mathcal{T}$ go to $\mathbb{R}^m$, $\mathbb{R}^{m \times n_y}$, and $\mathbb{R}^{m \times m'}$, respectively, and B_t is an m'-dimensional Brownian motion that is independent of $\mathcal{F}_t^{X,Y}$ (e.g., independent noise in particle filters). By differentiating w.r.t. θ, we find the corresponding SDE for $M_t^\theta(\theta)$

$$\begin{aligned} dM_t^\theta(\theta) = \mathcal{R}'(M_t(\theta), M_t^\theta(\theta), \theta)dt \\ + \mathcal{S}'(M_t(\theta), M_t^\theta(\theta), \theta)dY_t \\ + \mathcal{T}'(M_t(\theta), M_t^\theta(\theta), \theta)dB_t \end{aligned} \tag{12}$$

where the tensor fields $\mathcal{R}', \mathcal{S}', \mathcal{T}'$ are given by

$$\mathcal{R}'(M_t(\theta), M_t^\theta(\theta), \theta) = \partial_\theta \mathcal{R}(M_t(\theta), \theta) + \partial_M \mathcal{R}(M_t(\theta), \theta) M_t^\theta(\theta) \quad (13)$$

and analogously for $\mathcal{S}$ and $\mathcal{T}$. In Section IV, we will present examples of both exact and approximate filters for which these calculations will be made explicit.

These equations can be conveniently summarized in a single SDE

$$d\mathcal{X}_t(\theta) = \Phi(\mathcal{X}_t(\theta), \theta)dt + \Sigma(\mathcal{X}_t(\theta), \theta)d\mathcal{B}_t. \quad (14)$$

Here, $\mathcal{X}_t(\theta)$ is a D-dimensional process defined by concatenating the state X_t, the filter representation, and all the filter derivatives as follows:

$$\begin{aligned}\mathcal{X}_t(\theta) &= \mathcal{C}(X_t, M_t(\theta), M_t^\theta(\theta)) \\ &\doteq \Big(X_{t,1}, \ldots, X_{t,n}, M_{t,1}(\theta), \ldots, M_{t,m}(\theta), \\ &\quad M_{t,1}^{\theta_1}(\theta), \ldots, M_{t,m}^{\theta_1}(\theta), \ldots, M_{t,1}^{\theta_p}(\theta), \ldots, M_{t,m}^{\theta_p}(\theta)\Big)^\top\end{aligned} \quad (15)$$

where $D = n + m + mp, \mathcal{C} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times p} \to \mathbb{R}^D$ is the concatenation, and $\mathcal{B}_t$ is the Wiener process defined by $\mathcal{B}_t = (W_t, V_t, B_t)$.

B. Online Algorithm

Instead of integrating the gradient of the log-likelihood function up to time t, an SGA uses the integrand of the gradient of the log-likelihood (evaluated with the current parameter estimate) to update the parameter estimate online as new data are reaching the observer. The SDE for the SGA takes the form

$$d\tilde{\theta}_t = \begin{cases}\gamma_t F(\tilde{\mathcal{X}}_t, \tilde{\theta}_t)dt + \gamma_t H(\tilde{\mathcal{X}}_t, \tilde{\theta}_t)^\top dV_t, & \tilde{\theta}_t \in \Theta \\ 0, & \tilde{\theta}_t \notin \Theta\end{cases} \quad (16)$$

where $\tilde{\mathcal{X}}_t \doteq \mathcal{C}(X_t, \tilde{M}_t, \tilde{M}_t^\theta)$ is a diffusion process with SDE

$$d\tilde{\mathcal{X}}_t = \Phi(\tilde{\mathcal{X}}_t, \tilde{\theta}_t)dt + \Sigma(\tilde{\mathcal{X}}_t, \tilde{\theta}_t)d\mathcal{B}_t \quad (17)$$

consisting of the state as well as the filter and filter derivatives integrated with the online parameter estimate. The functions l, F, and H, which go from $\mathbb{R}^D \times \Theta$ to $\mathbb{R}$, $\mathbb{R}^p$, and $\mathbb{R}^{n_y \times p}$, respectively, are defined as

$$l(x, \theta) = \psi_h(M, \theta) \cdot \left[h(X, \theta_0) - \tfrac{1}{2}\psi_h(M, \theta)\right] \quad (18)$$

$$F(x, \theta) \doteq H(x, \theta)^\top \left[h(X, \theta_0) - \psi_h(M, \theta)\right] \quad (19)$$

$$H(x, \theta) \doteq \partial_\theta \psi_h(M, \theta) + \partial_M \psi_h(M, \theta) M' \quad (20)$$

where $(X, M, M') = \mathcal{C}^{-1}(x)$ are the components of x. The function l will be used later on (Eq. (26) and following).

III. Convergence Analysis

As in any stochastic gradient method, convergence relies on being able to control the errors of estimating the gradient. This is usually done by assuming ergodicity of the system and applying regularity results on a Poisson equation, as in the treatments of related problems presented in [17] (discrete time), as well as [18] (continuous time but fully observed). In our case, the ergodic system consists of the hidden state, the filter, and the filter derivative. We therefore need to find the assumptions that guarantee that this system is ergodic, with appropriate regularity results. We attack this problem in Section III-A by giving conditions directly in terms of the finite-dimensional approximation. However, this means that these conditions have to be checked on a case-by-case basis in order to obtain convergence results. Once the question of ergodicity is settled, the remainder of the proof is very similar to the one in [18].

Besides this direct verification approach, the only hope of otherwise obtaining ergodicity seems to be via the *optimal* filter. This is due to the fact that the finite-dimensional system is usually highly degenerate, such that the standard theory which was used, e.g., in [18], does not apply.[3] Ergodicity of the *optimal* filter for a stochastic dynamical system of the form of (1) and (2) follows from the ergodicity of the hidden state process and the nondegeneracy of observations[4] (see [25]–[28]). The problem is then to extend these results to the derivative of the optimal filter with respect to the parameters, and to transfer them to *approximate* finite-dimensional representations of the filter, given some bounds on the accuracy of the approximation. The question of transferring ergodicity of the *exact* filter and filter derivative to the approximate ones, as well as the ergodicity of the filter derivative, remains open.

A. Direct Conditions for the Ergodicity of the Approximate Filter

Here, we give sufficient conditions directly in terms of the approximate filtering equation. Before stating the conditions, we introduce the following notation: We say that a function $G : \mathbb{R}^d \times \Theta \to \mathbb{R}$ has the polynomial growth property (PGP) if there are $q, K > 0$ such that for all $\theta \in \Theta$

$$|G(x, \theta)| \le K(1 + ||x||^q). \quad (21)$$

Let $\mathbb{G}^d$ be the function space defined by all functions $G : \mathbb{R}^d \times \Theta \to \mathbb{R}$ such that the following conditions hold:

1) $G(\cdot, \theta) \in C(\mathbb{R}^d)$;
2) $G(x, \cdot) \in C^2(\Theta)$; and
3) $\partial_\theta G(x, \cdot)$ and $\partial_\theta^2 G(x, \cdot)$ are Hölder continuous with exponent $\alpha > 0$.

Let $\mathbb{G}_c^d$ be the subset consisting of all $G \in \mathbb{G}^d$ that are centered, i.e., $\int_{\mathbb{R}^d} G(x, \theta)\mu_\theta(dx) = 0$. Let $\bar{\mathbb{G}}^d$ be the subset consisting of all $G \in \mathbb{G}^d$ such that G and all its first and second derivatives w.r.t. θ satisfy the PGP.

Now, we may state the conditions on the following processes.

Condition 1:

i) The process $\mathcal{X}_t(\theta)$ is ergodic under P_{θ_0}, with a unique invariant probability measure μ_θ on $(\mathbb{R}^D, \mathbb{B}_D)$, where $\mathbb{B}_D$ is the Borel σ-algebra on $\mathbb{R}^D$.

[3] The theory for elliptic diffusions [19]–[21] is clearly not applicable, and it is not clear how to apply hypoellipticity or Hörmander's condition [22]–[24] in general. For example, in the linear-Gaussian case, the process $\mathcal{X}_t$ does not satisfy the parabolic Hörmander condition.

[4] Note that the work in [25] contained a gap that has been fully closed by the work presented in [26].

ii) For any $q > 0$ and $\theta \in \Theta$, there is a constant $K_q > 0$ such that

$$\int_{\mathbb{R}^D} (1 + ||x||^q)\, \mu_\theta(dx) \le K_q. \tag{22}$$

iii) Define the finite signed measures $\nu_{\theta,i} = \partial_{\theta_i} \mu_\theta,\ i = 1, \ldots, p$, and let $|\nu_{\theta,i}(dx)|$ be their total variation. For any $q > 0$ and $\theta \in \Theta$, there is a constant $K'_q > 0$ such that

$$\int_{\mathbb{R}^D} (1 + ||x||^q)\, |\nu_{\theta,i}(dx)| \le K'_q. \tag{23}$$

iv) Let $\mathcal{A}_\mathcal{X}$ be the infinitesimal generator of $\mathcal{X}_t(\theta)$ and let $G \in \mathbb{G}_c^D$. Then, the Poisson equation $\mathcal{A}_\mathcal{X} v(x,\theta) = G(x,\theta)$ has a unique solution $v(x,\theta)$ that lies in $\mathbb{G}^D$, with $v(\cdot,\theta) \in C^2(\mathbb{R}^D)$. Moreover, if $G \in \bar{\mathbb{G}}^D$, then $v \in \bar{\mathbb{G}}^D$ and also $\partial_x \partial_\theta v$ has the PGP.

v) For all $q > 0$, $\mathbb{E}[||\tilde{\mathcal{X}}_t||^q] < \infty$ and there is a $K > 0$ such that for t large enough

$$\forall \theta \in \Theta \quad \mathbb{E}_{\theta_0}\left[\sup_{s \le t} ||\mathcal{X}_s(\theta)||^q\right] \le K\sqrt{t} \tag{24}$$

$$\mathbb{E}_{\theta_0}\left[\sup_{s \le t} ||\tilde{\mathcal{X}}_s||^q\right] \le K\sqrt{t}. \tag{25}$$

■

Condition 2: The function F is in $\bar{\mathbb{G}}^d$ (componentwise). The function ψ_h is in $\mathbb{G}^m$ and has the PGP (componentwise). In addition, $l(x,\theta)$, $H(x,\theta)$, and Σ have the PGP (componentwise). ■

Finally, the following condition on the learning rate is imposed.

Condition 3: $\int_0^\infty \gamma_t dt = \infty$, $\int_0^\infty \gamma_t^2 dt = 0$, and there is an $r > 0$ such that $\lim_{t\to\infty} \gamma_t^2 t^{1/2+2r} = 0$. ■

B. Results

Let the approximate (in the sense of using the approximate filter representation) incomplete-data log-likelihood be given by

$$\mathcal{L}_t(\theta) = \int_0^t l(\mathcal{X}_s(\theta), \theta) ds + \int_0^t \psi_h(M_s(\theta), \theta) \cdot dV_s. \tag{26}$$

Under the above-mentioned Conditions 1–3, we have the following.

Proposition 1 (Regularity of the asymptotic likelihood):

i) The process $\frac{1}{t}\mathcal{L}_t(\theta)$ converges almost surely (a.s.) to $\tilde{\mathcal{L}}(\theta)$, which is given by

$$\tilde{\mathcal{L}}(\theta) = \int_{\mathbb{R}^D} l(x,\theta)\mu_\theta(dx). \tag{27}$$

ii) The asymptotic likelihood function $\tilde{\mathcal{L}}(\theta)$ is in $C^2(\Theta)$, and the gradient g and Hessian $\mathcal{H}$ of the asymptotic likelihood are given in terms of the invariant measure μ_θ and its derivative ν_θ as

$$g(\theta) \doteq \partial_\theta \tilde{\mathcal{L}}(\theta) = \int_{\mathbb{R}^D} F(x,\theta)^\top \mu_\theta(dx) \tag{28}$$

$$\mathcal{H}(\theta) \doteq \partial_\theta^\top \partial_\theta \tilde{\mathcal{L}}(\theta) = \int_{\mathbb{R}^D} \partial_\theta F(x,\theta)\mu_\theta(dx) + \int_{\mathbb{R}^D} F(x,\theta)\nu_\theta(dx). \tag{29}$$

iii) The function $G(x,\theta) \doteq F(x,\theta) - g(\theta)$ is in $\bar{\mathbb{G}}^D \cap \mathbb{G}_c^D$.

iv) There is a constant $C > 0$ such that

$$\tilde{\mathcal{L}}(\theta) + \|g(\theta)\| + \|\mathcal{H}(\theta)\| \le C. \tag{30}$$

Proof: See Appendix A. ■

Now we can formulate our main result. Its proof relies on several lemmas that are given in Appendix B.

