

Adapting the Number of Particles in Sequential Monte Carlo Methods through an Online Scheme for Convergence Assessment

Víctor Elvira^{*}, Joaquín Míguez[†], Petar M. Djurić[‡]

Abstract

Particle filters are broadly used to approximate posterior distributions of hidden states in state-space models by means of sets of weighted particles. While the convergence of the filter is guaranteed when the number of particles tends to infinity, the quality of the approximation is usually unknown but strongly dependent on the number of particles. In this paper, we propose a novel method for assessing the convergence of particle filters in an online manner, as well as a simple scheme for the online adaptation of the number of particles based on the convergence assessment. The method is based on a sequential comparison between the actual observations and their predictive probability distributions approximated by the filter. We provide a rigorous theoretical analysis of the proposed methodology and, as an example of its practical use, we present simulations of a simple algorithm for the dynamic and online adaptation of the number of particles during the operation of a particle filter on a stochastic version of the Lorenz system.

Index Terms

Particle filtering, sequential Monte Carlo, convergence assessment, predictive distribution, convergence analysis, computational complexity, adaptive complexity.

I. INTRODUCTION

Many problems in science and engineering can be described by dynamical models where hidden states of the systems change over time and observations that are functions of the states are available. Often, the observations are sequentially acquired and the interest is in making recursive inference on the hidden states. In many applications, the Bayesian approach to the problem is adopted because it allows for optimal inclusion of prior knowledge of the unknown state in the estimation process [1], [2]. In this case, the prior information and the likelihood function that relates the hidden state and the observation are combined yielding a posterior distribution of the state.

Exact Bayesian inference, however, is only possible in a small number of scenarios, including linear Gaussian state-space models (using the Kalman filter [3], [4]) and finite state-space hidden Markov models (HMM filters [5]). Therefore, in many other practical problems, only approximate inference methods can be used. One class of suboptimal methods is particle filtering, which is also known as sequential Monte Carlo sampling [6], [7], [8], [9], [10]. Since the publication of [11], where the sampling importance resampling (SIR) filter was introduced, particle filtering has received outstanding attention in research and practice.

^{*}Víctor Elvira is with the Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés 28911 (Spain). Email: velvira@tsc.uc3m.es

[†]Joaquín Míguez is with the School of Mathematical Sciences, Queen Mary University, London E1 4NS (UK). Email: j.miguez@qmul.ac.uk

[‡]Petar M. Djurić is with the Dept. of Electrical and Computer Engineering, Stony Brook University, New York 11794 (USA). Email: djuric@ece.sunysb.edu

Particle filters approximate posterior distributions of the hidden states sequentially and recursively. They do it by exploiting the principle of importance sampling and by using sets of weighted particles [6], [7], [12].

The key parameter of particle filters is the number of particles. It can be proved that the rate of convergence of the approximate probability distribution towards the true posterior is inversely proportional to the square root of the number of particles used in the filter [13], [12]. This, too, entails that the filter “perfectly” approximates the posterior distribution when the number of particles tends to infinity. However, since the computational cost grows with the number of particles, practitioners must choose a specific number of particles in the design of their filters.

In many applications, the observations arrive sequentially, and there is a strict deadline for processing each new observation. Then, one could argue that the best solution in terms of filter performance is to increase the number of particles as much as possible and keep it fixed. Also, in some hardware implementations, the number of particles is a design parameter that cannot be modified during implementation. Nevertheless, in many other applications where resources are scarce or are shared with a dynamical allocation and/or with energy restrictions, one might be interested in adapting the number of particles in a smart way. One would use enough particles to achieve a certain performance requirement but without wasting resources by using many more particles if they do not translate into a significant improvement of the filter performance.

The selection of the number of particles, however, is often a delicate subject because, (1) the performance of the filter (the quality of the approximation) cannot usually be described in advance as a function of the number of particles, and (2) the mismatch between the approximation provided by the filter and the unknown posterior distribution is obviously also unknown. Therefore, although there is a clear trade-off between performance and computational cost, this relation is not straightforward; e.g., increasing the number of particles over a certain value may not significantly improve the quality of the approximation while decreasing the number of particles below some other value can dramatically affect the performance of the filter.

