LOG-NORMALIZATION CONSTANT ESTIMATION USING THE ENSEMBLE KALMAN–BUCY FILTER WITH APPLICATION TO HIGH-DIMENSIONAL MODELS

DAN CRISAN,* Imperial College London
PIERRE DEL MORAL,** Institut de Mathématiques de Bordeaux
AJAY JASRA*** **** AND
HAMZA RUZAYQAT *****, King Abdullah University of Science and Technology

Abstract

In this article we consider the estimation of the log-normalization constant associated to a class of continuous-time filtering models. In particular, we consider ensemble Kalman–Bucy filter estimates based upon several nonlinear Kalman–Bucy diffusions. Using new conditional bias results for the mean of the aforementioned methods, we analyze the empirical log-scale normalization constants in terms of their $L_n$-errors and $L_n$-conditional bias. Depending on the type of nonlinear Kalman–Bucy diffusion, we show that these are bounded above by terms such as $C(n)\left[t^{1/2}/N^{1/2} + t/N\right]$ or $C(n)/N^{1/2}$ ($L_n$-errors) and $C(n)\left[t + t^{1/2}\right]/N$ or $C(n)/N$ ($L_n$-conditional bias), where $t$ is the time horizon, $N$ is the ensemble size, and $C(n)$ is a constant that depends only on $n$, not on $N$ or $t$. Finally, we use these results for online static parameter estimation for the above filtering models and implement the methodology for both linear and nonlinear models.

Keywords: Kalman–Bucy filter; Riccati equations; normalizing constant; nonlinear Markov processes; parameter estimation

2020 Mathematics Subject Classification: Primary 60G35; 62M20
Secondary 65C40; 65C20; 93E10

1. Introduction

The filtering problem concerns the recursive estimation of a partially observed Markov process conditioned on a path of observations. It is found in a wide class of real applications, including finance, applied mathematics, and engineering; see e.g. [2, 8, 20]. This article focuses upon the computation of the normalizing constant for certain classes of filtering problems, to be introduced below. These normalizing constants can be of interest in statistics and engineering for model selection and/or parameter estimation.

In this article we study linear and Gaussian models in continuous time. It is well known that, for such models, under some minimal assumptions, the filter is Gaussian, with mean satisfying the Kalman–Bucy (KB) equations and the covariance matrix obeying a Riccati equation. In
many cases of practical interest, these latter equations may not be computable, for instance if one does not have access to an entire trajectory of data. Instead one can use some suitable approximations obtained via time-discretization methods. However, even then, if the dimension $r_1$ of the state vector is very large, the numerical approximation of the KB and Riccati equations can become computationally prohibitive, often requiring a computational effort of at least $O(r_1^2)$ per update. The case where $r_1$ is large occurs in many applications, including ocean and atmosphere science (e.g. [1]) and oil reservoir simulations (e.g. [18]). A rather elegant and successful solution to this problem, in discrete time, was developed in [11] in the form of the ensemble Kalman filter (EnKF), which can reduce the cost to $O(r_1)$.

To develop an analogue of the EnKF in continuous time, the starting point is often a class of conditional McKean–Vlasov diffusions (the KB diffusion). These latter diffusions share the same law as the filter associated to the linear and Gaussian model in continuous time. Hence a possible alternative to recursively solving the KB and Riccati equations is to generate $N \in \mathbb{N}$ independent copies from the KB diffusion and use a simple Monte Carlo estimate for expectations with respect to the filter. However, the diffusion process cannot be simulated exactly, but can be approximated in a principled way by allowing the $N$ samples to interact; precise details are given in Section 2. The resulting method, called the ensemble Kalman–Bucy filter (EnKBF), is by now rather well understood, with several contributions on its convergence (as $N \to \infty$) analysis; see for instance [5, 3, 9, 14, 21].

In this work we focus upon using several versions of EnKBF for online estimation of the normalization constant. In particular the contributions of this work are as follows:

1. New results on the conditional bias of the mean using the EnKBF.
2. A derivation of an estimate of the normalization constant using the EnKBF.
3. A proof that the $L_n$-error of the estimate on the log-scale is bounded above by $C(n)[t^{1/2}/N^{1/2} + t/N]$ or $C(n)/N^{1/2}$, depending on the nonlinear KB diffusion, with $t$ the time parameter and $C(n)$ a constant that depends only on $n$, not on $N$ or $t$.
4. A proof that the $L_n$-conditional bias of the estimate on the log-scale is bounded above by $C(n)[t + t^{1/2}]/N$ or $C(n)/N$, depending on the nonlinear KB diffusion.
5. A development of a method that uses the normalization constant estimate to perform online static parameter estimation.

The result in item 1 is of independent interest, but is used directly in item 4. To the best of our knowledge the estimate in item 2 is new and can be computed with a little extra computational cost over applying an EnKBF (under time discretization). Whilst several authors have used the EnKF for normalization constant estimation, e.g. [10], we have not seen the EnKBF version investigated in the literature.

In addition to the contributions 3 and 4, the results establish the decay of the mean squared error (MSE), for instance, of the log-normalizing constant estimate as the time parameter increases. This rate is expected to occur in practice, as we will see in simulations, and parallels other results found in the literature for particle estimates of the normalization constant (e.g. [7]). In relation to item 5, if one assumes that the model of interest has several unknown and time-homogeneous parameters, $\theta$ say, then one is interested in estimating such parameters, for instance using likelihood-based methods. We show how our estimate can be leveraged for this latter task. In this paper we use the simultaneous perturbation stochastic approximation (SPSA) method [19], which is popular in engineering applications. SPSA is based upon a finite difference estimator of the gradient, with respect to $\theta$, of the log-normalizing constant. It
constructs a stochastic approximation scheme for the estimation of static parameters. We are not aware of this method being used for the EnKBF. As it is a zero-order optimization method, we expect it to be computationally less expensive than resorting to using other estimates of the gradient (of the log-normalizing constant). It should be noted that our work focuses upon the standard or ‘vanilla’ EnKBF, the deterministic EnKBF [17] and deterministic-transport-type EnKBFs [16]; other extensions are possible, for instance, to the feedback particle filter [22].

This article is structured as follows. In Section 2 we provide details on the class of filtering problems that we address, as well as a review on the enKBF. The conditional bias of the mean associated to various EnKBFs is also presented there. In Section 3 we discuss how to compute the normalizing constant estimate, as well as its $L_n$-error and conditional bias. The latter is supported by numerical results checking that the order of the convergence rate indeed holds even under Euler discretization. In Section 4 we show how the normalizing constant estimate can be used in online static parameter estimation problems.

2. Description of the model and algorithms

2.1. The Kalman–Bucy filter

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space together with a filtration $(\mathcal{G}_t)_{t \geq 0}$ which satisfies the usual conditions. On $(\Omega, \mathcal{F}, \mathbb{P})$ we consider two $\mathcal{G}_t$-adapted processes $X = \{X_t, \ t \geq 0\}$ and $Y = \{Y_t, \ t \geq 0\}$ that form a time-homogeneous linear-Gaussian filtering model of the following form:

$$
\begin{align*}
\dot{X}_t &= AX_t dt + R_1^{1/2} dW_t, \\
\dot{Y}_t &= CX_t dt + R_2^{1/2} dV_t,
\end{align*}
$$

(1)

$Y_0 = 0$. In the above display, we have the following:

- $(W_t, V_t)$ is a $\mathcal{G}_t$-adapted $(r_1 + r_2)$-dimensional Brownian motion.
- Independent of $(W_t, V_t)$, $X_0$ is a $\mathcal{G}_0$-measurable $r_1$-valued Gaussian random vector with mean and covariance matrix $(\mathbb{E}(X_0), P_0)$.
- The symmetric matrices $R_1^{1/2}$ and $R_2^{1/2}$ are invertible.
- $A$ is a square $r_1 \times r_1$ matrix; $C$ is an $r_2 \times r_1$ matrix.

We let $\mathcal{F}_t = \sigma(Y_s, s \leq t)$ be the filtration generated by the observation process. It is well known that the conditional distribution $\eta_t$ of the signal state $X_t$ given $\mathcal{F}_t$ is an $r_1$-dimensional Gaussian distribution with mean and covariance matrix

$$
\hat{X}_t := \mathbb{E}(X_t | \mathcal{F}_t) \quad \text{and} \quad P_t := \mathbb{E}\left((X_t - \mathbb{E}(X_t | \mathcal{F}_t))(X_t - \mathbb{E}(X_t | \mathcal{F}_t))'\right)
$$

given by the KB and Riccati equations

$$
\begin{align*}
\dot{\hat{X}}_t &= A \hat{X}_t dt + P_t C' R_2^{-1} (dY_t - C \hat{X}_t dt), \\
\partial_t P_t &= \text{Ricc}(P_t),
\end{align*}
$$

(2)

The Riccati drift function in (3), from $S_{r_1}^+$ into $S_{r_1}$, is given by

$$
\text{Ricc}(Q) = AQ + QA' - QSQ + R \quad \text{with} \quad R = R_1 \quad \text{and} \quad S := C' R_2^{-1} C,
$$

(4)

where $S_{r_1}^+$ (resp. $S_{r_1}$) is the collection of symmetric and positive definite (resp. semi-definite) $r_1 \times r_1$ matrices defined for any $Q \in S_{r_1}^+$. 