Theorem 1 (main theorem): Assume Conditions 1–3 and let $\tilde{\theta}_0 \in \Theta$. Then, with probability one

$$\lim_{t\to\infty} \left\|g(\tilde{\theta}_t)\right\| = 0 \quad \text{or} \quad \tilde{\theta}_t \to \partial\Theta. \tag{31}$$

Proof: See Appendix C. ■

IV. Examples and Numerical Validation

Here, we consider two different example filtering problems and show explicitly how the parameter learning rules are derived. We also study the numerical performance of the learning method. Since under suitable conditions on the decay of the learning rate, convergence is guaranteed by the results in the preceding section, we do not study this case. Instead, we study whether the method also converges with constant learning rate, i.e., when violating Condition 3. A constant learning rate is a sensible choice when the system parameters are expected to change.

All numerical experiments use the Euler–Maruyama method to integrate the SDEs. We evaluate the performance of the learned filter by the mean-squared error (MSE), normalized by the variance of the hidden process.

A. One-Dimensional (1-D) Kalman–Bucy Filter (Linear Filtering Problem)

We shall first consider the simple case of the linear filtering problem, for which it is possible to obtain an exact finite-dimensional filter as well as exact expressions for the asymptotic likelihood. Here, we have a 3-D parameter vector $\theta = (a, \sigma, w)$, where $a, \sigma > 0$ and $w \in \mathbb{R}$, and we have $f(x,\theta) = -ax$, $g(x,\theta) = \sigma$, and $h(x,\theta) = wx$, such that the filtering problem reads

$$dX_t = -aX_t dt + \sigma dW_t, \quad dY_t = wX_t dt + dV_t. \tag{32}$$

Assuming a Gaussian initialization, i.e., $X_0 \sim \mathcal{N}(0, \sigma^2/2a)$, the optimal filter has a Gaussian distribution with mean μ_t and variance P_t (the Kalman–Bucy filter [2]). This is a 2-D representation with $M_t(\theta) = (\mu_t(\theta), P_t(\theta))^\top$, which can be expressed as

$$dM_t(\theta) = \begin{pmatrix} -a\mu_t(\theta) - w^2\mu_t(\theta)P_t(\theta) \\ \sigma^2 - 2aP_t(\theta) - w^2P_t(\theta)^2 \end{pmatrix} dt + \begin{pmatrix} wP_t(\theta) \\ 0 \end{pmatrix} dY_t. \tag{33}$$

We have $\psi_h(M_t(\theta),\theta) = w\mu_t(\theta)$.

Let us first calculate the asymptotic log-likelihood. It follows from the above that $P_t(\theta)$ (and its derivatives with respect to θ) will tend to a unique steady state given by

$$P_\infty(\theta) = \frac{1}{w^2}\left(\sqrt{a^2+w^2\sigma^2} - a\right). \tag{34}$$

By initializing the filter with this steady-state value, the representation can be made 1-D, i.e.,

$$\begin{aligned} dM_t(\theta) = &\left(-a\mu_t(\theta) - w^2\mu_t(\theta)P_\infty(\theta)\right)dt \\ &+ wP_\infty(\theta)dY_t. \end{aligned} \tag{35}$$

The process $\mathcal{X}_t(\theta)$ consisting of X_t, $\mu_t(\theta)$, and the filter derivatives $\mu_t^a(\theta), \mu_t^\sigma(\theta), \mu_t^w(\theta)$, therefore admits the SDE representation

$$d\mathcal{X}_t(\theta) = A\mathcal{X}_t(\theta)dt + B\begin{pmatrix} dW_t \\ dV_t \end{pmatrix} \tag{36}$$

with matrices, (37) shown at the bottom of this page and

$$B = \begin{pmatrix} \sigma_0 & 0 \\ 0 & wP \\ 0 & wP^a \\ 0 & wP^\sigma \\ 0 & P + wP^w \end{pmatrix} \tag{38}$$

where P is a shorthand for $P_\infty(\theta)$ and P^a, etc., are partial derivatives of $P_\infty(\theta)$.

The process $\mathcal{X}_t(\theta)$ is ergodic, and its unique invariant probability measure is multivariate Gaussian with zero mean and covariance matrix K given by the solution to

$$BB^\top + AK + KA^\top = 0. \tag{39}$$

In terms of this, the asymptotic log-likelihood reads

$$\begin{aligned} \tilde{\mathcal{L}}(\theta) &= ww_0K_{12} - \frac{1}{2}w^2K_{22} \\ &= \frac{P_\infty(\theta)w^2\sigma_0^2w_0^2(2a + P_\infty(\theta)w^2)}{4a_0(a+P_\infty(\theta)w^2)(a+a_0+P_\infty(\theta)w^2)} \\ &\quad - \frac{P_\infty(\theta)^2w^4}{4(a+P_\infty(\theta)w^2)}. \end{aligned} \tag{40}$$

With suitable boundaries of the parameter space, all the items from Condition 1 can be verified.

This model is nonidentifiable from the observations. The set of critical points of the asymptotic likelihood is characterized by

$$\partial_\theta\tilde{\mathcal{L}}(\theta) = 0 \Leftrightarrow \theta = \left(a_0, \sigma, \frac{w_0\sigma_0}{\sigma}\right)^\top, \quad \sigma > 0 \tag{41}$$

i.e., convergence can be guaranteed to one of these points only, and not to the ground truth parameters $\theta_0 = (a_0, \sigma_0, w_0)^\top$. The model becomes identifiable if either σ_0 or w_0 is known. Alternatively, one may fix a parameterization for which X_t has unit variance (i.e., $\sigma = \sqrt{2a}$).

Let us now derive the parameter update equations. The filtering equations for the mismatched filter, expressed in terms of the online parameter estimates, read

$$d\mu_t = -\tilde{a}_t\mu_t dt + \tilde{w}_tP_t(dY_t - \tilde{w}_t\mu_t dt), \quad \mu_0 = 0 \tag{42}$$

$$dP_t = \left(\tilde{\sigma}_t^2 - 2\tilde{a}_tP_t - \tilde{w}_t^2P_t^2\right)dt, \quad P_0 = \frac{\tilde{\sigma}_0^2}{2\tilde{a}_0} \tag{43}$$

where the initialization of P_0 reflects the prior belief of the variance of X_0 based on the initial parameter estimates.

The online parameter update equations read

$$d\tilde{a}_t = \gamma_a\tilde{a}_t\tilde{w}_t\mu_t^a\left(dY_t - \tilde{w}_t\mu_t dt\right) \tag{44}$$

$$d\tilde{\sigma}_t = \gamma_\sigma\tilde{\sigma}_t\tilde{w}_t\mu_t^\sigma\left(dY_t - \tilde{w}_t\mu_t dt\right) \tag{45}$$

$$d\tilde{w}_t = \gamma_w\tilde{w}_t\left(\mu_t + \tilde{w}_t\mu_t^w\right)\left(dY_t - \tilde{w}_t\mu_t dt\right). \tag{46}$$

In order to prevent sign changes of the parameters, we chose time-dependent learning rates that are proportional to the parameters ($\tilde{a}_t$ has to stay nonnegative because the filter equations turn unstable otherwise; for $\tilde{\sigma}_t$ and $\tilde{w}_t$, it is because of identifiability, i.e., the signs of σ and w are not identifiable from $\mathcal{F}_t^Y$). Here, we introduced the filter derivatives μ_t^a, μ_t^σ, and μ_t^w of the mean, which, together with the filter derivatives of the variance, satisfy the coupled system of SDEs

$$\begin{aligned} d\mu_t^a = &-\left[\mu_t + \left(\tilde{a}_t + \tilde{w}_t^2P_t\right)\mu_t^a + \tilde{w}_t^2\mu_tP_t^a\right]dt \\ &+ \tilde{w}_tP_t^a dY_t \end{aligned} \tag{47}$$

$$dP_t^a = -\left[2P_t + 2\left(\tilde{a}_t + \tilde{w}_t^2P_t\right)P_t^a\right]dt \tag{48}$$

$$\begin{aligned} d\mu_t^\sigma = &-\left[\left(\tilde{a}_t + \tilde{w}_t^2P_t\right)\mu_t^\sigma + \tilde{w}_t^2\mu_tP_t^\sigma\right]dt \\ &+ \tilde{w}_tP_t^\sigma dY_t \end{aligned} \tag{49}$$

$$dP_t^\sigma = \left[2\tilde{\sigma}_t - 2\left(\tilde{a}_t + \tilde{w}_t^2P_t\right)P_t^\sigma\right]dt \tag{50}$$

$$\begin{aligned} d\mu_t^w = &-\left[2\tilde{w}_t\mu_tP_t + \left(\tilde{a}_t + \tilde{w}_t^2P_t\right)\mu_t^w\right]dt \\ &- \tilde{w}_t^2\mu_tP_t^w dt + \left[P_t + \tilde{w}_tP_t^w\right]dY_t \end{aligned} \tag{51}$$

$$dP_t^w = -\left[2\tilde{w}_tP_t^2 + 2\left(\tilde{a}_t + \tilde{w}_t^2P_t\right)P_t^w\right]dt \tag{52}$$

$$\mu_0^a = \mu_0^\sigma = \mu_0^w = 0 \tag{53}$$

$$P_0^a = -\frac{\tilde{\sigma}_0^2}{2\tilde{a}_0^2}, \quad P_0^\sigma = \frac{\tilde{\sigma}_0}{\tilde{a}_0}, \quad P_0^w = 0. \tag{54}$$

The right-hand sides (RHSs) of the filter derivative equations and the initial conditions of the filter derivatives are obtained

$$A = \begin{pmatrix} -a_0 & 0 & 0 & 0 & 0 \\ ww_0P & -a - w^2P & 0 & 0 & 0 \\ ww_0P^a & -1 - w^2P^a & -a - w^2P & 0 & 0 \\ ww_0P^\sigma & -w^2P^\sigma & 0 & -a - w^2P & 0 \\ w_0(P + wP^w) & -w^2P^w & 0 & 0 & -a - w^2P \end{pmatrix} \tag{37}$$

from the corresponding equations of the filtered mean and variance and their initial conditions by differentiating with respect to each of the parameters (see Section II for details).

First, we investigated one of the cases where the model is identifiable, i.e., the parameter w was assumed to be known and we set $\tilde{w}_0 = w_0 = 3$ and $\gamma_w = 0$. The performance of the algorithm is visualized in Fig. 1 where the learning process is shown in a single trial, and in Fig. 2, where we show trial-averaged learning curves for the MSE and the parameter estimates. For both figures, the ground truth parameters were set to $a_0 = 1$ and $\sigma_0 = 2$, and the initial parameter estimates were $\tilde{a}_0 = 10$ and $\tilde{\sigma}_0 = \sqrt{0.2}$, making for a strongly mismatched model that produces an MSE close to 1 without learning, i.e., with all learning rates set to zero. With constant learning rates $\gamma_a = \gamma_\sigma = 0.03$, the filter performance can be improved to almost optimal performance within a time frame of $T = 1000$, after which the parameter estimates approach the ground truth. The log-likelihood function is not globally concave, but it has a single global maximum [see Fig. 3 (left)].

The comparison to online expectation maximization (EM) (see Section V-C) is shown in Fig. 3 (center and right). Here, only the parameter a is learned, while $\sigma = \sigma_0 = 2$ and $w = w_0 = 3$. The simulations suggest that online is slightly faster in the beginning, but the order of convergence is similar for SGA and online EM. This goes along with very similar computational complexity: the number of SDEs that has to be integrated is the same for SGA and online EM.