Few papers in the wide literature have addressed the problem of online assessment of the filter convergence for the purpose of adapting the number of particles. In [14], the number of particles is selected so that the bound on the approximation error does not exceed a threshold with certain probability. The latter error is defined as the Kullback-Leibler divergence (KLD) between the approximate filter distribution and a grid-discretized version of the true one (which is itself a potentially-costly approximation with an unknown error). In [15], an adaptation of the number of particles is proposed, based on the KLD approach of [14] and an estimate of the variance of the estimators computed via the particle filter, along with an improvement of the proposal distributions. In [16], the adaptation of the number of particles is based on the effective sample size. To our best knowledge, all existing methods are heuristic: they do not enjoy any theoretical guarantees (in the assessment of the approximation errors made by the particle filter) and the allocation of particles, therefore, cannot be ensured to be optimal according to any probabilistic criterion.

In this paper, we introduce a model-independent methodology for the online assessment of the convergence of particle filters and carry out a rigorous analysis that ensures the consistency of the proposed scheme under fairly standard assumptions. The method is an extension of our previous work presented in [17]. In the proposed scheme, the observations are processed one at a time and the filter performance is assessed by measuring the discrepancy between the actual observation at each time step and a number of fictitious data-points drawn from the particle approximation of the predictive probability distribution of the observations. The method can be exploited to adjust the number of particles dynamically when the performance of the filter degrades below a certain desired level. This would allow a practitioner to select the operation point by considering performance-computational cost tradeoffs. Based on the method, we propose a simple and efficient algorithm that adjusts the number of particles in real time. We demonstrate the performance of the algorithm numerically by running it for a stochastic

version of the 3-dimensional Lorenz 63 system.

Let us point out that the adaptive procedure for the online selection of the number of particles described herein is only one of many that can exploit the results of the convergence analysis. In other words, our analysis opens the door for development of new family of algorithms for online adaptation of the number of particles by way of online convergence assessment.

The rest of the paper is organized as follows. In Section II we describe the class of state space Markov models and provide a basic background on the well-known bootstrap particle filter of [11]. The theoretical results that enable the online assessment of particle filters are stated in Section III, with full details and proofs contained in Appendix A. The proposed methodology for online convergence assessment of the particle filter is introduced in Section IV. Furthermore, this section provides a simple algorithm for the dynamic, online adaptation of the number of particles. In Section V, we illustrate the validity of the method by means of computer simulations for a stochastic Lorenz 63 model. Finally, Section VI contains a summary of results and some concluding remarks.

II. PARTICLE FILTERING

In this section we describe the class of state space models of interest and then present the standard particle filter (PF), which is the basic building block for the methods to be introduced later.

A. State space models and stochastic filtering

Let us consider discrete-time, Markov dynamic systems in state-space form described by the triplet¹

$$\mathbf{X}_0 \sim p(\mathbf{x}_0), \quad (1)$$

$$\mathbf{X}_t \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (2)$$

$$\mathbf{Y}_t \sim p(\mathbf{y}_t | \mathbf{x}_t), \quad (3)$$

where

- $t \in \mathbb{N}$ denotes discrete time;
- \mathbf{X}_t is the $d_x \times 1$ -dimensional (random) system state at time t , which takes variables in the set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$,
- $p(\mathbf{x}_0)$ is the a priori pdf of the state, while
- $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ denotes the conditional density of the state \mathbf{X}_t given $\mathbf{X}_{t-1} = \mathbf{x}_{t-1}$;
- \mathbf{Y}_t is the $d_y \times 1$ -dimensional observation vector at time t , which takes values in the set $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and is assumed to be conditionally independent of all other observations given the state \mathbf{X}_t ,
- $p(\mathbf{y}_t | \mathbf{x}_t)$ is the conditional pdf of \mathbf{Y}_t given $\mathbf{X}_t = \mathbf{x}_t$. It is often referred to as the *likelihood* of \mathbf{x}_t , when it is viewed as a function of \mathbf{x}_t given \mathbf{y}_t .