Log-normalization constant estimation using EnKBF

1141
2.2. Nonlinear Kalman–Bucy diffusions

We now consider three conditional nonlinear McKean–Vlasov-type diffusion processes

\[ d\bar{X}_t = A\bar{X}_t dt + R_1^{1/2} d\bar{W}_t + \mathcal{P}_{\eta_t} C' R_2^{-1} \left[ dY_t - \left( C\bar{X}_t dt + R_2^{1/2} d\bar{V}_t \right) \right], \quad (5) \]

\[ d\bar{X}_t = A\bar{X}_t dt + R_1^{1/2} d\bar{W}_t + \mathcal{P}_{\eta_t} C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C[\bar{X}_t + \eta_t] dt \right) \right], \quad (6) \]

\[ d\bar{X}_t = A\bar{X}_t dt + R_1 P^{-1}_t (\bar{X}_t - \eta_t) dt + \mathcal{P}_{\eta_t} C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C[\bar{X}_t + \eta_t] dt \right) \right], \quad (7) \]

where \((\bar{W}_t, \bar{V}_t, X_0)\) are independent copies of \((W_t, V_t, X_0)\) (thus independent of the signal and the observation path). In the above formulas, \(\mathcal{P}_{\eta_t}\) stands for the covariance matrix

\[ \mathcal{P}_{\eta_t} = \eta_t[(e - \eta_t(e))(e - \eta_t(e))^\top] \quad \text{with} \quad \eta_t := \text{Law}(\bar{X}_t | F_t) \quad \text{and} \quad e(x) := x, \quad (8) \]

and we use the generic notation \(\eta_t(f) = \int_{\mathbb{R}^2} f(x) \eta_t(dx)\) for any \(f : \mathbb{R}^d \to \mathbb{R}\) that is \(\eta_t\)-integrable (in particular, \(r = r_1\) in \((6)\) and \((7)\) and \(r = r_2^2\) in \((8)\)). Any of these probabilistic models will be commonly referred to as KB (nonlinear) diffusion processes. In the following we will denote by \(G_t\) the augmented filtration generated by \(X_0\) and the triplet \((Y_t, \bar{W}_t, \bar{V}_t)\). The process \((5)\) corresponds to the vanilla-type EnKBF that is typically used in the literature. The process \((6)\) is associated with the so-called deterministic EnKBF \([17]\), and \((7)\) is a deterministic-transport-inspired equation \([16]\).

We have the following result, which is considered, for instance, in \([9]\).

**Lemma 2.1.** Let \(\bar{X}_t\) be a process such that \(\mathbb{E}[|X_0|^2] < \infty\), and that satisfies any one of \((5)\)–\((7)\). Then the conditional mean and the conditional covariance matrix (given \(F_t\)) of any of the nonlinear KB diffusions \((5)\)–\((7)\) satisfy Equations \((2)\) and \((3)\), respectively.

Lemma 2.1 enables us to approximate the mean and covariance associated to the linear filtering problem in \((1)\) by simulating \(N\) independent and identically distributed (i.i.d.) samples from one of the processes \((5)\)–\((7)\) exactly and then use the sample mean and sample covariance to approximate the mean and covariance of the filter. Since this exact simulation is seldom possible, we consider how one can generate \(N\)-particle systems whose sample mean and sample covariance can be used to replace the exact simulation.

**Remark 2.1.** Let us observe that Equations \((5)\)–\((7)\) indeed have a unique solution in the class of processes \(Z_t\) such that \(\mathbb{E}[\sup_{t \in [0, T]} |Z_t|^2] < \infty\). To show existence, one considers the (linear version of) the equations \((5)\)–\((7)\) with the conditional expectations \(\eta_t(e)\) and \(\eta_t[(e - \eta_t(e))(e - \eta_t(e))^\top]\) replaced by the solution of the equations \((2)\) and \((3)\), respectively, that is,

\[ d\hat{X}_t = A\hat{X}_t dt + R_1^{1/2} d\bar{W}_t + P_t C' R_2^{-1} \left[ dY_t - \left( C\hat{X}_t dt + R_2^{1/2} d\bar{V}_t \right) \right], \quad (9) \]

\[ d\tilde{X}_t = A\tilde{X}_t dt + R_1^{1/2} d\bar{W}_t + P_t C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C[\tilde{X}_t + \hat{X}_t] dt \right) \right], \quad (10) \]

\[ d\check{X}_t = A\check{X}_t dt + R_1 P^{-1}_t (\check{X}_t - \hat{X}_t) dt + P_t C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C[\check{X}_t + \hat{X}_t] dt \right) \right]. \quad (11) \]

Equations \((9)\)–\((11)\) have a unique solution (as they are linear) that indeed satisfies the corresponding (nonlinear) equations \((5)\)–\((7)\). To show the uniqueness of the solutions of
Equations (5)–(7), one observes first that any solution \( \eta_t \) of the (nonlinear) equations (5)–(7) has its conditional expectations \( \eta_t(e) \) and \( \eta_t'(e-\eta_t(e))' \) uniquely characterized by the equations (2) and (3). Therefore they satisfy the corresponding linear versions of the equations (5)–(7), with the conditional expectations \( \eta_t(e) \) and \( \eta_t'(e-\eta_t(e))' \) replaced by the solution of the equations (2) and (3). In other words, they satisfy Equations (9)–(11) and therefore they are unique as the corresponding linear equations (9)–(11) have a unique solution.

**Remark 2.2.** If one modifies (1) to

\[
\begin{align*}
    dX_t &= f(X_t) \, dt + R_1^{1/2} \, dW_t, \\
    dY_t &= C X_t \, dt + R_2^{1/2} \, dV_t,
\end{align*}
\]

for some nonlinear function \( f : \mathbb{R}^{r_1} \to \mathbb{R}^{r_1} \), one can consider a modification of any of (5)–(7). For instance in the case (5)

\[
dX_t = f(X_t) \, dt + R_1^{1/2} \, dW_t + \mathcal{P}_{\eta_t} C' R_2^{-1} \left[ dY_t - \left( C X_t \, dt + R_2^{1/2} \, dV_t \right) \right].
\]

We note, however, that any approximation of this process, as alluded to above, does not typically provide an approximation of the nonlinear filter. Nonetheless it is considered in many works in the field.

**2.3. Ensemble Kalman–Bucy filters**

We now consider three ensemble Kalman–Bucy filters that correspond to the mean-field particle interpretation of the nonlinear diffusion processes (5)–(7). To be more precise we let \( (\bar{W}_t, \bar{V}_t, \xi_{i}^{0})_{1 \leq i \leq N} \) be \( N \) independent copies of \( (\bar{W}_t, \bar{V}_t, \bar{X}_0) \). In this notation, we have the three McKean–Vlasov-type interacting diffusion processes for \( i \in \{1, \ldots, N\} \),

\[
\begin{align*}
    d\xi_{i}^{t} &= A \xi_{i}^{t} \, dt + R_1^{1/2} \, d\bar{W}_t + p_t C' R_2^{-1} \left[ dY_t - \left( C \xi_{i}^{t} \, dt + R_2^{1/2} \, d\bar{V}_i \right) \right], \quad (12) \\
    d\xi_{i}^{t} &= A \xi_{i}^{t} \, dt + R_1^{1/2} \, d\bar{W}_t + p_t C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C [\xi_{i}^{t} + \eta_{i}^{N}(e)] \right) \right], \quad (13) \\
    d\xi_{i}^{t} &= A \xi_{i}^{t} \, dt + R_{1}(p_t)^{-1} \left( \xi_{i}^{t} - \eta_{i}^{N}(e) \right) \, dt + p_t C' R_2^{-1} \left[ dY_t - \left( \frac{1}{2} C [\xi_{i}^{t} + \eta_{i}^{N}(e)] \right) \right], \quad (14)
\end{align*}
\]

with the rescaled particle covariance matrices

\[
p_t := \left( 1 - \frac{1}{N} \right)^{-1} \mathcal{P}_{\eta_t} = \frac{1}{N-1} \sum_{1 \leq i \leq N} \left( \xi_{i}^{t} - m_t \right) \left( \xi_{i}^{t} - m_t \right)^{'}
\]

and the empirical measures

\[
\eta_{i}^{N} := \frac{1}{N} \sum_{1 \leq i \leq N} \delta_{\xi_{i}^{t}} \quad \text{and the sample mean} \quad m_t := \frac{1}{N} \sum_{1 \leq i \leq N} \xi_{i}^{t}.
\]

Note that for (14) one needs \( N \geq r_1 \) for the almost sure invertibility of \( p_t \), but it is also sufficient to use the pseudo-inverse instead of the inverse.
These processes have been thoroughly studied in the literature, and a review can be found in [5]. We have, as shown in [9, Theorem 3.1], the evolution equations for (12) and (13),

\[ dm_t = A m_t dt + p_t C'R_2^{-1} (dY_t - Cm_t dt) + \frac{1}{\sqrt{N} + 1} dM_t, \]

\[ dp_t = \text{Ricc}(p_t) dt + \frac{1}{\sqrt{N}} dM_t, \]

with pairwise orthogonal martingales \( M_t \) and \( \mathcal{M}_t \) described in [9, Theorem 3.1]. Note that for any \( t \geq 0 \), the absolute moments of \( \mathcal{M}_t \) and \( M_t \) are uniformly, with respect to \( N \), bounded above. We remark that the mathematical theory for (14) in the linear-Gaussian setting reverts to that of the standard KB filter as detailed in [3].

**Remark 2.3.** Returning to the context of Remark 2.2, one could, for instance, run the following version of (12) for \( i \in \{1, \ldots, N\} \):

\[ d\xi^i_t = f(\xi^i_t) dt + R_1^{1/2} d\tilde{W}_t^i + p^N_t C'R_2^{-1} \left[ dY_t - \left( C\xi^i_t dt + R_2^{1/2} d\tilde{V}_t^i \right) \right]. \]

Again, one should note that this will typically not produce a consistent approximation of the nonlinear filter, as would be the case for (12) with the model (1). A substantial discussion of the application and theory of the EnKBF in the nonlinear setting can be found in [5, Section 9].