B. Bimodal State and Linear Observation Model With (Approximate) Projection Filter

Consider the following system with four positive parameters (a, b, σ, w):

$$dX_t = X_t\left(a - bX_t^2\right)dt + \sigma dW_t \tag{55}$$

$$dY_t = wX_t dt + dV_t. \tag{56}$$

In this problem, the hidden state X_t has a bimodal stationary distribution with modes at $x = \pm\sqrt{a/b}$. Since the observation model is linear like in Section IV-A, the parameter learning rules are expressed in terms of the posterior mean $\mu_t = \hat{X}_t$ as

$$d\tilde{a}_t = \gamma_a \tilde{a}_t \tilde{w}_t \mu_t^a\left(dY_t - \tilde{w}_t\mu_t dt\right) \tag{57}$$

$$d\tilde{b}_t = \gamma_b \tilde{b}_t \tilde{w}_t \mu_t^b\left(dY_t - \tilde{w}_t\mu_t dt\right) \tag{58}$$

$$d\tilde{\sigma}_t = \gamma_\sigma \tilde{\sigma}_t \tilde{w}_t \mu_t^\sigma\left(dY_t - \tilde{w}_t\mu_t dt\right) \tag{59}$$

$$d\tilde{w}_t = \gamma_w \tilde{w}_t\left(\mu_t + \tilde{w}_t\mu_t^w\right)\left(dY_t - \tilde{w}_t\mu_t dt\right). \tag{60}$$

We have made the learning rules proportional to the parameters in order to prevent sign changes, i.e., to guarantee that all parameters remain positive. In contrast to the linear model in Section IV-A, the filtering problem is not exactly solvable. We use the projection filter on the manifold of Gaussian densities introduced by [12], or equivalently, the Gaussian ADF in Stratonovich calculus. The mean μ_t and variance P_t of the Gaussian approximation to the filter evolve as

$$\begin{aligned} d\mu_t &= \left[\tilde{a}_t\mu_t - \tilde{b}_t\mu_t^3 - \left(3\tilde{b}_t + \tilde{w}_t^2\right)\mu_t P_t\right]dt \\ &\quad + \tilde{w}_t P_t dY_t, \quad \mu_0 = 0 \end{aligned} \tag{61}$$

$$dP_t = \left[\tilde{\sigma}_t^2 + \left(2\tilde{a}_t - \tilde{w}_t^2 P_t^2 - 6\tilde{b}_t(\mu_t^2 + P_t)\right)P_t\right]dt \tag{62}$$

where the initial variance as a function of the initial parameter estimates is the variance of the stationary distribution obtained by solving the equation $\mathcal{A}^\dagger = 0$

$$P_0 = \Gamma\left(\tilde{a}_0, \tilde{b}_0, \tilde{\sigma}_0\right) = \frac{\int_{-\infty}^{\infty} x^2 e^{\tilde{\sigma}_0^{-2}\left(\tilde{a}_0 x^2 - \frac{1}{2}\tilde{b}_0 x^4\right)}dx}{\int_{-\infty}^{\infty} e^{\tilde{\sigma}_0^{-2}\left(\tilde{a}_0 x^2 - \frac{1}{2}\tilde{b}_0 x^4\right)}dx}. \tag{63}$$

By differentiating (61) and (62) with respect to the parameters, we obtain the following equations for the filter derivatives:

$$d\mu_t^a = \left[\mu_t + \alpha_t\mu_t^a + \beta_t P_t^a\right]dt + \tilde{w}_t P_t^a dY_t \tag{64}$$

$$dP_t^a = \left[2P_t + A_t\mu_t^a + B_t P_t^a\right]dt \tag{65}$$

$$\begin{aligned} d\mu_t^b &= \left[-\mu_t\left(\mu_t^2 + 3P_t\right) + \alpha_t\mu_t^b + \beta_t P_t^b\right]dt \\ &\quad + \tilde{w}_t P_t^a dY_t \end{aligned} \tag{66}$$

$$dP_t^b = \left[-6P_t\left(\mu_t^2 + P_t\right) + A_t\mu_t^b + B_t P_t^b\right]dt \tag{67}$$

$$d\mu_t^\sigma = \left[\alpha_t\mu_t^\sigma + \beta_t P_t^\sigma\right]dt + \tilde{w}_t P_t^\sigma dY_t \tag{68}$$

$$dP_t^\sigma = \left[2\tilde{\sigma}_t + A_t\mu_t^\sigma + B_t P_t^\sigma\right]dt \tag{69}$$

$$\begin{aligned} d\mu_t^w &= \left[-2\tilde{w}_t\mu_t P_t + \alpha_t\mu_t^w + \beta_t P_t^w\right]dt \\ &\quad + \left[P_t + \tilde{w}_t P_t^w\right]dY_t \end{aligned} \tag{70}$$

$$dP_t^w = \left[-2\tilde{w}_t P_t^2 + A_t\mu_t^w + B_t P_t^w\right]dt \tag{71}$$

$$\mu_0^a = \mu_0^b = \mu_0^\sigma = \mu_0^w = 0 \tag{72}$$

$$P_0^a = \frac{\partial}{\partial\tilde{a}_0}\Gamma\left(\tilde{a}_0, \tilde{b}_0, \tilde{\sigma}_0\right), \quad P_0^b = \frac{\partial}{\partial\tilde{b}_0}\Gamma\left(\tilde{a}_0, \tilde{b}_0, \tilde{\sigma}_0\right) \tag{73}$$

$$P_0^\sigma = \frac{\partial}{\partial\tilde{\sigma}_0}\Gamma\left(\tilde{a}_0, \tilde{b}_0, \tilde{\sigma}_0\right), \quad P_0^w = 0 \tag{74}$$

where we introduced the following auxiliary processes:

$$\alpha_t = \tilde{a}_t - \tilde{w}_t^2 P_t - 3\tilde{b}_t\left(\mu_t^2 + P_t\right) \tag{75}$$

$$\beta_t = -\left(\tilde{w}_t^2 + 3\tilde{b}_t\right)\mu_t \tag{76}$$

$$A_t = -12\tilde{b}_t\mu_t P_t \tag{77}$$

$$B_t = 2\tilde{a}_t - 2\tilde{w}_t^2 P_t - 6\tilde{b}_t\left(\mu_t^2 + 2P_t\right). \tag{78}$$

We numerically tested the learning algorithm for this nonlinear model by simulating a system with $a_0 = 4$, $b_0 = 3$, $\sigma_0 = 1$, and $w_0 = 2$, leading to a variance $\mathrm{Var}(X_t) = 1.17$. Initial parameter estimates were set to a permutation of the ground truth, i.e., $\tilde{a}_0 = 1$, $\tilde{b}_0 = 2$, $\tilde{\sigma}_0 = 3$, and $\tilde{w}_0 = 4$ and the simulations lasted $T = 2000$ (due to the longer time scale compared to the linear model) with a time step of $dt = 10^{-3}$. In Fig. 4, we show an example of the learning process.

In this case, the suboptimality of the Gaussian approximation inherent in the projection filter allows the filter error (MSE) to

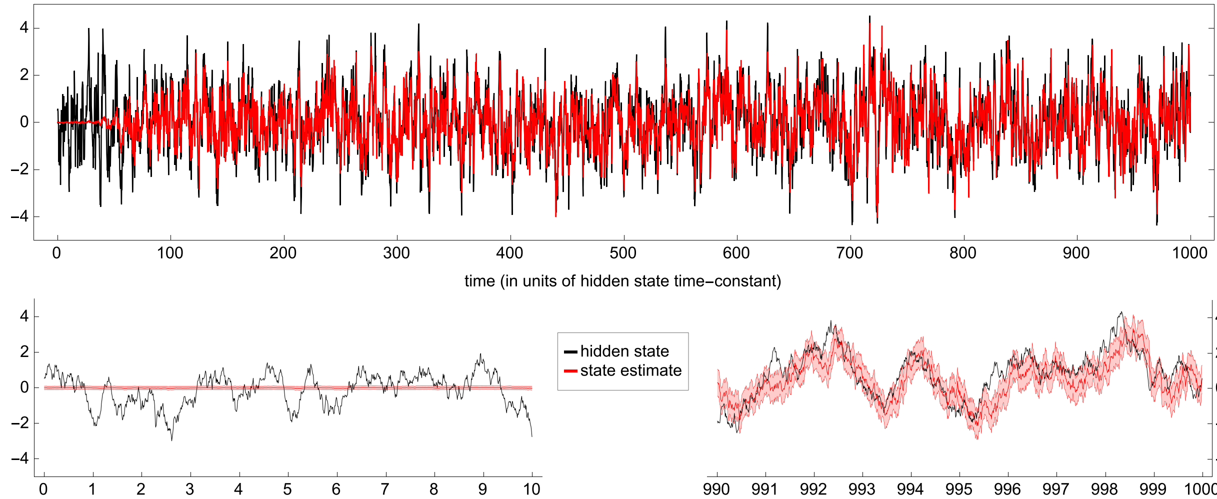

Fig. 1. *Online learning and filtering in the linear model.* The hidden state X_t (black) and Kalman–Bucy state estimate μ_t [red, shaded region shows $\mu_t \pm$ one standard deviation $\sqrt{P_t}$, cf., (42) and (43)] are shown for the linear model of Section IV-A with parameters $a_0 = 1$, $\sigma_0 = 2$, and $w_0 = 3$. The time step is $dt = 10^{-3}$, initial parameter estimates are $\tilde{a}_0 = 10$, $\tilde{\sigma}_0 = \sqrt{0.2}$, $\tilde{w}_0 = 3$ (i.e., the parameter w_0 is known), and the learning rates are $\gamma_a = \gamma_\sigma = 0.03$ and $\gamma_w = 0$. Top: The entire learning period of $T = 1000$ shows a gradual improvement of the performance of the filter. Bottom left: During the first 10 s, the model is still strongly mismatched. Bottom right: During the last 10 s, the filter optimally tracks the hidden state.

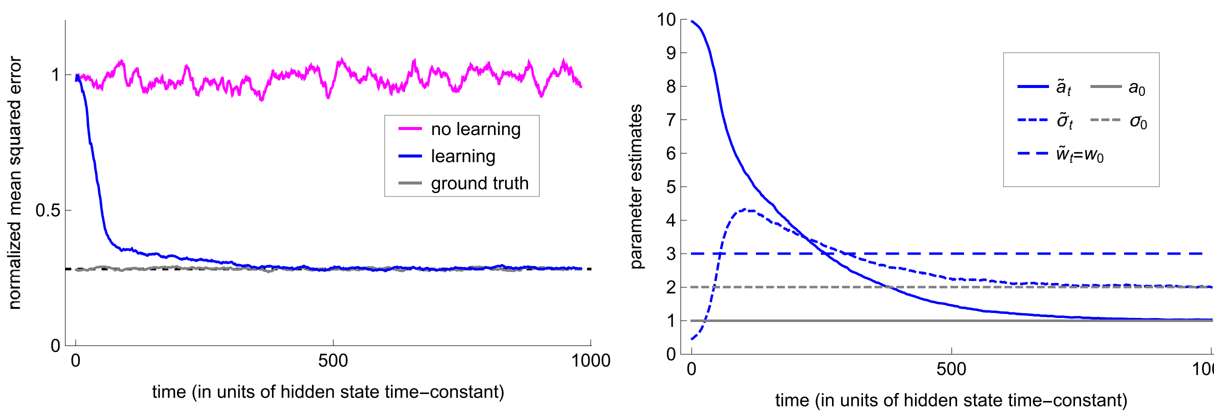

Fig. 2. *Online learning and filtering in the linear model.* The time evolution of the MSE and parameter estimates is shown for the linear model of Section IV-A (see Fig. 1 caption for details). Left: The moving average of the normalized MSE (time window of 20 s) shows how the learning algorithm leads to a gradual improvement of the performance of the filter, which eventually reaches the performance of an optimal Kalman–Bucy filter with ground truth parameters. The black, dashed line shows the theoretical result for the performance of the Kalman–Bucy filter. Right: The parameter estimates for the unknown parameters converge to the ground truth parameters. All curves are trial averaged ($N = 100$ trials).

be lower with learning than with the ground truth parameters in the absence of learning, getting close to the performance of the optimal filter. This is shown in Fig. 5 in terms of trial-averaged learning curves. The normalized MSE with learning decreases within the time frame of $T = 2000$ and converges below the MSE for the projection filter with fixed parameters set to the ground truth. The optimal performance was estimated by running a particle filter with prior importance function, resampling at every time step, 1000 particles, and parameters set to the ground truth [29].

V. Related Approaches

In this section, we attempt to review similar approaches for online maximum-likelihood estimation, and their relations to our method. We note that most of the literature on this topic is formulated for discrete-time systems, and we realize that the list of reviewed works is not exhaustive. Some of the approaches for hidden Markov models (HMMs) discussed here are also surveyed in more detail in [30]–[32].