The model described by Eqs. (1)–(3) includes a broad class of systems, both linear and nonlinear, with Gaussian or non-Gaussian perturbations. Here we focus on the case where all the model parameters are known. However, the proposed method can also be used for models with unknown parameters for which suitable particle filtering methods are available [18], [19], [20]. We assume that the prior distribution of the state $p(\mathbf{x}_0)$ is also known.

¹In most of the paper we abide by a simplified notation where $p(x)$ denotes the probability density function (pdf) of the random variables X . This notation is argument-wise, hence if we have two random variables X and Y , then $p(x)$ and $p(y)$ denote the corresponding density functions, possibly different; $p(x, y)$ denotes the joint pdf and $p(x|y)$ is the conditional pdf of X given $Y = y$. A more accurate notation, which avoids any ambiguities, is used for the analysis and the statement of the theoretical results. Vectors are denoted by bold-face letters, e.g., \mathbf{x} , while regular-face is used for scalars, e.g., x .

The stochastic filtering problem consists in the computation of the sequence of posterior probability distributions given by the so-called *filtering* densities $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, $t = 1, 2, \dots$. The pdf $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is closely related to the one-step-ahead predictive state density $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$, which is of major interest in many applications and can be written down by way of the Chapman-Kolmogorov equation,

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}. \quad (4)$$

Using Bayes' theorem together with Eq. (4), we obtain the well-known recursive factorization of the filtering pdf

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}.$$

For conciseness and notational accuracy, we use the measure-theoretic notation

$$\pi_t(d\mathbf{x}_t) := p(\mathbf{x}_t|\mathbf{y}_{1:t})d\mathbf{x}_t, \quad \xi_t(d\mathbf{x}_t) := p(\mathbf{x}_t|\mathbf{y}_{1:t-1})d\mathbf{x}_t$$

to represent the filtering and the predictive posterior probability distributions of the state, respectively. Note that π_t and ξ_t are probability measures, hence, given a Borel set $A \subset \mathcal{X}$, $\pi_t(A) = \int_A \pi_t(d\mathbf{x}_t)$ and $\xi_t(A) = \int_A \xi_t(d\mathbf{x}_t)$ denote the posterior probability of the event $\mathbf{X}_t \in A$ conditional on $\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}$ and $\mathbf{Y}_{1:t-1} = \mathbf{y}_{1:t-1}$, respectively.

However, the object of main interest for the convergence assessment method to be introduced in this paper is the predictive pdf of the observations, namely the function $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ and the associated probability measure

$$\mu_t(d\mathbf{y}_t) := p(\mathbf{y}_t|\mathbf{y}_{1:t-1})d\mathbf{y}_t.$$

The density $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ is the normalization constant of the filtering density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, and it is related to the predictive state pdf $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ through the integral

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})d\mathbf{x}_t. \quad (5)$$

It also plays a key role in model assessment [17] and model inference problems [19], [20], [21].

B. The standard particle filter

A PF is an algorithm that processes the observations $\{\mathbf{y}_t\}_{t \geq 1}$ sequentially in order to compute Monte Carlo approximations of the sequence of probability measures $\{\pi_t\}_{t \geq 1}$. The simplest algorithm is the so-called *bootstrap* filter (BF) [11] (see also [22]), which consists of a recursive importance sampling procedure and a resampling step. The term ‘‘particle’’ refers to a Monte Carlo sample in the state space \mathcal{X} , which is assigned an importance weight. Below, we outline the BF algorithm with M particles.

Algorithm 1. Bootstrap filter.