Below we use \( \| \cdot \| \) to denote the \( L_2 \)-norm for vectors. For a square matrix \( B \), say, \( B_{\text{sym}} = \frac{1}{2}(B + B') \) and \( \mu(B) \) is the largest eigenvalue of \( B_{\text{sym}} \). In the cases of (12) and (13), the authors of [9] consider the following assumption (recall (4)): we have \( \mu(S) > 0 \) with

\[ S = \mu(S)\text{Id}, \tag{16} \]

where \( \text{Id} \) is the \( r_1 \times r_1 \) identity matrix (in general we use \( \text{Id} \) to denote the identity matrix of ‘appropriate dimension’, appropriate depending on the context). The paper [9] proves the following time-uniform convergence theorem for the mean.

**Theorem 2.1.** Consider the cases of (12) and (13). Assume that (16) holds and that \( \mu(A) < 0 \). Then for any \( n \geq 1 \) and \( N \) sufficiently large, we have that

\[ \sup_{i \geq 0} \mathbb{E}\left[ \| m_t - \hat{X}_t^i \|^n \right]^{\frac{1}{n}} \leq \frac{\mathcal{C}(n)}{\sqrt{N}}, \]

where \( \mathcal{C}(n) < \infty \) is a constant that does not depend upon \( N \).

Below we assume that \( \left( A, R_1^{1/2} \right) \) and \( (A, C) \) are controllable and observable pairs; see for instance [5, Equation (2.7)]. Under this assumption, \( P_\infty \), the solution to \( \text{Ricc}(P_\infty) = 0 \), exists and is unique. Set \( \mu(A - P_\infty S) \) as the largest eigenvalue of \( \frac{1}{2}(A - P_\infty S + (A - P_\infty S)^\top) \). For the case of (13), the following result, presented in [5], can be shown.

**Theorem 2.2.** Consider the case of (13). Assume that \( \left( A, R_1^{1/2} \right) \) and \( (A, C) \) are controllable and observable pairs and that \( \mu(A - P_\infty S) < 0 \) and \( S \in \mathbb{S}_{r_1}^+ \). Then for any \( n \geq 1 \), \( N \geq 2 \) there exists a \( t(n, N) > 0 \) with \( t(n, N) \to \infty \) as \( N \to \infty \), such that for any \( t \in [0, t(n, N)] \)

\[ \mathbb{E}\left[ \| m_t - \hat{X}_t^i \|^n \right]^{\frac{1}{n}} \leq \frac{\mathcal{C}(n)}{\sqrt{N}}, \]

where \( \mathcal{C}(n) < \infty \) is a constant that does not depend upon \( N \). The quantity \( t(n, N) \) is characterized in [4].
We note that $t(n, N)$ is $O(\log(N))$; see [4, Theorem 2.1] with $\epsilon = 1/\sqrt{N}$ ($\epsilon$ is as in [4, Theorem 2.1]).

For the case of (14) we have the following result. Also note that we use the inverse of $p_t$ in (14), but one could use the pseudo-inverse instead, and then the constraint on $N$ in the below statement would not be required.

**Theorem 2.3.** Consider the case of (14). Assume that $\left(A, R_1^{1/2}\right)$ and $(A, C)$ are controllable and observable pairs. Then for any $n \geq 1$ there exist $C(n) < \infty$, $\kappa \in (0, 1)$ such that for any $N \geq r_1$ and $t \geq 0$ we have that

$$\mathbb{E}\left[\|m_t - \hat{X}_t\|^n\right]^{1/n} \leq \frac{C(n)\kappa^t}{\sqrt{N}},$$

where $C(n) < \infty$ is a constant that does not depend upon $N$.

**2.4. Conditional bias result**

We set

$$\hat{m}_t := \mathbb{E}(m_t \mid \mathcal{F}_t), \quad \hat{p}_t := \mathbb{E}(p_t \mid \mathcal{F}_t).$$

We consider bounds on

$$\mathbb{E}\left(\|\hat{m}_t - \hat{X}_t\|^n\right)^{1/n}.$$

These bounds do not currently exist in the literature and will allow us to investigate our new estimator for the normalization constant, later on in the article. In the case (14) we have (see e.g. [3])

$$\hat{m}_t = \hat{X}_t \quad \text{and} \quad \hat{p}_t = P_t,$$

so we focus only on (12) and (13).

Using the second-order Taylor-type expansions presented in [6] when the pair $\left(A, R_1^{1/2}\right)$ is stabilizable and $(A, S_1^{1/2})$ is detectable, we have the uniform almost-sure bias and the $\mathbb{L}_n$-mean error estimates ($n \geq 1$)

$$0 \leq P_t - \hat{p}_t \leq C_1/N \quad \text{and} \quad \mathbb{E}\left(\|P_t - p_t\|^n \mid \mathcal{F}_t\right)^{1/n} \leq C_2(n)/\sqrt{N}$$

as soon as $N$ is chosen sufficiently large, for some deterministic constants $C_1, C_2(n)$ that do not depend on the time horizon or $N$ (but $C_2(n)$ does depend upon $n$). Whenever $\mu(A) < 0$, Theorem 2.1 yields the rather crude estimate

$$\mathbb{E}\left(\|\hat{m}_t - \hat{X}_t\|^n\right)^{1/n} = \mathbb{E}\left(\|m_t - \hat{X}_t \mid \mathcal{F}_t\|^n\right)^{1/n} \leq \mathbb{E}\left(\|m_t - \hat{X}_t\|^n\right)^{1/n} \leq \frac{C(n)}{\sqrt{N}}.$$

Our objective is to improve the estimate (19). We observe that

$$d(m_t - \hat{X}_t) = A(m_t - \hat{X}_t)dt + (p_t - P_t) C'R_2^{-1} (dY_t - Cm_t dt) + P_tC'S^{-1} (dY_t - Cm_t dt) - P_tC'R_2^{-1} (dY_t - C\hat{X}_t dt) + \frac{1}{\sqrt{N} + 1} d\tilde{M}_t,$$

which yields

$$d(\hat{m}_t - \hat{X}_t) = (A - P_tS) (\hat{m}_t - \hat{X}_t)dt + (\hat{p}_t - P_t) C'R_2^{-1} (dY_t - C\hat{X}_t dt) + \mathbb{E}\left((P_t - p_t) S(m_t - \hat{X}_t \mid \mathcal{F}_t) dt\right).$$
Equation (21) is deduced by conditioning both terms in (20) with respect to $\mathcal{F}_t$, exchanging the (stochastic) integration with the conditional expectation for all the terms on the right-hand side of (20). To justify this Fubini-like exchange one can use, for example, Lemma 3.21 in [2]. One also needs to use the fact that, for arbitrary $0 \leq s \leq t$, $\hat{m}_s = \mathbb{E}(m_s \mid \mathcal{F}_s) = \mathbb{E}(m_s \mid \mathcal{F}_t)$ and $\hat{\eta}_s := \mathbb{E}(\eta_s \mid \mathcal{F}_s) = \mathbb{E}(\eta_s \mid \mathcal{F}_t)$. To justify this, one can use, for example, Proposition 3.15 in [2]. We note that both Proposition 3.15 and Lemma 3.21 hold true when the conditioning is side of (20). To justify this Fubini-like exchange one can use, for example, Lemma 3.21 in $\mathbb{P}$, they remain independent of $Y$ under $\hat{\mathbb{P}}$, and therefore conditioning the processes $\tilde{\xi}_i$ under $\hat{\mathbb{P}}$ is the same as conditioning them under $\mathbb{P}$. Note that this argument does not apply to the original signal $X$.

Let $\mathcal{E}_{s,t}$ be the exponential semigroup (or the state transition matrix) associated with the smooth flow of matrices $t \mapsto (A - P_t)S$ defined for any $s \leq t$ by the forward and backward differential equations
\[
\partial_t \mathcal{E}_{s,t} = (A - P_t)S \mathcal{E}_{s,t} \quad \text{and} \quad \partial_s \mathcal{E}_{s,t} = -\mathcal{E}_{s,t}(A - P_s)S,
\]
with $\mathcal{E}_{s,s} = \text{Id}$. Equivalently, in terms of the matrices $E_t := \mathcal{E}_{0,t}$, we have $\mathcal{E}_{s,t} = E_tE_s^{-1}$. Under some observability and controllability conditions, we have that the drift matrices $(A - P_s)$ deliver a uniformly stable time-varying linear system in the sense that
\[
||\mathcal{E}_{s,t}|| \leq c e^{-\lambda(t-s)} \quad \text{for some} \quad \lambda > 0. \tag{22}
\]
See for instance Equation (2.13) in the review article [5]. In this notation, recalling that $\hat{m}_0 = \hat{X}_0$, we have the bias formulae
\[
\hat{m}_t - \hat{X}_t = \int_0^t \mathcal{E}_{s,t} \left( \tilde{\eta}_s - P_s \right) C R_2^{-1} (dY_s - C \hat{X}_s \, dt)
\]
\[+ \int_0^t \mathcal{E}_{s,t} \mathbb{E}((P_s - P_s) S (m_s - \hat{X}_s) \mid \mathcal{F}_s) \, ds.
\]
Note that the claim that $(\tilde{\eta}_s - P_s) \in [-c_1/N, 0]$ almost everywhere can be justified via (18). Combining the Burkholder–Davis–Gundy inequality with the almost sure and uniform bias estimate (18) and the exponential decay (22), we obtain the uniform estimate
\[
\mathbb{E} \left( \left\| \int_0^t \mathcal{E}_{s,t} \left( \tilde{\eta}_s - P_s \right) C R_2^{-1} (dY_s - C \hat{X}_s \, dt) \right\|^n \right)^{1/n} \leq \frac{C(n)}{N}
\]
for some deterministic constant $C(n)$ that does not depend upon $t$ or $N$. Conversely, combining Hölder’s inequality with the uniform variance estimate (18), we have the inequality
\[
\mathbb{E} \left( \left\| \mathbb{E}((P_s - P_s) S (m_s - \hat{X}_s) \mid \mathcal{F}_s) \right\|^n \right)^{1/n} \leq \frac{C(n)}{N}
\]
for some deterministic constant $C(n)$ that does not depend upon $t$ or $N$. Applying the generalized Minkowski inequality, we have thus proved the following theorem, which is the $\mathbb{L}_n$-conditional bias (we shall also say, interchangeably, the conditional bias) of the mean.