A. Recursive Maximum-Likelihood Approaches

This paper is the continuous-time analog of the online SGA algorithm of [17] and[33] for HMMs. The behavior of the algorithm is analyzed by casting it in the Robbins–Monro framework of stochastic approximations. We used a similar approach to studying convergence in Section III. More recently, the convergence of discrete-time stochastic gradient algorithms for parameter estimation in HMMs was studied under more general conditions [34]. To our knowledge, it is an open problem to obtain a similarly general result for continuous-time models such

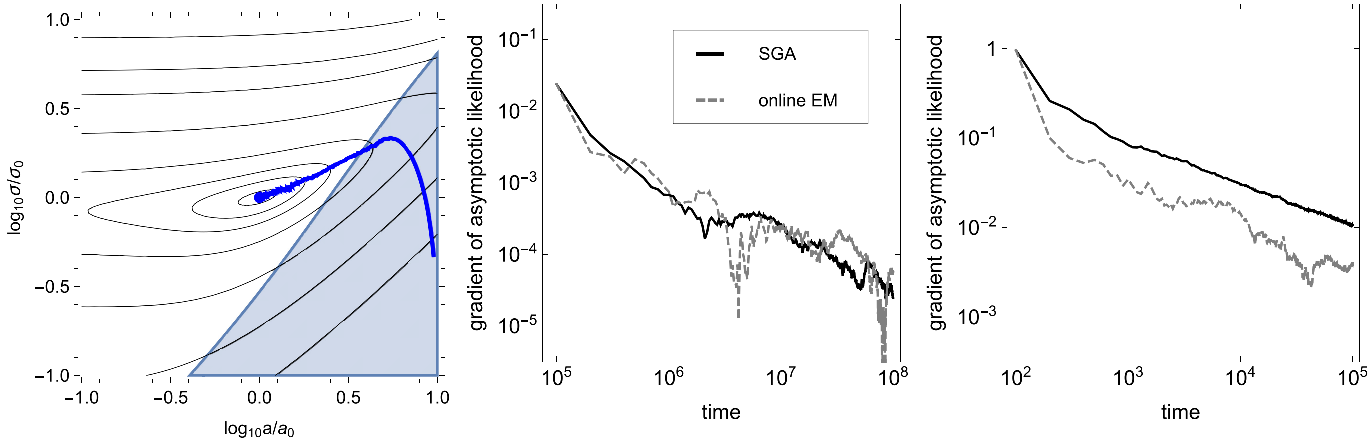

Fig. 3. *Parameter estimation in the linear model*. Left: The asymptotic log-likelihood function from (40) in the parameter subspace spanned by a and σ for $w = w_0 = 3$ has a single global maximum near $a = a_0$ and $\sigma = \sigma_0$. The shading shows the region where the function is nonconcave, and the thick black line is the trial-averaged learning trajectory from Fig. 2. Center: Single-trial convergence of the absolute value of the gradient of the asymptotic log-likelihood evaluated at the online parameter estimate given by SGA and online EM, respectively. The rate of convergence is similar for both algorithms. Right: An average over 10 samples reveals that the online EM is initially faster and more variable, but the order of convergence is the same. Time is measured in units of the intrinsic time constant $1/a_0$.

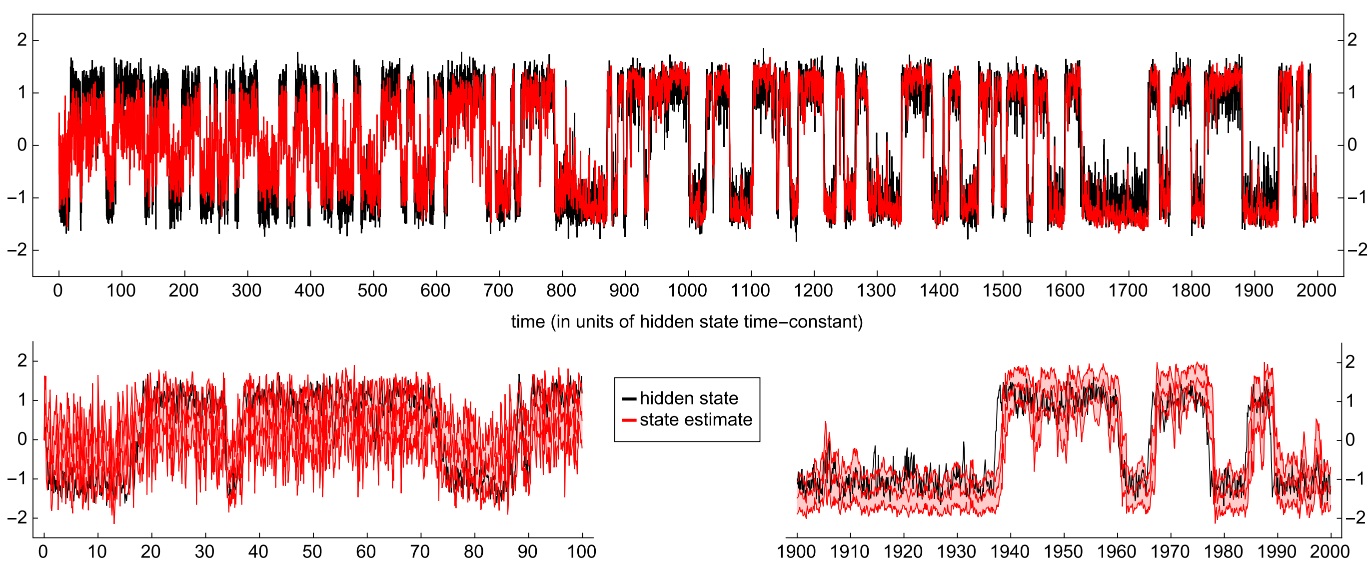

Fig. 4. *Online learning and filtering in the nonlinear model*. The hidden state X_t (black) and mean μ_t of the projection filter are shown for the bimodal model of Section IV-B with parameters $a_0 = 4$, $b_0 = 3$, $\sigma_0 = 1$, $w_0 = 2$, $\tilde{a}_0 = 1$, $\tilde{b}_0 = 2$, $\tilde{\sigma}_0 = 3$, $\tilde{w}_0 = 4$, $\gamma_a = \gamma_b = \gamma_w = 10^{-1}$, and $\gamma_\sigma = 0.04$. Top: The entire learning period of $T = 2000$ shows an improvement in both step size between the two attractors and the variability within both attractors. Bottom left: During the first 100 s, the filter is too sensitive to observations and has an incorrect spacing between attractors. Bottom right: During the last 100 s, the filter shows good tracking performance.

as the one in this paper. Regarding estimation in a discretized diffusion model, particle algorithms have been discussed in [35] and [36].

B. Prediction Error Algorithms

Another stochastic approximation scheme is the recursive minimum prediction error scheme (see [17] and [37]) for HMMs. Instead of finding maxima of the likelihood, it finds minima of the average (squared) prediction error, i.e., the error between the observations and the predicted observations. In our continuous-time model, the prediction error is given by the infinitesimal pseudoinnovation increment $dY_t - \tilde{h}_t dt$. Formal differentiation of $(dY_t - \tilde{h}_t dt)^2$ with respect to the parameter yields the same parameter update rule as that derived in Section II. While a rigorous analysis has not been done, it seems natural to conjecture that recursive maximum likelihood and recursive minimum prediction error are equivalent in continuous time.

C. Online EM

EM is a well-known method for offline parameter learning in partially observed stochastic systems [38], [39]. It is based on

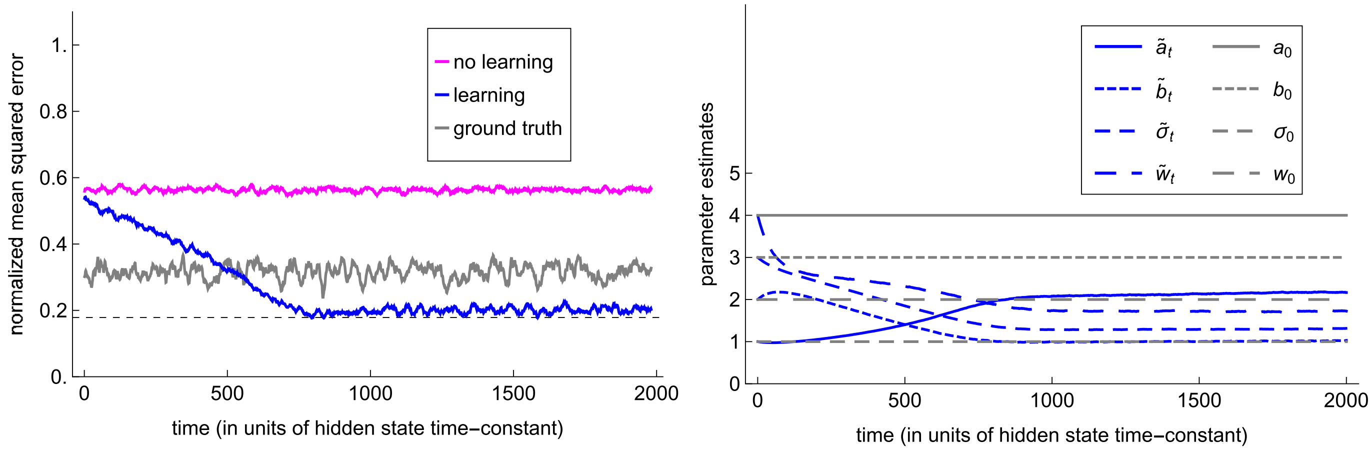

Fig. 5. *Online learning and filtering in the nonlinear model.* The time evolution of the MSE and parameter estimates is shown for the bimodal model of Section IV-B (see Fig. 4 caption for details). Left: The moving average of the normalized MSE (time window of 20 s) shows how the learning algorithm allows the filter performance to improve to a level that is better than that of a filter with fixed parameters set to the ground truth. However, it is still slightly worse than an optimal filter; the dashed black line shows the performance of a particle filter with 1000 particles with parameters set to the ground truth. Right: Despite the low filter error, the parameter estimates do *not* converge to the ground truth. All curves are trial averaged ($N = 100$ trials).

the following application of Jensen's inequality:

$$\begin{aligned}\mathcal{L}_t(\theta) - \mathcal{L}_t(\tilde{\theta}) &= \log \mathbb{E}_{\tilde{\theta}}\left[\frac{dP_\theta}{dP_{\tilde{\theta}}}\middle|\mathcal{F}_t^Y\right] \\ &\geq \mathbb{E}_{\tilde{\theta}}\left[\log\frac{dP_\theta}{dP_{\tilde{\theta}}}\middle|\mathcal{F}_t^Y\right] \doteq Q_t(\theta,\tilde{\theta}).\end{aligned} \tag{79}$$

Since $Q_t(\tilde{\theta},\tilde{\theta}) = 0$, by maximizing $Q_t(\theta,\tilde{\theta})$ with respect to θ (for fixed $\tilde{\theta}$), we obtain a nonnegative change in the likelihood. EM thus produces a sequence of parameter estimates $\tilde{\theta}_k$, $k = 0, 1, 2, \ldots$, with nondecreasing likelihood by iterating the following procedure: compute the quantity $Q_t(\theta, \tilde{\theta}_k)$ (the "expectation" or "E" step in EM), then set $\tilde{\theta}_{k+1} = \text{argmax}_\theta Q_t(\theta, \tilde{\theta}_k)$ (the "maximization" or "M" step in EM).

If a parameterization is chosen such that the complete-data log-likelihood[5] takes the form of an exponential family, i.e., $\Psi(\theta)\cdot S_t$, where Ψ is a vector-valued function of the parameters and S_t is a vector of functionals of the hidden state and observation trajectories, then $Q_t(\theta,\tilde{\theta}) = \Psi(\theta)\cdot\hat{S}_t(\tilde{\theta}) + R(\tilde{\theta})$, where

$$\hat{S}_t(\tilde{\theta}) = \mathbb{E}_{\tilde{\theta}}\left[S_t\middle|\mathcal{F}_t^Y\right] \tag{80}$$

and $R(\tilde{\theta})$ is independent of θ. The "M" step can be done explicitly if the equation $\partial_\theta\Psi(\theta)\cdot\hat{S}_t(\tilde{\theta}) = 0$ has a unique closed-form solution. Meanwhile, the "E" step consists of computing $\hat{S}_t(\tilde{\theta})$, which involves certain nonlinear smoothed functionals of the forms

$$\mathbb{E}_{\tilde{\theta}}\left[\int_0^t \varphi_1(X_s)dX_s\middle|\mathcal{F}_t^Y\right]$$
$$\mathbb{E}_{\tilde{\theta}}\left[\int_0^t \varphi_2(X_s)dY_s\middle|\mathcal{F}_t^Y\right]$$
$$\mathbb{E}_{\tilde{\theta}}\left[\int_0^t \varphi_3(X_s)ds\middle|\mathcal{F}_t^Y\right]$$

with possibly distinct integrands φ_1, φ_2, and φ_3. In general, these smoothed functionals are computed using a forward–backward smoothing algorithm, which is not suitable for online learning. In a few select cases, the smoothed functionals admit a finite-dimensional solution (see [41] and the remarks on [39, p. 99]), or even a finite-dimensional *recursive* solution (see [40], [42], and [43]).