- 1) Initialization. At time $t = 0$, draw M i.i.d. samples, $\mathbf{x}_0^{(m)}$, $m = 1, \dots, M$, from the prior $p(\mathbf{x}_0)$.
- 2) Recursive step. Let $\{\mathbf{x}_{t-1}^{(m)}\}_{m=1}^M$ be the particles at time $t - 1$. At time t , proceed with the two steps below.
 - a) For $m = 1, \dots, M$, draw $\bar{\mathbf{x}}_t^{(m)}$ from the model transition pdf $p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(m)})$. Then compute the normalized importance weights

$$w_t^{(m)} = \frac{p(\mathbf{y}_t|\bar{\mathbf{x}}_t^{(m)})}{\sum_{k=1}^M p(\mathbf{y}_t|\bar{\mathbf{x}}_t^{(k)})}, \quad m = 1, \dots, M. \quad (6)$$

- b) Resample M times with replacement: for $m = 1, \dots, M$, let $x_t^{(m)} = \bar{x}_t^{(k)}$ with probability $w_t^{(k)}$, where $k \in \{1, \dots, M\}$.

For the sake of simplicity, in step 2.(b) above we assume that multinomial resampling [7] is carried out for every $t \geq 1$. The results and methods to be presented in subsequent sections remain valid when resampling is carried out periodically and/or using alternative schemes such as residual [6], stratified [23] or minimum-variance [24] resampling (see also [25]).

The simple BF yields several useful approximations. After sampling at step 2.(a), the predictive state probability measure ξ_t can be approximated as

$$\xi_t^M(dx_t) = \frac{1}{M} \sum_{m=1}^M \delta_{\bar{x}_t^{(m)}}(dx_t),$$

where δ_x denotes the Dirac delta measure located at $x \in \mathcal{X}$. The filter measure π_t can be similarly approximated, either using the particles and weights computed at step 2.(a) or the resampled particles after step 2.(b), i.e.,

$$\bar{\pi}_t^M = \sum_{m=1}^M w_t^{(m)} \delta_{\bar{x}_t^{(m)}} \quad \text{and} \quad \pi_t^M = \frac{1}{M} \sum_{m=1}^M \delta_{x_t^{(m)}},$$

respectively. In addition, the BF yields natural approximations of the predictive pdf's of \mathbf{X}_t and \mathbf{Y}_t given the earlier observations $\mathbf{Y}_{1:t-1} = \mathbf{y}_{1:t-1}$. If we specifically denote these functions as $\tilde{p}_t(\mathbf{x}_t) := p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$ and $p_t(\mathbf{y}_t) := p(\mathbf{y}_t | \mathbf{y}_{1:t-1})$, then we readily obtain their respective estimates as mixture distributions with M mixands, or,

$$\begin{aligned} \tilde{p}_t^M(\mathbf{x}_t) &:= \sum_{m=1}^M w_{t-1}^M p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(m)}), \quad \text{and} \\ p_t^M(\mathbf{y}_t) &:= \frac{1}{M} \sum_{m=1}^M p(\mathbf{y}_t | \bar{x}_t^{(m)}), \end{aligned}$$

for any $\mathbf{x}_t \in \mathcal{X}$ and $\mathbf{y}_t \in \mathcal{Y}$.

III. A NOVEL ASYMPTOTIC CONVERGENCE RESULT

The convergence of the approximate measures, e.g., ξ_t^M , towards the true ones is usually assessed in terms of the estimates of 1-dimensional statistics of the corresponding probability distribution. To be specific, let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a real integrable function in the state space and denote²

$$(f, \xi_t) := \int f(\mathbf{x}_t) \xi_t(dx_t).$$

Under mild assumptions on the state space model, it can be proved that

$$\lim_{M \rightarrow \infty} (f, \xi_t^M) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M f(\mathbf{x}_t^{(m)}) = (f, \xi_t) \quad (7)$$

almost surely (a.s.) [26], [12].