**Theorem 2.4.** Consider the cases of (12) and (13). Assume that (16) holds and that $\mu(A) < 0$. Then for any $n \geq 1$ and $N$ sufficiently large, we have
\[
\sup_{t \geq 0} \mathbb{E} \left( \left\| \hat{m}_t - \hat{X}_t \right\|^n \right)^{1/n} \leq \frac{C(n)}{N},
\]
where $C(n) < \infty$ is a constant that does not depend upon $N$. 
3. Computing the normalizing constant

3.1. Estimation

Consider the time interval \([0, t]\), where \(t > 0\) is arbitrary; then the law of \((X, W)\) is absolutely continuous with respect to the law of the process \((X, Y)\), and its Radon–Nikodym derivative is \(Z_t^{-1}\). That is,

\[
\mathbb{E}[f(X_{0:t})g(Y_{0:t})Z_t(X, Y)^{-1}] = \mathbb{E}[f(X_{0:t})g(W_{0:t})].
\]

One can show (see [2, Exercise 3.14, p. 56]) that

\[
Z_t(X, Y) = \exp\left[ \int_0^t \left( \langle CX_s, R_s^{-1} dY_s \rangle - \frac{1}{2} \langle X_s, SX_s \rangle ds \right) \right].
\]

Following a standard approach (see, e.g., [2, Chapter 3]), we introduce a probability measure \(\tilde{P}\) by specifying its Radon–Nikodym derivative with respect to \(P\) to be given by

\[
d\tilde{P} = dP \bigg| \bigg| G_t = Z_t^{-1}.
\]

Under \(\tilde{P}\), the observation process \(Y\) is a scaled Brownian motion independent of \(X\); additionally, the law of the signal process \(X\) under \(\tilde{P}\) is the same as its law under \(P\). The proof of this statement is an immediate application of Girsanov’s theorem and follows in a similar manner to the proof of Proposition 3.13 in [2].

Moreover, for every \(f\) defined on the signal path space, we have the Bayes formula

\[
\mathbb{E}(f((X_s)_{s \in [0, t]}) | \mathcal{F}_t) = \frac{\tilde{\mathbb{E}}(f((X_s)_{s \in [0, t]}) Z_t(X, Y) | \mathcal{F}_t)}{\tilde{\mathbb{E}}(Z_t(X, Y) | \mathcal{F}_t)},
\]

(23)

where \(\tilde{\mathbb{E}}(\cdot)\) means expectation with respect to \(\tilde{P}\). The formula (23) is commonly known as the Kallianpur–Striebel formula. For a proof, see, e.g., Proposition 3.16 in [2]. Since, under \(\tilde{P}\), \(X\) and \(Y\) are independent, we can interpret \(\tilde{\mathbb{E}}(f((X_s)_{s \in [0, t]}) Z_t(X, Y) | \mathcal{F}_t)\) and \(\tilde{\mathbb{E}}(Z_t(X, Y) | \mathcal{F}_t)\) as integrals with respect to the law of the signal, where the integrand has the observation process path fixed as \(Y\). This interpretation can be made rigorous; see e.g. the robust representation formulation of the filtering problem described in Chapter 5 in [2]. We can therefore let \(Z_t(Y)\) be the likelihood function defined by

\[
Z_t(Y) := \mathbb{E}_Y(Z_t(X, Y)),
\]

where \(\mathbb{E}_Y(\cdot)\) stands for the expectation with respect to the signal process when the observation is fixed and independent of the signal. In this notation, the Bayes formula (23) takes the form

\[
\mathbb{E}(f((X_s)_{s \in [0, t]}) | \mathcal{F}_t) = \frac{\mathbb{E}_Y(f((X_s)_{s \in [0, t]}) Z_t(X, Y))}{\mathbb{E}_Y(Z_t(X, Y))}.
\]

We have

\[
dZ_t(X, Y) = \left[ \langle CX_t, R_t^{-1} dY_t \rangle - \frac{1}{2} \langle X_t, SX_t \rangle dt \right] Z_t + \frac{1}{2} \langle X_t, SX_t \rangle Z_t dt = Z_t(X, Y) \langle CX_t, R_t^{-1} dY_t \rangle.
\]
This implies that
\[
\begin{align*}
\mathbb{Z}_t(Y) &= 1 + \int_0^t \mathbb{Z}_s(Y)\mathbb{Z}_s(Y)^{-1} \mathbb{E}_Y \left( Z_s(X, Y) \{ C \hat{X}_s, R_2^{-1}dY_s \} \right) \\
&= 1 + \int_0^t \mathbb{Z}_s(Y) \{ C \hat{X}_s, R_2^{-1}dY_s \},
\end{align*}
\]
from which we conclude that
\[
\mathbb{Z}_t(Y) = \exp \left[ \int_0^t \left( \{ C \hat{X}_s, R_2^{-1}dY_s \} - \frac{1}{2} \{ \hat{X}_s, S \hat{X}_s \} ds \right) \right]
\]
or equivalently
\[
d\mathbb{Z}_t(Y) = \mathbb{Z}_t(Y) \{ C \hat{X}_t, R_2^{-1}dY_t \}.
\]
This suggests the estimator
\[
\mathbb{Z}_t^N(Y) = \exp \left[ \int_0^t \left( \{ C \hat{X}_s, R_2^{-1}dY_s \} - \frac{1}{2} \{ m_s, S m_s \} ds \right) \right],
\]
which satisfies the equation
\[
d\mathbb{Z}_t^N(Y) = \mathbb{Z}_t^N(Y) \{ C m_t, R_2^{-1}dY_t \}.
\]
We remark that any of the three processes (12)–(14) can be used to compute \( \mathbb{Z}_t^N(Y) \).

### 3.2. Justification of the estimator

We now seek to justify the estimator \( \mathbb{Z}_t^N(Y) \). We have
\[
\begin{align*}
\mathbb{Z}_t(X, Y) &= \mathbb{Z}_t(Y)^{-1} \mathbb{Z}_t(X, Y) \\
&= \exp \left[ \int_0^t \left( \{ C(X_s - \hat{X}_s), R_2^{-1}(dY_s - C \hat{X}_s ds) \} \\
- \left[ \frac{1}{2} \{ X_s, S X_s \} - \frac{1}{2} \{ \hat{X}_s, S \hat{X}_s \} - \{ X_s - \hat{X}_s, S \hat{X}_s \} \right] ds \right) \right] \\
&= \exp \left[ \int_0^t \left( \{ C(X_s - \hat{X}_s), R_2^{-1}(dY_s - C \hat{X}_s ds) \} - \frac{1}{2} \{ (X_s - \hat{X}_s), S(X_s - \hat{X}_s) \} ds \right) \right],
\end{align*}
\]
from which we conclude that
\[
d\mathbb{Z}_t(X, Y) = \mathbb{Z}_t(X, Y) \{ C(X_s - \hat{X}_s), R_2^{-1}(dY_s - C \hat{X}_s ds) \}.
\]
Now, we have
\[
Z_t^N(Y)Z_t^{-1}(Y) = \exp \left[ \int_0^t \left[ (C(m_s - \hat{X}_s), R_2^{-1}dY_s) - \frac{1}{2} \left[ (m_s, Sm_s) - (\hat{X}_s, S\hat{X}_s) \right] ds \right] \right]
\]
\[
= \exp \left[ \int_0^t \left[ (C(m_s - \hat{X}_s), R_2^{-1}dY_s) - \frac{1}{2} \left[ (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \right] ds \right] \right]
\times \exp \left[ \int_0^t \left[ \frac{1}{2} \left[ (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \right] ds \right] \right]
\]
\[
= \exp \left[ \int_0^t \left[ (C(m_s - \hat{X}_s), R_2^{-1}dY_s) - \frac{1}{2} \left[ (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \right] ds \right] \right]
\times \exp \left[ - \int_0^t \left[ C(m_s - \hat{X}_s), R^{-1}C\hat{X}_s \right] ds \right]
\]
\[
= \exp \left[ \int_0^t \left[ (C(m_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_sds)) - \frac{1}{2} \left[ (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \right] ds \right] \right].
\]

Observe that \( dY_s - C\hat{X}_sds \) is an \( \mathcal{F}_s \)-martingale increment. We also have
\[
d(Z_t^N(Y)Z_t^{-1}(Y)) = Z_t^N(Y)Z_t^{-1}(Y) \left\langle C(m_t - \hat{X}_t), R_2^{-1}(dY_t - C\hat{X}_tdt) \right\rangle. \tag{26}
\]

Now to conclude, it follows that \( Z_t^N(Y)Z_t^{-1}(Y) \) is a positive local martingale and therefore a supermartingale; hence
\[
\mathbb{E}(Z_t^N(Y)Z_t^{-1}(Y)) \leq \mathbb{E}(Z_0^N(Y)Z_0^{-1}(Y)) = 1
\]
for all \( t \geq 0 \). To justify the martingale property of \( Z_t^N(Y)Z_t^{-1}(Y) \) one can use, for example, an argument based on [12, Chapter 3, Corollary 5.14]. As a result we can deduce that \( Z_t^N(Y) \) is in some sense well-defined (i.e. \( Z_t^N(Y)Z_t^{-1}(Y) \) is almost surely finite), but a biased estimator of the normalization constant. Therefore, we will focus on the logarithm of the normalization constant, as it is this latter quantity that is used in practical algorithms. We begin by investigating the conditional bias.