In [40], the smoothed functionals of the linear-Gaussian model are expressed (using the Fisher identity) in terms of derivatives of the incomplete-data log-likelihood, or a generalization thereof. This enables a recursive computation of the smoothed quantities of interest, and the auxiliary variables that need to be integrated (called sensitivity equations) are very similar to (47)–(52). The relation between smoothed functionals and the sensitivity equations has been known for a long time (see [30, Sec. 10.2] and [44]).

Several authors [45]–[48] have introduced the idea of a fully recursive form of EM, called *online EM*. In the papers presented above, online EM has been explicitly formulated for HMMs and state-space models by integrating the recursive smoothing algorithm using the online parameter estimate. This stochastic approximation approach to EM is thus very similar to the gradient-based approach used here and in the references discussed in Section V-A.

In continuous-time diffusion models such as studied in this paper, the recursions found by [40],[42], and [43] can be

[5]We note that a limitation of EM in the continuous-time model is that the identification of parameters of the diffusion term g_θ has to be treated differently from that of drift parameters in f_θ and h_θ. This is due to the fact that there is no reference measure for the complete model that is independent of the diffusion parameters. The parameters of the diffusion term are therefore not included in θ, but are estimated separately from the quadratic variations of hidden state and observation. This issue is discussed in more detail in [40], see Section IV-B. This issue is avoided in the gradient-based method here because the reference measure restricted to the observations is independent of all parameters, including the ones of the diffusion term.

directly applied if the model is linear. We did this in order to do the comparison of SGA and online EM shown in Fig. 3. For this particular model, SGA and online EM are comparable in terms of computational complexity and rate of convergence. In nonlinear models, online EM can be formulated by making use of recursive particle approximations of the smoothing functionals (e.g., by applying the methods in [49] and [50] to a suitable time discretization of the SDEs). As an alternative, ADFs or projection filters can be used to approximate the recursive smoothed functionals. The full development of online EM in continuous time, as well as its convergence analysis, remains a topic for future research.

D. State Augmentation Algorithms

The idea is to treat the unknown parameter as a random variable that is either static ($d\theta_t = 0$) or has dynamics that are coupled to the hidden state. In both cases, the parameter may be estimated online by solving the filtering problem for the augmented state (X_t, θ_t). While this presents clear advantages for known dynamics of the hidden parameter, it introduces a new parameter estimation problem for the parameters of the dynamics of θ_t, called *hyperparameters*. A static prior for θ_t is problematic because the resulting filter will usually not be stable, with negative implications (see [51]) on the behavior of particle filters that are needed to solve the augmented filtering problem (but see [52], where stability conditions are discussed for the discrete-time case). In addition, for many interesting models, the parameter space may be of much higher dimension than the state space, introducing high computational costs for filtering of the augmented state.

E. Maximum-Likelihood Filtering and Identification

The opposite of state augmentation was explored in [53], where the hidden state is also estimated via maximum likelihood, instead of the usual filtering paradigm using minimum MSE. Equations for the maximum-likelihood state and parameter estimates are then derived. Although these equations are not directly suitable for recursive identification, they are very similar to the ones obtained by us in Section II. It remains a curiosity that the approach of [53] has rarely been cited and has not been further developed.

VI. Conclusion

The problem of estimating parameters in partially observed systems is old and relevant to many applications. However, the majority of the literature on this subject is written for discrete-time processes and for offline learning, whereas, despite of its enormous importance for filtering and control theory, the continuous-time case has received little attention. Online gradient ascent in continuous time has only recently been studied in [18]. The use of a change of measure in order to express the likelihood function in terms of the filter is not new, but it seems to be underexploited. To the best of our knowledge, its only use in parameter estimation is in the technical report presented in [53]. We found it appropriate to revisit this approach and to extend the work of [18] to the partially observable case. Recently, the above results for the fully observed model have been strengthened to a central limit theorem for the parameter estimate, see [54].

The main difficulty and open problem is to find conditions on the generative model that are easy to verify, sufficient for the convergence of the algorithm, and not too restrictive. Currently, the most promising avenue for obtaining such conditions is by settling open questions regarding the ergodicity of the approximate filter in terms of the exact one, and then using the general theory that guarantees ergodicity of the exact filter. The latter is relatively easy to check compared to the explicit conditions on the approximate filter that we currently give in Section III-A. We hope that these open questions will be addressed in the future.

Let us briefly comment on the numerical examples that we provided. As we showed numerically, the algorithm is capable of improving filter performance even if the models are unidentifiable and the learning rate constant, even though this cannot be expected. In addition, the second numerical example showed that the performance of the filter can be improved even beyond what is possible with fixed parameters. This result could lead to new ways of improving the performance of approximate filters by using the additional degrees of freedom given by the online parameter estimates for both adaptation (learning) and reduced filter error. It remains to be explored whether this feature applies to a large enough class of approximate filters to be useful for practical applications.

As we showed also in comparison with the online EM algorithm, these naïve methods exhibit rather slow convergence rates and cannot compete with fast offline methods such as second-order optimization methods or Nesterov's accelerated gradient. However, the main aim of this paper is to advance the theoretical understanding of convergence using continuous-time theory. Based on this, it remains a topic for future research to study the convergence of more elaborate algorithms such as the ones mentioned above.

Appendix A
Proof of Proposition 1

i) We have

$$\frac{1}{t}\mathcal{L}_t(\theta) = \frac{1}{t}\int_0^t l(\mathcal{X}_s(\theta), \theta)ds + \frac{1}{t}\int_0^t \psi_h(\theta, M_s(\theta)) \cdot dV_s. \quad (81)$$

By Condition 1(i), the first term on the RHS converges to $\int_{\mathbb{R}^D} l(x,\theta)\mu_\theta(dx) = \tilde{\mathcal{L}}(\theta)$ a.s. as $t \to \infty$. Consider the local martingale $\mathfrak{M}_t = \int_0^t \psi_h(\theta, M_s(\theta)) \cdot dV_s$. From Itô isometry, and Conditions 2 and 1(v), it follows that for t

large enough

$$\begin{aligned}\mathbb{E}&\left[\left(\int_0^t \psi_h(\theta, M_s(\theta))\cdot dV_s\right)^2\right]\\ &= \mathbb{E}\left[\int_0^t \|\psi_h(\theta, M_s(\theta))\|^2\, ds\right]\\ &\le \mathbb{E}\left[\int_0^t C(1+||M_s(\theta)||^q)ds\right]\\ &\le \mathbb{E}\left[\int_0^t C(1+||\mathcal{X}_s(\theta)||^q)ds\right]\\ &\le Ct\left(1+\mathbb{E}[\sup_{s\le t}||\mathcal{X}_s(\theta)||^q]\right)\\ &\le Ct(1+C'\sqrt{t}).\end{aligned} \tag{82}$$

In short, for t large enough, we have $\text{Var}[\mathfrak{M}_t] \le Kt^{3/2}$ for some $K > 0$. Therefore

$$\text{Var}\left[\tfrac{1}{t}\mathfrak{M}_t\right] \le Kt^{-1/2} \to 0, \quad t\to\infty \tag{83}$$

which means that the second term on the RHS of (81) converges to zero in L_2.

Now consider the process $\tilde{\mathfrak{M}}_t = \frac{1}{t}\mathfrak{M}_t + \int_0^t \frac{1}{s^2}\mathfrak{M}_s ds$. By Itô's lemma, this process is the local martingale given by $\int_0^t \frac{1}{s}\psi_h(M_s(\theta),\theta)\cdot dV_s$. By applying Itô isometry, and Conditions 1(v) and 2, we obtain

$$\begin{aligned}\sup_{t>0}\mathbb{E}\left[||\tilde{\mathfrak{M}}_t||^2\right] &\le \int_0^\infty \frac{\mathbb{E}\left[||\psi_h(M_s(\theta),\theta)||^q\right]}{s^2}ds\\ &\le K\int_0^\infty \frac{1}{s^2}\left(1+\mathbb{E}\left[||\mathcal{X}_s(\theta)||^2\right]\right)ds < \infty.\end{aligned} \tag{84}$$

By the martingale convergence theorem, there is a finite random variable $\tilde{\mathfrak{M}}_\infty$ such that $\tilde{\mathfrak{M}}_t \to \tilde{\mathfrak{M}}_\infty$ a.s. and in L_2. Therefore also $\frac{1}{t}\mathfrak{M}_t$ converges a.s.

ii) We have that $\partial_\theta \tilde{\mathcal{L}}(\theta) = \lim_{t\to\infty}\frac{1}{t}\partial_\theta\mathcal{L}_t(\theta)$, if the derivative exists and the limit exists a.s. Due to Condition 2, the derivative

$$\begin{aligned}\frac{1}{t}\partial_\theta\mathcal{L}_t(\theta) &= \frac{1}{t}\int_0^t \partial_\theta l(\mathcal{X}_s(\theta),\theta)ds\\ &\quad+\frac{1}{t}\int_0^t \partial_\theta\psi_h(M_s(\theta),\theta)dV_s\\ &= \frac{1}{t}\int_0^t F(\mathcal{X}_s(\theta),\theta)^\top ds\\ &\quad+\frac{1}{t}\int_0^t dV_s^\top H(M_s(\theta),\theta)\end{aligned} \tag{85}$$

exists. This converges to $\int_{\mathbb{R}^D} F(x,\theta)^\top\mu_\theta(dx)$ by an argument analogous to the one in (i).

The representation of $\mathcal{H}$ in terms of the invariant measure and its derivative follows from Conditions 1(iii) and 2.

iii) This follows from (i) and the fact that F is in $\bar{\mathbb{G}}$ (see Condition 2).

iv) By Condition 2, $q, K > 0$ can be chosen such that the functions $l, F, \partial_\theta F, H$ grow at most as $K(1+||x||^q)$ for all $\theta\in\Theta$. From this and the first part of the present Lemma, it follows that

$$\begin{aligned}\tilde{\mathcal{L}}(\theta) &= \int_{\mathbb{R}^D} l(x,\theta)\mu_\theta(dx)\\ &\le K\int_{\mathbb{R}^D}(1+||x||^q)\mu_\theta(dx) \le KK_q.\end{aligned} \tag{86}$$

By a similar calculation, we have

$$\|g(\theta)\| \le KK_q. \tag{87}$$

For $\|\mathcal{H}(\theta)\|$, observe that

$$\begin{aligned}\|\mathcal{H}(\theta)\| &\le \left\|\int_{\mathbb{R}^D}\partial_\theta F(x,\theta)\mu_\theta(dx)\right\|\\ &\quad+\left\|\int_{\mathbb{R}^D}F(x,\theta)\nu_\theta(dx)\right\|\\ &\le KK_q + \left\|\int_{\mathbb{R}^D}F(x,\theta)\nu_\theta(dx)\right\|\end{aligned} \tag{88}$$

where the first term on the RHS was treated in the same way as in the bound for $\tilde{\mathcal{L}}(\theta)$ and $\|g(\theta)\|$. For the second term, we observe that

$$\begin{aligned}&\left\|\int_{\mathbb{R}^D}F(x,\theta)\nu_\theta(dx)\right\|^2 = \sum_{i,j=1}^p\left(\int_{\mathbb{R}^D}F_i(x,\theta)\nu_{\theta,j}(dx)\right)^2\\ &\le \sum_{i,j=1}^p\left(\int_{\mathbb{R}^D}|F_i(x,\theta)|\,|\nu_{\theta,j}(dx)|\right)^2\\ &\le \sum_{i,j=1}^p\left(\int_{\mathbb{R}^D}||F(x,\theta)||\,|\nu_{\theta,j}(dx)|\right)^2 \le p^2K^2K_q'^2.\end{aligned} \tag{89}$$

The claimed inequality (30) then follows by setting $C = 3KK_q + pKK_q'$. ■

Appendix B
Lemmas

Here, we adapt the lemmas of [18] to fit the present setting. As in [18], the proofs of the lemmas require results from [21], but in a slightly more general form than what was needed in [18]. Despite the strong similarities between our proofs and the proofs in [18], for the convenience of the reader we shall write them out in full detail and in the appropriate notation.