According to (5), the predictive observation pdf $p_t(\mathbf{y}_t)$ is an integral w.r.t. ξ_t and, as a consequence, Eq. (7) implies that $\lim_{M \rightarrow \infty} p_t^M(\mathbf{y}) = p_t(\mathbf{y})$ a.s. and point-wise for every $\mathbf{y} \in \mathcal{Y}$ under mild assumptions [26]. However, existing theoretical results *do not* ensure that $p_t^M(\mathbf{y})$ can converge *uniformly* on \mathcal{Y} towards $p_t(\mathbf{y})$ and this fact prevents us from claiming that $\lim_{M \rightarrow \infty} \int h(\mathbf{y}) p_t^M(\mathbf{y}) d\mathbf{y} = \int h(\mathbf{y}) p_t(\mathbf{y}) d\mathbf{y} = (h, \mu_t)$ in some proper sense for integrable real functions $h(\mathbf{y})$.

The first contribution of this paper is to prove that, under mild regularity assumptions on the state space model, the continuous random probability measure

$$\mu_t^M(d\mathbf{y}) := p_t^M(\mathbf{y}) d\mathbf{y}$$

converges a.s. to μ_t and provide explicit error rates. To express this result rigorously, we need to introduce some notation:

²Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be a measurable space, where $\mathcal{Z} \subset \mathbb{R}^d$ for some integer $d \geq 1$ and $\mathcal{B}(\mathcal{Z})$ is the Borel σ -algebra of subsets of \mathcal{Z} . If α is a measure on $\mathcal{B}(\mathcal{Z})$ and the function $h : \mathcal{Z} \rightarrow \mathbb{R}$ is integrable with respect to (w.r.t.) α , then we use the shorthand notation $(f, \alpha) := \int f(z) \alpha(dz)$.

- For each $t \geq 1$, let us define the function $g_t(\mathbf{y}_t, \mathbf{x}_t) := p(\mathbf{y}_t | \mathbf{x}_t)$, i.e., the conditional pdf of \mathbf{y}_t given \mathbf{x}_t . When this function is used as a likelihood, we write $g_t^{y_t}(\mathbf{x}_t) := g_t(\mathbf{y}_t, \mathbf{x}_t)$ to emphasize that it is a function of \mathbf{x}_t .
- Let $f : \mathcal{Z} \rightarrow \mathbb{R}$ be a real function on some set \mathcal{Z} . We denote the absolute supremum of f as $\|f\|_\infty := \sup_{z \in \mathcal{Z}} |f(z)|$. The set of bounded real functions on \mathcal{Z} is $B(\mathcal{Z}) := \{f : \mathcal{Z} \rightarrow \mathbb{R} \text{ such that } \|f\|_\infty < \infty\}$.
- Let $\mathbf{a} = (a_1, \dots, a_d)$ be a multi-index, where each $a_i, i = 1, 2, \dots, d$, is a non-negative integer. Let $f : \mathcal{Z} \rightarrow \mathbb{R}$ be a real function on a d -dimensional set $\mathcal{Z} \subseteq \mathbb{R}^d$. We use $D_z^{\mathbf{a}} f(\mathbf{z})$ to denote the partial derivative of f w.r.t. the variable \mathbf{z} determined by the entries of \mathbf{a} , namely,

$$D_z^{\mathbf{a}} f(\mathbf{z}) = \frac{\partial^{a_1} \dots \partial^{a_d} f}{\partial z_1^{a_1} \dots \partial z_d^{a_d}}(\mathbf{z}).$$

The order of the derivative operator $D_z^{\mathbf{a}}$ is $|\mathbf{a}| = \sum_{i=1}^d a_i$.

- The minimum out of two scalar quantities, $a, b \in \mathbb{R}$, is denoted $a \wedge b$.

We make the following assumptions on the likelihood function g_t and the predictive observation measure $\mu_t(d\mathbf{y}_t) = p_t(\mathbf{y}_t)d\mathbf{y}_t$.