### 3.3. Conditional bias

We now consider the estimate of the logarithm of the normalizing constant
\[
U_t^N(Y) := \log(Z_t^N(Y))
\]
with the notation \( U_t(Y) = \log(Z_t(Y)) \).
Set \( \hat{U}_t^N(Y) := \mathbb{E}(U_t^N(Y) \mid \mathcal{F}_t) \). Then, using (26), we have

\[
U_t^N(Y) - U_t(Y) = \log(\mathbb{Z}_t^N(Y)/\mathbb{Z}_t(Y)) = \int_0^t \left[ \langle C(m_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_sds) \rangle - \frac{1}{2} \langle (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \rangle \right] ds.
\]

This yields the bias formula

\[
\hat{U}_t^N(Y) - U_t(Y) = \int_0^t \left[ \langle C(\hat{m}_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_sds) \rangle - \frac{1}{2} \mathbb{E}(\langle (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \rangle \mid \mathcal{F}_s) ds \right].
\]

Taking the expectation in the above display, when (16) holds and \( \mu(A) < 0 \), and applying Theorem 2.1, we readily check that

\[
0 \leq \mathbb{E}(U_t(Y)) - \mathbb{E}(\hat{U}_t^N(Y)) = \frac{\mu(S)}{2} \int_0^t \mathbb{E}(\|m_s - \hat{X}_s\|^2) ds \leq \frac{Ct}{N},
\]

which is the bias, but is not of particular interest. So we will focus on the conditional bias

\[
\mathbb{E}\left(\left[ \hat{U}_t^N(Y) - U_t(Y) \right]^n \right)^{1/n}.
\]

In the case (14), recall (17). In this context, using the generalized Minkowski inequality and under the assumptions of Theorem 2.3, we find that

\[
\sup_{t \geq 0} \mathbb{E}\left(\left[ \hat{U}_t^N(Y) - U_t(Y) \right]^n \right)^{1/n} \leq \frac{C(n)}{N},
\]

where \( C(n) < \infty \) is a constant that does not depend upon \( N \).

For the cases of (12) and (13), we start with the fact that the conditional bias in (27) is decomposed into two terms,

\[
\alpha_t := \int_0^t \langle C(\hat{m}_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_sds) \rangle
\]

and

\[
\beta_t := \frac{1}{2} \int_0^t \mathbb{E}(\langle (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \rangle \mid \mathcal{F}_s) ds.
\]

Using the uniform estimates presented in Section 2.3 we have

\[
\mathbb{E}(\beta_t^n) \leq \frac{C(n)t}{N}
\]

for some deterministic constants \( C(n) \) that do not depend on the time horizon. On the other hand, combining the Burkholder–Davis–Gundy inequality with the uniform bias estimate (19), we have the rather crude estimate

\[
\mathbb{E}(\alpha_t^n) \leq C(n) \left( \frac{t}{N} \right)^{n/2}
\]
for some deterministic constants $C(n)$ that do not depend on the time horizon. Arguing as for the proof of Theorem 2.4, we check that

$$E(|\alpha_t|^n) \leq C(n) \left( \frac{\sqrt{t}}{N} \right)^n.$$  

Observe that the above improves the crude estimate stated in (28). This yields the following corollary.

**Corollary 3.1.** Consider the cases of (12) and (13). Assume that (16) holds and that $\mu(A) < 0$. Then for any $n \geq 1$, $t \geq 0$ and $N$ sufficiently large, we have that

$$E\left( \left| U^N_t(Y) - U_t(Y) \right|^n \right)^{1/n} \leq \frac{C(n)(t + \sqrt{t})}{N},$$

where $C(n) < \infty$ is a constant that does not depend upon $t$ or $N$.

### 3.4. $\mathbb{L}_n$-error

We now consider the $\mathbb{L}_n$-error of the estimate of the log of the normalizing constant, $U^N_t(Y) = \log(\tilde{Z}^N_t(Y))$.

In the cases of (12) and (13) we have the following.

**Proposition 3.1.** Consider the cases of (12) and (13). Assume that (16) holds and that $\mu(A) < 0$. Then for any $n \geq 1$, $t \geq 0$ and $N$ sufficiently large, we have that

$$E\left( \left| U^N_t(Y) - U_t(Y) \right|^n \right)^{1/n} \leq \frac{C(n)\sqrt{t} + t}{N},$$

where $C(n) < \infty$ is a constant that does not depend upon $t$ or $N$.

**Proof.** We have that

$$U^N_t(Y) - U_t(Y) = \int_0^t \left[ \langle C(m_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_s ds) \rangle - \frac{1}{2} \langle (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \rangle \right] ds.$$  

So one can apply the Minkowski inequality to yield that

$$E\left( \left| U^N_t(Y) - U_t(Y) \right|^n \right)^{1/n} \leq E\left( \left| \int_0^t \langle C(m_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_s ds) \rangle ds \right|^n \right)^{1/n} \leq \frac{C(n)\sqrt{t} + t}{N}.$$ 

For the left term on the right-hand side of (29) one can use the Burkholder–Davis–Gundy inequality along with Theorem 2.3 to yield that

$$E\left( \left| \int_0^t \langle C(m_s - \hat{X}_s), R_2^{-1}(dY_s - C\hat{X}_s ds) \rangle ds \right|^n \right)^{1/n} \leq \frac{C(n)\sqrt{t}}{\sqrt{N}}.$$
for some $C(n) < \infty$ which is a constant that does not depend upon $t$ or $N$. For the right term on the right-hand side of (29) one can use the generalized Minkowski inequality and Theorem 2.1 to give
\[
E\left( \left| \int_0^t \langle (m_s - \hat{X}_s), S(m_s - \hat{X}_s) \rangle ds \right|^n \right)^{1/n} \leq \frac{C(n) t}{N}
\]
for some $C(n) < \infty$ which is a constant that does not depend upon $t$ or $N$. The proof is now easily concluded. \hfill \Box

In the case of (13) we can refine further.

**Proposition 3.2.** Consider the case of (13). Assume that $A, R_1^{1/2}$ and $(A, C)$ are controllable and observable pairs and that $\mu(A - P_{\infty}S) < 0$ and $S \in \mathbb{S}_r^+$. Then for any $N \geq 2$ there exists a $t(N) > 0$, with $t(N) \to \infty$ as $N \to \infty$, such that for any $t \in [0, t(N)]$,
\[
E\left( |U_t^N(Y) - U_t(Y)|^n \right)^{1/n} \leq C(n) \left( \frac{\sqrt{t}}{N} + \frac{t}{N} \right),
\]
where $C(n) < \infty$ is a constant that does not depend upon $t$ or $N$.

**Proof.** The proof is essentially like that of Proposition 3.1, except one should use Theorem 2.2 instead of Theorem 2.1. \hfill \Box

In the case of (14) we have the following, whose proof is again similar to those of the above results.

**Proposition 3.3.** Consider the case of (14). Assume that $A, R_1^{1/2}$ and $(A, C)$ are controllable and observable pairs. Then for any $t \geq 0$ and $N \geq r_1$, we have that
\[
E\left( |U_t^N(Y) - U_t(Y)|^n \right)^{1/n} \leq \frac{C(n)}{\sqrt{N}},
\]
where $C(n) < \infty$ is a constant that does not depend upon $t$ or $N$.

Both Proposition 3.1 and Proposition 3.2 establish that one can estimate the log of the normalization constant using the ensemble KB-type filters, with an MSE (for instance) that grows at most linearly in time. Proposition 3.3 provides an error that is uniform in time, mostly following from the deterministic nature of the algorithm and the dependence on standard KB theory.

One can say more, as we now state, when considering the average estimation error over a window of time $w$. We restrict ourselves to the MSE below. In the cases of both (12) and (13) we have the time-uniform upper bound.

**Corollary 3.2.** Consider the case of (12) and (13). Assume that (16) holds and that $\mu(A) < 0$. Then for any $t > 0$, $0 < w < t$ and $N$ sufficiently large, we have that
\[
E \left[ \left( \frac{1}{w} \left\{ \log \left( \frac{Z_t^N(Y)}{Z_{t-w}^N(Y)} \right) - \log \left( \frac{Z_t(Y)}{Z_{t-w}(Y)} \right) \right\} \right)^2 \right] \leq \frac{C}{N},
\]
where $C < \infty$ is a constant that does not depend upon $t$, $w$ or $N$.

In the case of (13) we refine further and have the following time-uniform upper bound.
**Corollary 3.3.** Consider the case of \((13)\). Assume that \((A, R_1^{1/2})\) and \((A, C)\) are controllable and observable pairs and that \(\mu(A - P_\infty S) < 0\) and \(S \in \mathbb{S}_+^n\). Then for any \(N \geq 2\) there exists a \(t(N) > 0\), with \(t(N) \to \infty\) as \(N \to \infty\), such that for any \(t \in (0, t(N)]\), \(0 < w < t\), we have

\[
\mathbb{E}\left[ \frac{1}{w} \left\{ \log\left( \frac{Z_i^N(Y)}{Z_{t-w}(Y)} \right) - \log\left( \frac{Z_i(Y)}{Z_{t-w}(Y)} \right) \right\}^2 \right] \leq \frac{C}{N},
\]

where \(C < \infty\) is a constant that does not depend upon \(t, w\) or \(N\).