For Lemmas 1–4, we assume that Conditions 1–3 hold and that the first exit time from Θ is infinite (see the proof of Theorem 1). In addition, we define the following. Let $\kappa, \lambda > 0$ and define the $(P_{\theta_0}, \mathcal{F}_t)$-stopping times $\sigma_0 = 0$ and $\sigma_k, \tau_k, k \in \mathbb{N}$ as

$$\tau_k \doteq \inf\left\{t > \sigma_{k-1} : \|g(\tilde{\theta}_t)\| \geq \kappa\right\} \tag{90}$$

$$\begin{aligned}\sigma_k \doteq \sup\Big\{t > \tau_k : \tfrac{1}{2}\|g\left(\tilde{\theta}_{\tau_k}\right)\| \leq \|g(\tilde{\theta}_s)\| \\ \leq 2\|g\left(\tilde{\theta}_{\tau_k}\right)\|, s \in [\tau_k, t] \text{ and } \int_{\tau_k}^{t} \gamma_s ds \leq \lambda\Big\}.\end{aligned} \tag{91}$$

Lemma 1: Let $\eta > 0$ and define

$$\Gamma_{k,\eta} \doteq \int_{\tau_k}^{\sigma_k+\eta} \gamma_s \left(F\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) - g\left(\tilde{\theta}_s\right)^\top\right) ds. \tag{92}$$

Then, with probability one

$$\lim_{k\to\infty} \|\Gamma_{k,\eta}\| = 0. \tag{93}$$

Proof: Consider the function $G(x, \theta) = F(x, \theta) - g(\theta)^\top$. By definition, we have

$$\int_{\mathbb{R}^D} G(x, \theta)\mu_\theta(dx) = 0 \tag{94}$$

and by Condition 2 we have that the components of $G(x, \cdot)$ are in $\bar{\mathbb{G}}^D$. Therefore, by Condition 1(iv), the Poisson equation

$$\mathcal{A}_\mathcal{X} v(x, \theta) = G(x, \theta), \quad \int_{\mathbb{R}^D} v(x, \theta)\mu_\theta(dx) = 0 \tag{95}$$

has a unique twice differentiable solution with

$$||v(x, \theta)|| + ||\partial_\theta v(x, \theta)|| + ||\partial_\theta^2 v(x, \theta)|| \leq K'(1 + ||x||^{q'}). \tag{96}$$

Let $u(t, x, \theta) = \gamma_t v(x, \theta)$, and apply Itô's lemma to each component of u

$$\begin{aligned}u_i(\sigma, \tilde{\mathcal{X}}_\sigma, \tilde{\theta}_\sigma) - u_i(\tau, \tilde{\mathcal{X}}_\tau, \tilde{\theta}_\tau) = \int_\tau^\sigma \partial_s u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s) ds \\ + \int_\tau^\sigma \mathcal{A}_\mathcal{X} u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s) ds + \int_\tau^\sigma \mathcal{A}_\theta u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s) ds \\ + \int_\tau^\sigma \gamma_s \mathrm{tr}\left[\hat{\Sigma}\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \partial_\theta^\top \partial_x u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s)\right] ds \\ + \int_\tau^\sigma \partial_x u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s)\Sigma\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) d\mathcal{B}_s \\ + \int_\tau^\sigma \gamma_s \partial_\theta u_i(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s\end{aligned} \tag{97}$$

where $\mathcal{A}_\mathcal{X}$ and $\mathcal{A}_\theta$ are the infinitesimal generators of the processes $\mathcal{X}_t$ and $\tilde{\theta}_t$, respectively, $\hat{\Sigma}(x, \theta)$ denotes the $(D \times n_y)$-matrix consisting of the rows $n' + 1, n' + 2, \ldots, n' + n_y$ of the matrix $\Sigma(x, \theta)$, and $\partial_\theta^\top \partial_x u_k(s, x, \theta)_{ij} = \partial_{\theta_i} \partial_{x_j} u_k(s, x, \theta)$.

Using the Poisson equation and the previous identity, we obtain

$$\begin{aligned}\Gamma_{k,\eta} &= \int_{\tau_k}^{\sigma_k+\eta} \gamma_s \left(F\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) - g\left(\tilde{\theta}_s\right)\right)^\top ds \\ &= \int_{\tau_k}^{\sigma_k+\eta} \gamma_s G\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds = \int_{\tau_k}^{\sigma_k+\eta} \gamma_s \mathcal{A}_\mathcal{X} v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds \\ &= \int_{\tau_k}^{\sigma_k+\eta} \mathcal{A}_\mathcal{X} u\left(s, \tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds \\ &= \gamma_{\sigma_k+\eta} v(\tilde{\mathcal{X}}_{\sigma_k+\eta}, \tilde{\theta}_{\sigma_k+\eta}) - \gamma_{\tau_k} v(\tilde{\mathcal{X}}_{\tau_k}, \tilde{\theta}_{\tau_k}) \\ &\quad - \int_{\tau_k}^{\sigma_k+\eta} \dot{\gamma}_s v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds - \int_{\tau_k}^{\sigma_k+\eta} \gamma_s \mathcal{A}_\theta v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds \\ &\quad - \int_{\tau_k}^{\sigma_k+\eta} \gamma_s^2 \mathrm{tr}\Big[\hat{\Sigma}\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \\ &\quad \times \partial_\theta^\top \partial_x\Big] v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) ds \\ &\quad - \int_{\tau_k}^{\sigma_k+\eta} \gamma_s \partial_x v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \Sigma\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) d\mathcal{B}_s \\ &\quad - \int_{\tau_k}^{\sigma_k+\eta} \gamma_s^2 \partial_\theta v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s.\end{aligned} \tag{98}$$

Define

$$J_t^{(1)} \doteq \gamma_t \sup_{s\leq t} ||v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)||. \tag{99}$$

By using Condition 1, we have

$$\begin{aligned}\mathbb{E}\left[\left(J_t^{(1)}\right)^2\right] &= \mathbb{E}\left[\gamma_t^2 \sup_{s\leq t} ||v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)||^2\right] \\ &\leq K\gamma_t^2 \mathbb{E}\left[1 + \sup_{s\leq t} ||\tilde{\mathcal{X}}_s||^q\right] \\ &= K\gamma_t^2\left(1 + \mathbb{E}\left[\sup_{s\leq t} ||\tilde{\mathcal{X}}_s||^q\right]\right) \\ &\leq KK'\gamma_t^2(1 + \sqrt{t}) \\ &\leq K''\gamma_t^2\sqrt{t}\end{aligned} \tag{100}$$

where the first two inequalities use Conditions 1(iv) and (v), respectively. We choose an $r > 0$ such that $\gamma_t^2 t^{1/2+2r} \to 0$ for $t \to \infty$ (this is possible due to Condition 3), and we pick $T > 0$ large enough such that $\gamma_t^2 t^{1/2+2r} \leq 1$ for $t \geq T$. In addition, for each $0 < \delta < r$, we define the event $A_{t,\delta} \doteq \{J_t^{(1)} t^{r-\delta} \geq 1\}$. For $t \geq T$

$$\begin{aligned}\mathbb{P}(A_{t,\delta}) &\leq \mathbb{E}\left[J_t^{(1)} t^{r-\delta}\right] \leq \mathbb{E}\left[\left(J_t^{(1)}\right)^2\right] t^{2r-2\delta} \\ &\leq K''\gamma_t^2 t^{1/2+2r-2\delta} \leq K'' t^{-2\delta}\end{aligned} \tag{101}$$

where (100) was used in the second inequality.[6] We therefore have that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_{2^n,\delta}) < \infty. \tag{102}$$

By the Borel–Cantelli lemma, only finitely many events $A_{2^n,\delta}$ can occur. Therefore, there is a random index n_0 such that $J_{2^n}^{(1)} 2^{n(r-\delta)} \leq 1$ for all $n \geq n_0$. Alternatively, we can say that there is a finite positive random variable ξ and a deterministic $n_1 \in \mathbb{N}$ such that

$$J_{2^n}^{(1)} 2^{n(r-\delta)} \leq \xi, \quad n \geq n_1 \tag{103}$$

(e.g., choose $\xi = \max\{\max_{1\leq n' \leq n_0} J_{2^{n'}}^{(1)} 2^{n'(r-\delta)}, 1\}$). For $t \in [2^n, 2^{n+1}]$ and $n \geq n_1$, we therefore have

$$\begin{aligned} J_t^{(1)} &= \gamma_t \sup_{s\leq t} \left\|v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right\| \leq \gamma_{2^n} \sup_{s\leq t} \left\|v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right\| \\ &\leq \gamma_{2^n} \sup_{s\leq 2^{n+1}} \left\|v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right\| \\ &\leq K\gamma_{2^{n+1}} \sup_{s\leq 2^{n+1}} \left\|v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right\| \\ &= KJ_{2^{n+1}}^{(1)} \leq K\frac{\xi}{2^{(n+1)(r-\delta)}} \leq K\frac{\xi}{t^{r-\delta}} \end{aligned} \tag{104}$$

and as a consequence, $J_t^{(1)} \to 0$ a.s. as $t \to \infty$.

Next, define

$$\begin{aligned} J_t^{(2)} = \int_0^t &\left\|\dot{\gamma}_s v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) + \gamma_s \mathcal{A}_\theta v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right. \\ &\left. + \gamma_s^2 \mathrm{tr}\left[\hat{\Sigma}\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \partial_\theta^\top \partial_x\right] v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\right\| ds. \end{aligned} \tag{105}$$

Due to the PGP of H, $\hat{\Sigma}$, and v (see Conditions 1 and 2), we have

$$\begin{aligned} \sup_{t>0} \mathbb{E}\left[J_t^{(2)}\right] &\leq K\int_0^\infty \left(\dot{\gamma}_s + \gamma_s^2\right)\left(1 + \mathbb{E}[||\tilde{\mathcal{X}}_s||^q]\right) ds \\ &\leq KC \int_0^\infty \left(\dot{\gamma}_s + \gamma_s^2\right) ds < \infty. \end{aligned} \tag{106}$$

In the first inequality, we additionally used the fact that $\mathcal{A}_\theta$ contains at least a factor of γ_t, in the second one we relied on Condition 1(v), and in the third inequality we used Condition 3. Thus, $J_t^{(2)}$ converges to a finite random variable a.s.

Finally, we have the term

$$\begin{aligned} J_t^{(3)} &= \int_0^t \gamma_s \partial_x v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \Sigma\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) d\mathcal{B}_s \\ &+ \int_0^t \gamma_s^2 \partial_\theta v\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s. \end{aligned} \tag{107}$$

[6]The first inequality in (101) is elementary: For a nonnegative random variable Y with law p, we have

$$\mathbb{P}(Y \geq 1) = \int_1^\infty p(dy) \leq \int_1^\infty y p(dy) \leq \int_0^\infty y p(dy) = \mathbb{E}(Y).$$

By using Itô isometry and the same PGPs as in (106), we obtain

$$\begin{aligned} \sup_{t>0} \mathbb{E}\left[||J_t^{(3)}||^2\right] &= \int_0^\infty \gamma_s^2 \mathbb{E}\left[\|\partial_x v \Sigma\|^2\right] ds \\ &+ \int_0^\infty \gamma_s^4 \mathbb{E}\left[\left\|\partial_\theta v H^\top\right\|^2\right] ds \\ &+ 2\int_0^\infty \gamma_s^3 \mathrm{tr}\mathbb{E}\left[\partial_x v \hat{\Sigma} H^\top \partial_\theta^\top v^\top\right] ds \\ &\leq CK \int_0^\infty \left(\gamma_s^2 + \gamma_s^3 + \gamma_s^4\right)\left(1 + \mathbb{E}[||\tilde{\mathcal{X}}_s||^q]\right) ds \\ &\leq CKC' \int_0^\infty \left(\gamma_s^2 + \gamma_s^3 + \gamma_s^4\right) ds < \infty. \end{aligned} \tag{108}$$

Thus, by Doob's martingale convergence theorem, $J_t^{(3)}$ converges to a square integrable random variable a.s.

Finally, we note that

$$\begin{aligned} ||\Gamma_{k,\eta}|| &\leq J_{\sigma_k+\eta}^{(1)} + J_{\tau_k}^{(1)} + J_{\sigma_k+\eta}^{(2)} - J_{\tau_k}^{(2)} \\ &\quad + ||J_{\sigma_k+\eta}^{(3)} - J_{\tau_k}^{(3)}|| \to 0, \quad k \to \infty. \end{aligned} \tag{109}$$

■

Lemma 2: Let L be the Lipschitz constant of g. Choose $\lambda > 0$ such that for a given $\kappa > 0$ (this is the parameter of the stopping times τ_k) we have $3\lambda + \frac{\lambda}{4\kappa} = \frac{1}{2L}$. For k large enough and $\eta > 0$ small enough, $\int_{\tau_k}^{\sigma_k+\eta} \gamma_s ds > \lambda$. In addition, a.s., $\frac{\lambda}{2} \leq \int_{\tau_k}^{\sigma_k} \gamma_s ds \leq \lambda$.