Note that, as we require \(t \in (0, t(N)]\) and \(t(N) = O(\log(N))\) in Corollary 3.4, this may not be as useful as Corollary 3.2, but the assumption of a stable system in Corollary 3.3 is much stronger than the hypotheses in Corollary 3.3; see [5] for a discussion.

### 3.5. Simulation results

For \((i, k, L) \in \{1, \cdots, N\} \times \mathbb{N}_0 \times \mathbb{N}_0\), let \(\Delta_L = 2^{-L}\) and consider the Euler discretization of \((12)-(14)\),

\[
\begin{align*}
\xi_{(k+1)\Delta_L}^i &= \xi_{k\Delta_L}^i + A \xi_{k\Delta_L}^i \Delta_L + R_1^{1/2} \left[ \overline{W}_{(k+1)\Delta_L} - \overline{W}_{k\Delta_L} \right] + \frac{p_{k\Delta_L} C^2 R_2^{-1} \left( \left\{ Y_{(k+1)\Delta_L} - Y_{k\Delta_L} \right\} - \left[ C \xi_{k\Delta_L}^i \Delta_L + R_2^{1/2} \left( \overline{V}_{(k+1)\Delta_L} - \overline{V}_{k\Delta_L} \right) \right] \right)}, \\
\xi_{(k+1)\Delta_L}^i &= \xi_{k\Delta_L}^i + A \xi_{k\Delta_L}^i \Delta_L + R_1^{1/2} \left[ \overline{W}_{(k+1)\Delta_L} - \overline{W}_{k\Delta_L} \right] + \frac{p_{k\Delta_L} C^2 R_2^{-1} \left( \left\{ Y_{(k+1)\Delta_L} - Y_{k\Delta_L} \right\} - \left( \xi_{k\Delta_L}^i \Delta_L + m_{k\Delta_L} \right) \right)}, \\
\xi_{(k+1)\Delta_L}^i &= \xi_{k\Delta_L}^i + A \xi_{k\Delta_L}^i \Delta_L + R_1(p_{k\Delta_L})^{-1} \left( \xi_{k\Delta_L}^i - m_{k\Delta_L} \right) \Delta_L + \frac{p_{k\Delta_L} C^2 R_2^{-1} \left( \left\{ Y_{(k+1)\Delta_L} - Y_{k\Delta_L} \right\} - \left( \xi_{k\Delta_L}^i + m_{k\Delta_L} \right) \right)},
\end{align*}
\]

and the discretization of \(Z_T^N(Y)\), for \(T \in \mathbb{N}\),

\[
Z_T^{N, L}(Y) = \exp \left\{ \sum_{k=0}^{T/\Delta_L-1} \left( m_{k\Delta_L}, C^2 R_2^{-1} \left[ Y_{(k+1)\Delta_L} - Y_{k\Delta_L} \right] - \frac{1}{2} \langle m_{k\Delta_L}, S m_{k\Delta_L} \rangle \Delta_L \right) \right\},
\]

such that

\[
\begin{align*}
p_{k\Delta_L}^N &= \frac{1}{N-1} \sum_{i=1}^{N} \left( \xi_{k\Delta_L}^i - m_{k\Delta_L}^N \right) \left( \xi_{k\Delta_L}^i - m_{k\Delta_L}^N \right)^\top, \\
m_{k\Delta_L}^N &= \frac{1}{N} \sum_{i=1}^{N} \xi_{k\Delta_L}^i.
\end{align*}
\]

We take \(T \in \mathbb{N}\), as this is all that will be used in our numerical simulations.

Our objective is to show that the MSE of the estimate of the log of the normalization constant in the cases (F1) and (F2) is \(O\left( \frac{1}{N} \right)\) and in the case (F3) is \(O\left( \frac{1}{N^2} \right)\). We take \(r_1 = r_2 = 1\), \(A = -2, R_1^{1/2} = 1, R_2^{1/2} = 2, C\) a uniform random number in \((0, 1]\), and \(\xi_{0,i,i.d.} \sim \mathcal{N}(0.5, 0.2)\) (normal distribution with mean 0.5 and variance 0.2). In Tables 1–2 and Figure 1, we show that the
### Table 1. The MSE and MSE/(t/N) for \( N \in \{250, 500, 1000\} \) in the (F1) case.

| \( t \) | \( N = 1000 \) | \( N = 500 \) | \( N = 250 \) |
|---|---|---|---|
| MSE | MSE/(t/N) | MSE | MSE/(t/N) | MSE | MSE/(t/N) |
| 50 | 2.194E-04 | 4.4E-03 | 5.404E-04 | 5.4E-03 | 8.903E-04 | 4.5E-03 |
| 100 | 5.483E-04 | 5.5E-03 | 1.442E-03 | 7.2E-03 | 3.082E-03 | 7.7E-03 |
| 200 | 1.085E-03 | 5.4E-03 | 2.901E-03 | 7.3E-03 | 4.442E-03 | 5.6E-03 |
| 400 | 2.283E-03 | 5.7E-03 | 5.047E-03 | 6.3E-03 | 1.017E-02 | 6.4E-03 |
| 800 | 4.894E-03 | 6.1E-03 | 8.579E-03 | 5.4E-03 | 1.985E-02 | 6.2E-03 |
| 1600 | 9.716E-03 | 6.1E-03 | 2.039E-02 | 6.4E-03 | 2.904E-02 | 4.5E-03 |
| 3200 | 1.974E-02 | 6.2E-03 | 3.504E-02 | 5.5E-03 | 7.835E-02 | 6.1E-03 |
| 6400 | 3.571E-02 | 5.6E-03 | 7.087E-02 | 5.5E-03 | 1.599E-01 | 6.2E-03 |

### Table 2. The MSE and MSE/(t/N) for \( N \in \{250, 500, 1000\} \) in the (F2) case.

| \( t \) | \( N = 1000 \) | \( N = 500 \) | \( N = 250 \) |
|---|---|---|---|
| MSE | MSE/(t/N) | MSE | MSE/(t/N) | MSE | MSE/(t/N) |
| 50 | 1.137E-03 | 5.7E-03 | 5.127E-04 | 5.1E-03 | 3.317E-04 | 6.6E-03 |
| 100 | 2.157E-03 | 5.4E-03 | 1.128E-03 | 5.6E-03 | 8.892E-04 | 8.9E-03 |
| 200 | 4.601E-03 | 6.1E-03 | 2.708E-03 | 6.8E-03 | 1.260E-03 | 6.3E-03 |
| 400 | 9.777E-03 | 6.1E-03 | 5.250E-03 | 6.6E-03 | 2.436E-03 | 6.1E-03 |
| 800 | 1.979E-02 | 6.2E-03 | 9.949E-03 | 6.2E-03 | 4.639E-03 | 5.8E-03 |
| 1600 | 4.901E-02 | 7.7E-03 | 1.595E-02 | 5.0E-03 | 8.895E-03 | 5.6E-03 |
| 3200 | 9.593E-02 | 7.5E-03 | 3.001E-02 | 4.7E-03 | 1.925E-02 | 6.0E-03 |
| 6400 | 1.744E-01 | 6.8E-03 | 5.415E-02 | 4.2E-03 | 3.635E-02 | 5.7E-03 |

**Figure 1.** The MSE associated to the EnKBF in the (F1) (left) and (F2) (right) cases. This plots the MSE against the time parameter for \( N \in \{250, 500, 1000\} \). For each \( N = 250, 500, 1000 \), the MSE in (F1) is \( \approx 5.9E-03 \left( \frac{t}{N} \right) \), \( 6.1E-03 \left( \frac{t}{N} \right) \), \( 5.6E-03 \left( \frac{t}{N} \right) \), and in (F2) is \( \approx 6.4E-03 \left( \frac{t}{N} \right) \), \( 5.5E-03 \left( \frac{t}{N} \right) \), \( 6.4E-03 \left( \frac{t}{N} \right) \), respectively.

The rate of the MSE of the estimate in (31) for the cases (F1) and (F2) is \( \mathcal{O}(1/N) \), even though we have used a naive time discretization for our results. In Table 3 and Figure 2 we show that the rate in the (F3) case is \( \mathcal{O}(1/N) \).
TABLE 3. The MSE and $\text{MSE} \times N$ for $t = 100$ in the (F3) case.

| $N$  | MSE    | $\text{MSE} \times N$ |
|------|--------|------------------------|
| 50   | 1.073E-05 | 5.4E-04               |
| 100  | 4.951E-06 | 5.0E-04               |
| 200  | 2.867E-06 | 5.7E-04               |
| 400  | 1.492E-06 | 6.0E-04               |
| 800  | 8.157E-07 | 6.5E-04               |
| 1600 | 3.119E-07 | 5.0E-04               |
| 3200 | 1.773E-07 | 5.7E-04               |
| 6400 | 6.344E-08 | 4.1E-04               |

FIGURE 2. The MSE associated to the EnKBF in the (F3) case. This plots the MSE against the ensemble size $N$ for fixed time $t = 100$. MSE $\approx 5.4E-04/N$.