Proof: This proof goes through exactly like the proof of [18, Lemma 3.2], with the only modification that the martingale in that proof takes the form

$$\int_0^t \gamma_s \frac{g\left(\tilde{\theta}_s\right)}{R_s} H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s.$$

■

Lemma 3: Suppose that $\tilde{\theta}_t \in \Theta$ for $t \geq 0$ and that there is an infinite number of intervals $[\tau_k, \sigma_k)$. There is a $\beta > 0$ such that for $k > k_0$

$$\tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k}) - \tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) \geq \beta \tag{110}$$

a.s.

Proof: By using Itô's lemma and the parameter update SDE (16), we obtain four terms

$$\begin{aligned} \tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k}) - \tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) &= \int_{\tau_k}^{\sigma_k} \gamma_s \left\|g\left(\tilde{\theta}_s\right)\right\|^2 ds \\ &+ \int_{\tau_k}^{\sigma_k} \gamma_s g\left(\tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s \\ &+ \int_{\tau_k}^{\sigma_k} \frac{\gamma_s^2}{2} \mathrm{tr}\left[H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) \mathcal{H}\left(\tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top\right] ds \\ &+ \int_{\tau_k}^{\sigma_k} \gamma_s g\left(\tilde{\theta}_s\right) \cdot \left[F\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) - g\left(\tilde{\theta}_s\right)^\top\right] ds \\ &= \Omega_{1,k} + \Omega_{2,k} + \Omega_{3,k} + \Omega_{4,k} \end{aligned} \tag{111}$$

where $\mathcal{H}$ is used to denote the Hessian of $\tilde{\mathcal{L}}$, see (29). By virtue of the definition of the stopping times and Lemmas 1 and 2

$$\begin{aligned}\Omega_{1,k} &= \int_{\tau_k}^{\sigma_k} \gamma_s \left\| g\left(\tilde{\theta}_s\right)\right\|^2 ds \\ &\geq \frac{\left\| g\left(\tilde{\theta}_{\tau_k}\right)\right\|^2}{4}\int_{\tau_k}^{\sigma_k} \gamma_s ds \geq \frac{\left\| g\left(\tilde{\theta}_{\tau_k}\right)\right\|^2}{8}\lambda(\kappa).\end{aligned} \tag{112}$$

We define

$$R_t = \begin{cases} \|g\left(\tilde{\theta}_{\tau_k}\right)\|, & t\in[\tau_k,\sigma_k) \text{ for some } k\geq 1,\\ \kappa, & \text{else}\end{cases} \tag{113}$$

such that we can write

$$\begin{aligned}\Omega_{2,k} &= \int_{\tau_k}^{\sigma_k} \gamma_s g\left(\tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s \\ &= \left\| g\left(\tilde{\theta}_{\tau_k}\right)\right\| \int_{\tau_k}^{\sigma_k} \gamma_s \frac{g\left(\tilde{\theta}_s\right)}{R_s} H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s.\end{aligned} \tag{114}$$

Since $\|g\left(\tilde{\theta}_s\right)\|/R_s \leq 2$, it follows from the Itô isometry and Conditions 1(v) and 2 that

$$\begin{aligned}&\sup_{t\geq 0}\mathbb{E}\left[\left(\int_0^t \gamma_s \frac{g\left(\tilde{\theta}_s\right)}{R_s} H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s\right)^2\right] \\ &\leq \sup_{t\geq 0}\int_0^t \mathbb{E}\left[\gamma_s^2 \frac{\|g\left(\tilde{\theta}_s\right)\|^2}{R_s^2}\|H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\|^2\right] ds \\ &\leq 4\int_0^\infty \gamma_s^2 \mathbb{E}\left[\left\|H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\right\|^2\right] ds \\ &\leq 4K\int_0^\infty \gamma_s^2\left(1+\mathbb{E}\left[\|\tilde{\mathcal{X}}_s\|^q\right]\right) ds < \infty.\end{aligned} \tag{115}$$

By Doob's martingale convergence theorem, the martingale $\mathfrak{M}_t = \int_0^t \gamma_s \frac{g(\tilde{\theta}_s)}{R_s} H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s$ converges to a finite random variable $\mathfrak{M}$ as $t\to\infty$. Thus, for any $\epsilon>0$, there is a k_0 such that a.s. we have $\Omega_{2,k}\leq \|g(\tilde{\theta}_{\tau_k})\|\epsilon$ for all $k\geq k_0$.

Next, we consider $\Omega_{3,k}$. Using Conditions 1 and 2 and Proposition 1, we obtain

$$\begin{aligned}&\sup_{t\geq 0}\mathbb{E}\left[\left|\int_0^t \frac{\gamma_s^2}{2}\mathrm{tr}\left[H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\mathcal{H}\left(\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top\right] ds\right|\right] \\ &\leq \sup_{t\geq 0}\mathbb{E}\left[\int_0^t \frac{\gamma_s^2}{2}\left|\mathrm{tr}\left[H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\mathcal{H}\left(\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top\right]\right| ds\right] \\ &\leq \int_0^\infty \frac{\gamma_s^2}{2}\mathbb{E}\left[\|H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\|^2\|\mathcal{H}\left(\tilde{\theta}_s\right)\|\right] ds \\ &\leq K\int_0^\infty \frac{\gamma_s^2}{2}\left(1+\mathbb{E}\left[\|\tilde{\mathcal{X}}_s\|^q\right]\right) ds<\infty\end{aligned} \tag{116}$$

from which it follows that

$$\int_0^t \frac{\gamma_s^2}{2}\mathrm{tr}\left[H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\mathcal{H}\left(\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top\right] ds$$

converges to a finite random variable as $t\to\infty$. Thus, a.s., $\Omega_{3,k}$ must converge to zero as $k\to\infty$.

Finally, we consider the term $\Omega_{4,k}$ and define the function $G(x,\theta)=g(\theta)\cdot\left[F(x,\theta)-g(\theta)^\top\right]$, which by definition of g satisfies $\int_{\mathbb{R}^D} G(x,\theta)\mu_\theta(dx)=0$. By Condition 1(iv), for each $\theta\in\Theta$, the Poisson equation $\mathcal{A}_\mathcal{X} v(x,\theta)=G(x,\theta)$ (where $\mathcal{A}_\mathcal{X}$ is the infinitesimal generator of the process $\mathcal{X}_t$) has a unique solution v with $\int_{\mathbb{R}^D} v(x,\theta)\mu_\theta(dx)=0$. Let $u(t,x,\theta)\doteq\gamma_t v(x,\theta)$ and apply Itô's lemma

$$\begin{aligned}&u(\sigma,\tilde{\mathcal{X}}_\sigma,\tilde{\theta}_\sigma)-u(\tau,\tilde{\mathcal{X}}_\tau,\tilde{\theta}_\tau)=\int_\tau^\sigma \partial_s u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)ds \\ &+\int_\tau^\sigma \mathcal{A}_\mathcal{X} u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)ds+\int_\tau^\sigma \mathcal{A}_\theta u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)ds \\ &+\int_\tau^\sigma \gamma_s\mathrm{tr}\left[\hat{\Sigma}\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\partial_x\partial_\theta^\top u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)\right] ds \\ &+\int_\tau^\sigma \partial_x u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)\Sigma\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)d\mathcal{B}_s \\ &+\int_\tau^\sigma \gamma_s\partial_\theta u(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s\end{aligned} \tag{117}$$

where $\hat{\Sigma}(x,\theta)$ denotes the $(D\times n_y)$-matrix consisting of the rows $n'+1,n'+2,\ldots,n'+n_y$ of the matrix $\Sigma(x,\theta)$, and $\partial_x\partial_\theta^\top u(s,x,\theta)_{ij}=\partial_{\theta_i}\partial_{x_j}u(s,x,\theta)$. Using the Poisson equation, we obtain

$$\begin{aligned}\Omega_{4,k} &= \int_{\tau_k}^{\sigma_k} \gamma_s g\left(\tilde{\theta}_s\right)\cdot\left[F\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)-g\left(\tilde{\theta}_s\right)^\top\right] ds \\ &= \int_{\tau_k}^{\sigma_k}\gamma_s G\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right) ds=\int_{\tau_k}^{\sigma_k}\gamma_s\mathcal{A}_\mathcal{X} v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right) ds \\ &= \int_{\tau_k}^{\sigma_k}\mathcal{A}_\mathcal{X} u\left(s,\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right) ds\end{aligned}$$

which, by using the previous identity, turns into

$$\begin{aligned}&=\gamma_{\sigma_k} v(\tilde{\mathcal{X}}_{\sigma_k},\tilde{\theta}_{\sigma_k})-\gamma_{\tau_k} v(\tilde{\mathcal{X}}_{\tau_k},\tilde{\theta}_{\tau_k}) \\ &-\int_{\tau_k}^{\sigma_k}\partial_s\gamma_s v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right) ds-\int_{\tau_k}^{\sigma_k}\gamma_s\mathcal{A}_\theta v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right) ds \\ &-\int_{\tau_k}^{\sigma_k}\gamma_s^2\mathrm{tr}\left[\hat{\Sigma}\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\partial_\theta^\top\partial_x v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\right] ds \\ &-\int_{\tau_k}^{\sigma_k}\gamma_s\partial_x v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)\Sigma\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)d\mathcal{B}_s \\ &-\int_{\tau_k}^{\sigma_k}\gamma_s^2\partial_\theta v\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s,\tilde{\theta}_s\right)^\top dV_s.\end{aligned} \tag{118}$$

By following the steps in the proof of Lemma 1, we find that $\Omega_{4,k}\to 0$ as $k\to\infty$ a.s.

For all $\epsilon > 0$ a.s., we have for k large enough

$$\tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k}) - \tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) = \Omega_{1,k} + \Omega_{2,k} + \Omega_{3,k} + \Omega_{4,k}$$
$$\geq \Omega_{1,k} - ||\Omega_{2,k}|| - ||\Omega_{3,k}|| - ||\Omega_{4,k}||$$
$$\geq \frac{1}{8}\lambda(\kappa)||g\left(\tilde{\theta}_{\tau_k}\right)||^2 - \epsilon||g\left(\tilde{\theta}_{\tau_k}\right)|| - 2\epsilon. \tag{119}$$

The lemma then follows by choosing $\epsilon = \min\{\frac{\lambda(\kappa)\kappa^2}{32}, \frac{\lambda(\kappa)}{32}\}$ and $\beta = \frac{\lambda(\kappa)\kappa^2}{32}$. ■

Lemma 4: Under the conditions of Lemma 3, there is a $0 < \beta_1 < \beta$ such that for $k > k_0$

$$\tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) - \tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_{k-1}}) \geq -\beta_1 \tag{120}$$

a.s.

Proof: As in Lemma 3, we obtain

$$\tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) - \tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_{k-1}}) \geq \int_{\sigma_{k-1}}^{\tau_k} \gamma_s g\left(\tilde{\theta}_s\right) H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s$$
$$+ \int_{\sigma_{k-1}}^{\tau_k} \frac{\gamma_s^2}{2} \mathrm{tr}\left[H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)\mathcal{H}\left(\tilde{\theta}_s\right)H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top\right] ds$$
$$+ \int_{\sigma_{k-1}}^{\tau_k} \gamma_s g\left(\tilde{\theta}_s\right) \cdot \left[F\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right) - g\left(\tilde{\theta}_s\right)^\top\right] ds. \tag{121}$$

It is sufficient to show that the RHS converges to zero a.s. Due to (113), the first term can be rewritten as

$$\kappa \int_{\sigma_{k-1}}^{\tau_k} \gamma_s \frac{g\left(\tilde{\theta}_s\right)}{R_s} H\left(\tilde{\mathcal{X}}_s, \tilde{\theta}_s\right)^\top dV_s. \tag{122}$$

Using the argument from the proof of Lemma 3, this converges to zero a.s. as $k \to \infty$. The treatment of the second and third terms is identical to the treatment of the terms $\Omega_{3,k}$ and $\Omega_{4,k}$ in the proof of Lemma 3. ■

Appendix C
Proof of Theorem 1

First, define the first exit time from Θ

$$\tau = \inf\left\{t \geq 0 : \tilde{\theta}_t \notin \Theta\right\}. \tag{123}$$

If $\tau < \infty$, since the paths of $\tilde{\theta}_t$ are continuous, we have $\tilde{\theta}_\tau \in \partial\Theta$. Furthermore, since $d\tilde{\theta}_t = 0$ on $\partial\Theta$, we have $\tilde{\theta}_t \in \partial\Theta$ for all $t \geq \tau$.