We also compare the error-to-cost rates of the EnKBF estimation of the log-normalizing constant using (F1), (F2), and (F3). We take $r_1 = r_2 = 2, A = 2 I_d$,

$$R_1^{1/2} = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix},$$

$R_2^{1/2} = I_d, C$ a uniform random matrix, and $\xi_0^i \sim \mathcal{N}(0.1_1, 0.05 I_d)$. The estimate of the log-normalizing constant is the mean of 104 simulations computed at discretization levels $L \in \{8, 9, 10, 11\}$. The number of ensembles $N_L$ is different for each EnKBF variant and is chosen so that all three cases return approximately the same MSE. The reference value is the mean of 208 simulations at level 13 with ensemble size of 500. As we see in Figure 3, to attain a certain MSE, the cost of (F1) is larger than that of (F2) and the cost of (F2) is larger than that of (F1). This is expected as (F1) has two sources of noise, while (F2) has only one and (F3) is fully deterministic.
4. Application to static parameter estimation

4.1. Approach

We now assume that there are a collection of unknown parameters, \( \theta \in \Theta \subseteq \mathbb{R}^{d_{\theta}} \), associated to the model (1). For instance \( \theta \) could consist of some or all of the elements of \( A, C, R_1, R_2 \). We will then include an additional subscript \( \theta \) in each of the mathematical objects that have been introduced in the previous two sections. As an example, we would write

\[
Z_{t,\theta}(X, Y) = \exp \left[ \int_0^t \left( CX_s, R_2^{-1} dY_s \right) - \frac{1}{2} \langle X_s, S X_s \rangle ds \right].
\]

For \( 0 < s < t \) we introduce the notation

\[
Z_{s,t,\theta}(Y) := \frac{Z_{t,\theta}(Y)}{Z_{s,\theta}(Y)}
\]

with the obvious extension to \( Z_{s,t,\theta}^{N}(Y) := Z_{t,\theta}^{N}(Y)/Z_{s,\theta}^{N}(Y) \).

The basic idea of our approach is to consider the recursive estimation of \( \theta \) on the basis of the arriving observations. In particular, for notational convenience, we shall consider a method which will update our current estimate of \( \theta \) at unit times. Our objective is to follow a recursive maximum likelihood (RML) method (e.g. [13]) which is based upon the following update at any time \( t \in \mathbb{N} \):

\[
\theta_t = \theta_{t-1} + \zeta_t \nabla_{\theta} \log \left( Z_{t-1,t,\theta}(Y) \right) \bigg|_{\theta = \theta_{t-1}},
\]

where \( \{\zeta_t\}_{t \in \mathbb{N}} \) is a sequence of real numbers with \( \zeta_t > 0, \sum_{t \in \mathbb{N}} \zeta_t = \infty, \sum_{t \in \mathbb{N}} \zeta_t^2 < \infty \). Computing the gradient of \( Z_{t-1,t,\theta}(Y) \) can be computationally expensive, so our approach is to use a type of finite difference estimator via SPSA. We note that a classical finite difference estimator would require \( 2d_{\theta} \) evaluations of \( Z_{t-1,t,\theta}(Y) \), whereas the SPSA method keeps this evaluation down to 2; as computational cost is a concern, we prefer the aforementioned method.
Algorithm 1 Algorithm for Online Parameter Estimation

1. Initialization: Set an initial $\theta_0 \in \Theta$ and choose two step sizes $\{\zeta_t\}_{t \in \mathbb{N}}$ and $\{\nu_t\}_{t \in \mathbb{N}}$ such that $\zeta_t > 0$, $\zeta_t, \nu_t \to 0$, $\sum_{t \in \mathbb{N}} \zeta_t = \infty$, $\sum_{t \in \mathbb{N}} \frac{\zeta_t^2}{\nu_t^2} < \infty$. Set $t = 1$ and generate i.i.d. the initial ensemble from $\eta_{0,0}$.

2. Iteration:
   - For $k \in \{1, \ldots, d_\theta\}$, independently, sample $\Psi_t(k)$ from a Bernoulli distribution with success probability $1/2$ and support $\{-1, 1\}$.
   - Set $\theta_{t-1}^+ = \theta_{t-1} + \nu_t \Psi_t$, $\theta_{t-1}^- = \theta_{t-1} - \nu_t \Psi_t$.
   - Using the EnKBF, with samples simulated under $\theta_{t-1}$, generate estimates $Z_{t-1,t,\theta_{t-1}^+}(Y)$ and $Z_{t-1,t,\theta_{t-1}^-}(Y)$.
   - Set, for $k \in \{1, \ldots, d_\theta\}$, $\theta_t(k) = \theta_{t-1}(k) + \zeta_t \frac{1}{2 \nu_t \Psi_t(k)} \log \left( \frac{Z_{t-1,t,\theta_{t-1}^+}(Y)}{Z_{t-1,t,\theta_{t-1}^-}(Y)} \right)$.
   - Run the EnKBF from $t - 1$ (starting with samples simulated under $\theta_{t-1}$) up-to time $t$ using the parameter $\theta_t$. Set $t = t + 1$ and return to the start of 2.

Our approach is given in Algorithm 1, noting that we will use the argument $(k)$ to denote the $k$th element of a vector. We note that in practice, one must run a time-discretization of the EnKBF, rather than the EnKBF itself. We remark also that the use of SPSA for parameter estimation associated to hidden Markov models has been considered previously, for instance in [15].

4.2. Simulation results

We use Algorithm 1 along with the Euler discretization in (30) to estimate the parameters in three different models. In particular, we show that the algorithm works in both linear and nonlinear models. In all of our examples, the data are generated from the model with the true parameters. We also remark that in the nonlinear examples below we replace the term $A_{\xi_{k\Delta t}^i}$ in the cases (F1)–(F3) in (30) with $f(\xi_{k\Delta t}^i)$ for some nonlinear function $f$. Then the mean is approximated through the formula

$$m_{k\Delta t}^N = \frac{1}{N} \sum_{i=1}^N \xi_{k\Delta t}^i$$

and afterwards plugged into $Z_{t-1,t,\theta_{t-1}^\pm}(Y)$. 
4.2.1. Linear Gaussian model. In the first example, we take $A = \theta_1 \text{Id}$; $R_1^{1/2} = \theta_2 R$, where

$$R = \begin{bmatrix}
1 & 0.5 \\
0.5 & 1 & 0.5 \\
& 0.5 & \ddots & \ddots \\
& & \ddots & 0.5 \\
& & & 0.5 & 1
\end{bmatrix};$$

$C = \alpha_1(r_1, r_2) C^*$, where $C^*$ is a uniform random matrix and $\alpha_1(r_1, r_2)$ is a constant; and $R_2^{1/2} = \alpha_2(r_2) \text{Id}$, where $\alpha_2(r_2)$ is a constant. In Figures 4–9 we show the results for the parameter estimation of $\theta_1$ and $\theta_2$ in the cases $r_1 = r_2 = 2$ and $r_1 = r_2 = 100$. In all cases we take $N = 100$, except in the case when $r_1 = r_2 = 100$ in (F3), where we take $N = 200$ to ensure the invertibility of $p_N^N$; otherwise, the condition number of $p_N^N$ will be huge. The discretization level is $L = 8$ in all cases. The initial state is $X_0 \sim N(4 \mathbf{1}_{r_1}, \text{Id})$, where $\mathbf{1}_{r_1}$ is a vector of 1s in $\mathbb{R}^{r_1}$.

The results show that in a reasonable case, one can estimate low-dimensional parameters using RML via SPSA. We now consider a few nonlinear models.

4.2.2. Lorenz 63 model. We now consider the following nonlinear model with $r_1 = r_2 = 3$, which is a simplified mathematical model for atmospheric convection:

$$dX_t = f_1(X_t) \, dt + R_1^{1/2} \, dW_t,$$

$$dY_t = C X_t \, dt + R_2^{1/2} \, dV_t,$$

where

$$f_1(X_t) = \theta_1(X_t(2) - X_t(1)),$$
$$f_2(X_t) = \theta_2 X_t(1) - X_t(2) - X_t(1) X_t(3),$$
$$f_3(X_t) = X_t(1) X_t(2) - \theta_3 X_t(3),$$

with

$$C = \alpha_1(r_1, r_2) C^*,$$

$C^*$ is a uniform random matrix and $\alpha_1(r_1, r_2)$ is a constant; and $R_2^{1/2} = \alpha_2(r_2) \text{Id}$, where $\alpha_2(r_2)$ is a constant.
Figure 5. The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of \((\theta_1, \theta_2)\) in the case (F1) with \(r_1 = r_2 = 100\). The initial values of the parameters are \((-1, 2)\). The green horizontal lines represent the true parameter values \(\left(\theta^*_1, \theta^*_2\right)\) = \((-2, 1)\). We take \(\nu_t = t^{-0.1}\) and \(\zeta_t = 0.1\) when \(t \leq 400\) and \(\zeta_t = 3 \times t^{-0.601}\) otherwise.

Figure 6. The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of \((\theta_1, \theta_2)\) in the case (F2) with \(r_1 = r_2 = 2\). The initial values of the parameters are \((-1, 2)\). The green horizontal lines represent the true parameter values \(\left(\theta^*_1, \theta^*_2\right)\) = \((-2, 1)\). We take \(\nu_t = t^{-0.1}\) and \(\zeta_t = 0.09\) when \(t \leq 300\) and \(\zeta_t = t^{-0.7}\) otherwise.

Figure 7. The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of \((\theta_1, \theta_2)\) in the case (F2) with \(r_1 = r_2 = 100\). The initial values of the parameters are \((-1, 2)\). The green horizontal lines represent the true parameter values \(\left(\theta^*_1, \theta^*_2\right)\) = \((-2, 1)\). We take \(\nu_t = t^{-0.1}\) and \(\zeta_t = 0.08\) when \(t \leq 400\) and \(\zeta_t = 3 \times t^{-0.64}\) otherwise.
Figure 8. The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of \((\theta_1, \theta_2)\) in the case (F3) with \(r_1 = r_2 = 2\). The initial values of the parameters are \((-1, 2.2)\). The green horizontal lines represent the true parameter values \((\theta^*_1, \theta^*_2) = (-2, 1)\). We take \(\nu_t = t^{-0.1}\) and \(\zeta_t = 0.09\) when \(t \leq 100\) and \(\zeta_t = t^{-0.8}\) otherwise.