Next, consider the case when $\tau = \infty$, which implies that $\tilde{\theta}_t \in \Theta$ for all $t \geq 0$. Consider the case when there is a finite number of stopping times τ_k. Then, there is a finite T such that $||g(\tilde{\theta}_t)|| < \kappa$ for $t \geq T$. Therefore, since κ can be chosen arbitrarily small, $\lim_{t\to\infty} ||g(\tilde{\theta}_t)|| = 0$. Next, suppose that the number of stopping times τ_k is infinite. By Lemmas 3 and 4, there is a k_0 and constants $\beta > \beta_1 > 0$ such that for all $k \geq k_0$ a.s.

$$\tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k}) - \tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) \geq \beta \tag{124}$$

$$\tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) - \tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_{k-1}}) \geq -\beta_1 > -\beta. \tag{125}$$

Thus, we have

$$\tilde{\mathcal{L}}(\tilde{\theta}_{\tau_{n+1}}) - \tilde{\mathcal{L}}(\tilde{\theta}_{\tau_{k_0}})$$
$$= \sum_{k=k_0}^{n}\left[\tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k}) - \tilde{\mathcal{L}}\left(\tilde{\theta}_{\tau_k}\right) + \tilde{\mathcal{L}}(\tilde{\theta}_{\tau_{k+1}}) - \tilde{\mathcal{L}}(\tilde{\theta}_{\sigma_k})\right]$$
$$\geq (n + 1 - k_0)(\beta - \beta_1). \tag{126}$$

Since $\beta - \beta_1 > 0$, when $n \to \infty$, $\tilde{\mathcal{L}}(\tilde{\theta}_{\tau_{n+1}}) \to \infty$ a.s., and therefore $\tilde{\mathcal{L}}(\tilde{\theta}_t) \to \infty$ a.s. This is in contradiction to Proposition 1(iv), which states that $\tilde{\mathcal{L}}$ is bounded from above. Therefore, there are a.s. only a finite number of stopping times τ_k. ■

Acknowledgment

The authors would like to thank P. Harms, O. Zeitouni, K. Spiliopoulos, and S. Meyn for helpful discussions.

References

[1] A. Bain and D. Crişan, *Fundamentals of Stochastic Filtering* (Stochastic Modelling and Applied Probability), vol. 60. New York, NY, USA: Springer, 2009.
[2] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83, no. 1, pp. 95–108, 1961.
[3] V. E. Beneš, "Exact finite-dimensional filters for certain diffusions with nonlinear drift," *Stochastics*, vol. 5, no. 1/2, pp. 65–92, 1981.
[4] V. E. Beneš, "New exact nonlinear filters with large Lie algebras," *Syst. Control Lett.*, vol. 5, no. 4, pp. 217–221, 1985.
[5] F. E. Daum, "Exact finite-dimensional nonlinear filters," *IEEE Trans. Autom. Control*, vol. AC-31, no. 7, pp. 616–622, Jul. 1986.
[6] F. E. Daum, "Solution of the Zakai equation by separation of variables," *IEEE Trans. Autom. Control*, vol. AC-32, no. 10, pp. 941–943, Oct. 1987.
[7] C. D. Charalambous and R. J. Elliott, "New explicit filters and smoothers for diffusions with nonlinear drift and measurements," *Syst. Control Lett.*, vol. 33, no. 2, pp. 89–103, 1998.
[8] C. D. Charalambous and R. J. Elliott, "Classes of nonlinear partially observable stochastic optimal control problems with explicit optimal control laws," *SIAM J. Control Optim.*, vol. 36, no. 2, pp. 542–578, 1998.
[9] F. Gustafsson and G. Hendeby, "Some relations between extended and unscented Kalman filters," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 545–555, Feb. 2012.
[10] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adapt. Syst. Signal Process., Commun., Control Symp.*, 2000, pp. 153–158.
[11] D. Brigo, B. Hanzon, and F. LeGland, "A differential geometric approach to nonlinear filtering: The projection filter," *IEEE Trans. Autom. Control*, vol. 43, no. 2, pp. 247–252, Feb. 1998.
[12] D. Brigo, B. Hanzon, and F. LeGland, "Approximate nonlinear filtering by projection on exponential manifolds of densities," *Bernoulli*, vol. 5, no. 3, pp. 495–534, 1999.
[13] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*. London, U.K.: Oxford Univ. Press, Feb. 2011, pp. 656–704.
[14] G. Evensen, "The ensemble Kalman filter: Theoretical formulation and practical implementation," *Ocean Dyn.*, vol. 53, no. 4, pp. 343–367, 2003.
[15] D. Crişan and J. Xiong, "Approximate McKean-Vlasov representations for a class of SPDEs," *Stochastics*, vol. 82, no. 1–3, pp. 53–68, 2010.
[16] T. Yang, P. G. Mehta, and S. P. Meyn, "Feedback particle filter," *IEEE Trans. Autom. Control*, vol. 58, no. 10, pp. 2465–2480, Oct. 2013.
[17] F. LeGland and L. Mevel, "Recursive estimation in hidden Markov models," in *Proc. 36th IEEE Conf. Decis. Control*, 1997, pp. 3468–3473.
[18] J. Sirignano and K. Spiliopoulos, "Stochastic gradient descent in continuous time," *SIAM J. Financial Math.*, vol. 8, no. 1, pp. 933–961, 2017.
[19] A. Y. Veretennikov, "On polynomial mixing bounds for stochastic differential equations," *Stochastic Processes Appl.*, vol. 70, no. 1, pp. 115–127, 1997.

[20] E. Pardoux and A. Y. Veretennikov, "On the Poisson equation and diffusion approximation 1," *Ann. Probab.*, vol. 29, no. 3, pp. 1061–1085, 2001.
[21] E. Pardoux and A. Y. Veretennikov, "On the Poisson equation and diffusion approximation 2," *Annals Probab.*, vol. 31, no. 3, pp. 1166–1192, 2003.
[22] L. Hörmander, "Hypoelliptic second order differential equations," *Acta Mathematica*, vol. 119, pp. 147–171, 1967.
[23] J.-M. Bony, "Principe du maximum, inégalite de Harnack et unicité du problème de Cauchy pour les opérateurs elliptiques dégénérés," *Annales de l'Institut Fourier*, vol. 19, no. 1, pp. 277–304, 1969.
[24] K. Ichihara and H. Kunita, "A classification of the second order degenerate elliptic operators and its probabilistic characterization," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 30, no. 3, pp. 235–254, 1974.
[25] H. Kunita, "Asymptotic behavior of the nonlinear filtering errors of Markov processes," *J. Multivariate Anal.*, vol. 1, no. 4, pp. 365–393, 1971.
[26] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Ann. Probab.*, vol. 37, no. 5, pp. 1876–1925, 2009.
[27] H. Kunita, "Ergodic properties of nonlinear filtering processes," in *Spatial Stochastic Processes*. Boston, MA, USA: Birkhäuser, 1991, pp. 233–256.
[28] A. Budhiraja, "Asymptotic stability, ergodicity and other asymptotic properties of the nonlinear filter," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 39, no. 6, pp. 919–941, 2003.
[29] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[30] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models* (Springer Series in Statistics). New York, NY, USA: Springer, 2005.
[31] R. van Handel, *Hidden Markov Models* (Lecture Notes), Princeton, NJ, USA: Princeton Univ. Press, Jul. 2008.
[32] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "A survey of techniques for incremental learning of HMM parameters," *Inf. Sci.*, vol. 197, pp. 105–130, 2012.
[33] F. LeGland and L. Mevel, "Recursive identification of HMMs with observations in a finite set," in *Proc. 34th IEEE Conf. Decis. Control*, 1995, pp. 216–221.
[34] V. B. Tadic, "Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6406–6432, Dec. 2010.
[35] G. Poyiadjis, S. S. Singh, and A. Doucet, "Online parameter estimation for partially observed diffusions," in *Proc. IEEE Nonlinear Statist. Signal Process. Workshop*, 2006, pp. 197–200.
[36] G. Poyiadjis, A. Doucet, and S. S. Singh, "Particle approximations of the score and observed information matrix in state space models with application to parameter estimation," *Biometrika*, vol. 98, no. 1, pp. 65–80, 2011.
[37] I. B. Collings, V. Krishnamurthy, and J. B. Moore, "On-line identification of hidden Markov models via recursive prediction error techniques," *IEEE Trans. Signal Process.*, vol. 42, no. 12, pp. 3535–3539, Dec. 1994.
[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[39] A. Dembo and O. Zeitouni, "Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm," *Stochastic Processes Appl.*, vol. 23, no. 1, pp. 91–113, 1986.
[40] C. D. Charalambous and A. Logothetis, "Maximum likelihood parameter estimation from incomplete data via the sensitivity equations: The continuous-time case," *IEEE Trans. Autom. Control*, vol. 45, no. 5, pp. 928–934, May 2000.
[41] O. Zeitouni and B. Z. Bobrovsky, "On the joint nonlinear filtering-smoothing of diffusion processes," *Syst. Control Lett.*, vol. 7, no. 4, pp. 317–321, 1986.
[42] R. J. Elliott and V. Krishnamurthy, "Exact finite-dimensional filters for maximum likelihood parameter estimation of continuous-time linear Gaussian systems," *SIAM J. Control Optim.*, vol. 35, no. 6, pp. 1908–1923, 1997.
[43] C. D. Charalambous, R. J. Elliott, and V. Krishnamurthy, "Conditional moment generating functions for integrals and stochastic integrals," *SIAM J. Control Optim.*, vol. 42, no. 5, pp. 1578–1603, 2003.
[44] N. Gupta and R. Mehra, "Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 774–783, Dec. 1974.
[45] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557–2573, Aug. 1993.
[46] G. Mongillo and S. Deneve, "Online learning with hidden Markov models," *Neural Comput.*, vol. 20, no. 7, pp. 1706–1716, 2008.
[47] O. Cappé, "Online sequential Monte Carlo EM algorithm," in *Proc. IEEE/SP 15th Workshop Statist. Signal Process.*, 2009, pp. 37–40.
[48] O. Cappé, "Online EM algorithm for hidden Markov models," *J. Comput. Graph. Statist.*, vol. 20, no. 3, pp. 728–749, 2011.
[49] O. Cappé, "Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation," *Monte Carlo Methods Appl.*, vol. 7, no. 1/2, pp. 81–92, 2001.
[50] P. Del Moral, A. Doucet, and S. Singh, "Forward smoothing using sequential Monte Carlo," 2010, arXiv:1012.5390.
[51] P. Rebeschini and R. van Handel, "Can local particle filters beat the curse of dimensionality?" *Ann. Appl. Probab.*, vol. 25, no. 5, pp. 2809–2866, 2015.
[52] A. Papavasiliou, "Parameter estimation and asymptotic stability in stochastic filtering," *Stochastic Processes Appl.*, vol. 116, no. 7, pp. 1048–1065, 2006.
[53] J. M. F. Moura and S. K. Mitter, "Identification and filtering–optimal recursive maximum likelihood approach," Lab. Inf. Decis. Syst., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. LIDS-P, 1986.
[54] J. Sirignano and K. Spiliopoulos, "Stochastic gradient descent in continuous time: A central limit theorem," 2017, arXiv:1710.04273.

Simone Carlo Surace received the M.Sc. degree in theoretical physics and Ph.D. degree in neuroscience from the University of Bern, Bern, Switzerland, in 2011 and 2016, respectively.

He is currently a Postdoctoral Researcher with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zürich, Switzerland. His research interests include nonlinear filtering, parameter learning, information theory, information geometry, and their application to biology.

Jean-Pascal Pfister received the M.Sc. degree in physics and Ph.D. degree in computational neuroscience from the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2002 and 2006, respectively.

He is currently an Assistant Professor with the Institute of Neuroinformatics (INI), University of Zurich and ETH Zurich, Zürich, Switzerland. Prior joining the INI, he was a Postdoctoral Researcher with the University of Cambridge, Cambridge, U.K. and the University of Bern, Bern, Switzerland. His current research focuses on the interface between statistical learning and biological learning, including nonlinear filtering, synaptic plasticity, and spiking neuron models.