Figure 9. The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of \((\theta_1, \theta_2)\) in the case (F3) with \(r_1 = r_2 = 100\). The initial values of the parameters are \((-1, 2)\). The green horizontal lines represent the true parameter values \((\theta^*_1, \theta^*_2) = (-2, 1)\). We take \(\nu_t = t^{-0.1}\) and \(\zeta_t = 0.09\) when \(t \leq 200\) and \(\zeta_t = 2 \times t^{-0.601}\) otherwise.

where \(X_t(i)\) is the \(i\)th component of \(X_t\). We also have \(R_1^{1/2} = Id\),

\[
C_{ij} = \begin{cases} 
\frac{1}{2} & \text{if } i = j, \\
\frac{1}{2} & \text{if } i = j - 1, \\
0 & \text{otherwise}, 
\end{cases}
\]

and \((R_2^{1/2})_{ij} = 2q \left( \frac{2}{3} \min(|i - j|, r_2 - |i - j|) \right)\), \(i, j \in \{1, 2, 3\}\), where

\[
q(x) = \begin{cases} 
1 - \frac{3}{2}x + \frac{1}{2}x^3 & \text{if } 0 \leq x \leq 1, \\
0 & \text{otherwise}. 
\end{cases}
\]

In Figures 10–12 we show the results for the parameter estimation of \(\theta = (\theta_1, \theta_2, \theta_3)\) in the F1, F2, and F3 cases. The ensemble size is \(N = 100\) and the discretization level is \(L = 8\) in all cases. The initial state is \(X_0 \sim \mathcal{N}(1_{r_1}, \frac{1}{2}Id)\), where \(1_{r_1}\) is a vector of 1s in \(\mathbb{R}^{r_1}\).
Log-normalization constant estimation using EnKBF

4.2.3. Lorenz 96 model. Finally, we consider the nonlinear Lorenz 96 model, with $r_1 = r_2 = 40$. The solution of this model has a chaotic behavior and it describes the evolution of a scalar quantity on a circle of latitude. The model is as follows:

\[
\begin{align*}
    dX_t &= f(X_t) \, dt + R_1^{1/2} \, dW_t, \\
    dY_t &= X_t \, dt + R_2^{1/2} \, dV_t,
\end{align*}
\]

**FIGURE 10.** The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of $(\theta_1, \theta_2, \theta_3)$ in the case (F1). The initial values of the parameters are $(7.5, 26.7, 6.5)$. The green horizontal lines represent the true parameter values $(\theta^*_1, \theta^*_2, \theta^*_3) = \left(10, 28, \frac{8}{3}\right)$. We take $v_t = t^{-0.2}$ and $\zeta_t = 0.0314$ when $t \leq 100$ and $\zeta_t = t^{-0.71}$ otherwise.

**FIGURE 11.** The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of $(\theta_1, \theta_2, \theta_3)$ in the case (F2). The initial values of the parameters are $(7.5, 26.7, 6.5)$. The green horizontal lines represent the true parameter values $(\theta^*_1, \theta^*_2, \theta^*_3) = \left(10, 28, \frac{8}{3}\right)$. We take $v_t = t^{-0.2}$ and $\zeta_t = 0.0139$ when $t \leq 300$ and $\zeta_t = t^{-0.75}$ otherwise.

**FIGURE 12.** The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of $(\theta_1, \theta_2, \theta_3)$ in the case (F3). The initial values of the parameters are $(7.5, 26.7, 6.5)$. The green horizontal lines represent the true parameter values $(\theta^*_1, \theta^*_2, \theta^*_3) = \left(10, 28, \frac{8}{3}\right)$. We take $v_t = t^{-0.2}$ and $\zeta_t = 0.0314$ when $t \leq 100$ and $\zeta_t = t^{-0.71}$ otherwise.
The curves along with their average (gray dashed lines) are trajectories from the execution of Algorithm 1 for the estimation of $\theta$ in the case (F1) (left), (F2) (middle), and (F3) (right). The initial value of $\theta$ is 10. The green horizontal lines represent the true parameter value $\theta^* = 8$. We take $\nu_t = t^{-0.1}$ and $\zeta_t = 0.0314$ when $t \leq 50$ and $\zeta_t = t^{-0.75}$ otherwise.

where

$$f_i(X_t) = (X_t(i + 1) - X_t(i - 2))X_t(i - 1) - X_t(i) + \theta,$$

where $X_t(i)$ is the $i$th component of $X_t$, and it is assumed that $X_t(-1) = X_t(r_1 - 1)$, $X_t(0) = X_t(r_1)$, and $X_t(r_1 + 1) = X_t(1)$. We also have $R_1^{1/2} = \sqrt{2}I_d$ and $R_2^{1/2} = \frac{1}{2}I_d$. Here $\theta$ represents the external force in the system, while $(X_t(i + 1) - X_t(i - 2))X_t(i - 1)$ is the advection-like term and $-X_t(i)$ is the damping term. In Figure 13, we show the results for the parameter estimation of $\theta$ in the (F1), (F2), and (F3) cases. The ensemble size is $N = 100$ and the discretization level is $L = 8$ in all cases. In the (F1) and (F2) cases, $X_t$ is initialized as follows: $X_0(1) = 8.01$ and $X_0(k) = 8$ for $1 < k \leq 40$. In (F3), to avoid having the matrix $p_0^N$ equal to zero, we take $X_0 \sim N(81_{r_1}, 0.05I_d)$.

Acknowledgements

We thank an editor and two referees for very useful comments which have greatly improved the paper.

Funding information

A. J. and H. R. were supported by KAUST baseline funding. D. C. was partially supported by EU project STUOD—DLV-856408.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process for this article.

References

[1] Allen, J. I., Eknés, M., and Evensen, G. (2002). An Ensemble Kalman Filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. *Ann. Geophys.* **20**, 1–13.
[2] Bain, A. and Crisan, D. (2009). *Fundamentals of Stochastic Filtering*. Springer, New York.
[3] Bishop, A. N. and Del Moral, P. (2017). On the stability of Kalman–Bucy diffusion processes. *SIAM J. Control Optimization* **55**, 4015–4047.
[4] Bishop, A. N. and Del Moral, P. (2019). Stability Properties of Systems of Linear Stochastic Differential Equations with Random Coefficients. *SIAM J. Control Optimization* **57**, 1023–1042.
[5] Bishop, A. N. and Del Moral, P. (2020). On the mathematical theory of ensemble (linear-gaussian) Kalman–Bucy filtering. Preprint. Available at [https://arxiv.org/abs/2006.08843](https://arxiv.org/abs/2006.08843).
[6] Bishop, A. N. and Del Moral, P. (2020). On the stability of matrix-valued Riccati diffusions. *Electron. J. Prob.* **24**, 1–40.
[7] CÉROU, F., DEL MORAL, P. AND GUYADER, A. (2011). A non-asymptotic variance theorem for unnormalized Feynman–Kac particle models. *Ann. Inst. H. Poincaré Prob. Statist.* **47**, 629–649.

[8] DEL MORAL, P. (2013). *Mean Field Simulation for Monte Carlo Integration*. Chapman and Hall, London.

[9] DEL MORAL, P. AND TUGAUT, J. (2018). On the stability and the uniform propagation of chaos properties of ensemble Kalman–Bucy filters. *Ann. Appl. Prob.* **28**, 790–850.

[10] DROVANDI, C., EVERITT, R., GOLIGHTLY, A. AND PRANGLE, D. (2019). Ensemble MCMC: accelerating pseudo-marginal MCMC for state space models using the ensemble Kalman filter. Preprint. Available at [https://arxiv.org/abs/1906.02014v2](https://arxiv.org/abs/1906.02014v2).

[11] EVENSEN, G. (1994). Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162.

[12] KARATZAS, I. AND SHREVE, S. (1991). *Brownian Motion and Stochastic Calculus*, 2nd edn. Springer, New York.

[13] LE GLAND, F. AND MEVEL, M. (1997). Recursive identification in hidden Markov models. In *Proc. 36th IEEE Conference on Decision and Control*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 3468–3473.

[14] LE GLAND, F., MONBET, V. AND TRAN, V.-D. (2011). Large sample asymptotics for the ensemble Kalman filter. In *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, pp. 598–631.

[15] POYIADJIS, G., SINGH, S. S. AND DOUCET, A. (2006). Gradient-free maximum likelihood parameter estimation with particle filters. In *Proc. American Control Conference*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 6–9.

[16] REICH, S. AND COTTER, C. (2013). Ensemble filter techniques for intermittent data assimilation. In *Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences*, eds M. Cullen, M. A. Freitag, S. Kindermann and R. Scheichl, Walter de Gruyter, Berlin, pp. 91–134.

[17] SAKOV, P. AND OKE, P. R. (2008). A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A* **60**, 361–371.

[18] SKJERVHEIM, J.-A. *et al.* (2005). Incorporating 4D seismic data in reservoir simulation models using ensemble Kalman filter. *SPE* **12**, 95789.

[19] SPALL, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* **37**, 332–341.

[20] STUART, A. M., LAW, K. J. H., and ZYGALAKIS, K. (2015). *Data Assimilation: A Mathematical Introduction*. Springer, New York.

[21] TONG, X. T., MAJDA, A. J. AND KELLY, D. (2016). Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity* **29**, 657–691.

[22] YANG, T., MEHTA, P. AND MEYN, S. (2013). Feedback particle filters. *IEEE Trans. Automatic Control* **58**, 2465–2480.