- (\mathfrak{L}) For each $t \geq 1$, the function g_t is positive and bounded, i.e., $g_t(\mathbf{y}, \mathbf{x}) > 0$ for any $(\mathbf{y}, \mathbf{x}) \in \mathcal{Y} \times \mathcal{X}$ and $\|g_t\|_\infty = \sup_{(\mathbf{y}, \mathbf{x}) \in \mathcal{Y} \times \mathcal{X}} |g_t(\mathbf{y}, \mathbf{x})| < \infty$.
- (\mathfrak{D}) For each $t \geq 1$, the function $g_t(\mathbf{y}, \mathbf{x})$ is differentiable with respect to \mathbf{y} , with bounded derivatives up to order d_y , i.e., $D_y^{\mathbf{1}} g_t(\mathbf{y}, \mathbf{x}) = \frac{\partial^{d_y} g_t}{\partial y_1 \dots \partial y_{d_y}}(\mathbf{y}, \mathbf{x})$ exists and

$$\|D_y^{\mathbf{1}} g_t\|_\infty = \sup_{(\mathbf{y}, \mathbf{x}) \in \mathcal{Y} \times \mathcal{X}} |D_y^{\mathbf{1}} g_t(\mathbf{y}, \mathbf{x})| < \infty.$$

- (\mathfrak{C}) For any $0 < \beta < 1$ and any $p \geq 4$, the sequence of hypercubes

$$C_M := \left[-\frac{M^{\frac{\beta}{p}}}{2}, +\frac{M^{\frac{\beta}{p}}}{2} \right] \times \dots \times \left[-\frac{M^{\frac{\beta}{p}}}{2}, +\frac{M^{\frac{\beta}{p}}}{2} \right] \subset \mathbb{R}^{d_y}$$

satisfies the inequality $\mu_t(\overline{C_M}) \leq bM^{-\eta}$ for some constants $b > 0$ and $\eta > 0$ independent of M (yet possibly dependent on β and p), where $\overline{C_M} = \mathbb{R}^{d_y} \setminus C_M$ is the complement of C_M .

Remark 1. Assumptions (\mathfrak{L}) and (\mathfrak{D}) refer to regularity conditions (differentiability and boundedness) that the likelihood function of the state space model should satisfy. Models of observations, for example, of the form $\mathbf{y}_t = f(\mathbf{x}_t) + \mathbf{u}_t$, where f is a (possibly nonlinear) transformation of the state \mathbf{x}_t and \mathbf{u}_t is noise with some differentiable, exponential-type pdf (e.g., Gaussian or mixture-Gaussian), readily satisfy these assumptions. Typical two-sided heavy-tailed distributions, such as Student's t distribution, also satisfy (\mathfrak{L}) and (\mathfrak{D}).

Assumption (\mathfrak{C}) states that the tails of the pdf $p_t(\mathbf{y}_t) = p(\mathbf{y}_t | \mathbf{y}_{1:t-1})$ should not be too heavy. Being polynomial on M , this constraint is relatively weak and the assumption is satisfied for all exponential-type distributions as well as for many heavy-tailed distributions. For example, when $d_y = 1$, one can choose the constants b and η such that $bM^{-\eta}$ is an upper bound for the tails of the (heavy-tailed) Pareto, Weibull, Burr or Levy distributions.

Theorem 1. Assume that (\mathfrak{L}), (\mathfrak{D}) and (\mathfrak{C}) hold and the observations $\mathbf{y}_{1:t-1}$ are fixed (and otherwise arbitrary). Then, for every $h \in B(\mathcal{Y})$ and any $\epsilon \in (0, \frac{1}{2})$ there exists an a.s. finite r.v. W_t^ϵ , independent of M , such that

$$|(h, \mu_t^M) - (h, \mu_t)| \leq \frac{W_t^\epsilon}{M^{(\frac{1}{2}-\epsilon) \wedge \eta}}.$$

In particular,

$$\lim_{M \rightarrow \infty} (h, \mu_t^M) = (h, \mu_t) \quad \text{a.s.}$$

See Appendix A for a proof.

IV. ONLINE SELECTION OF THE NUMBER OF PARTICLES

In the sequel we assume scalar observations, hence $d_y = 1$ and $\mathbf{y}_t = y_t$ (while $d_x \geq 1$ is arbitrary). A discussion of how to proceed when $d_y > 1$ is provided in Section IV-E.

Our goal is to evaluate the convergence of the BF (namely, the accuracy of the approximation $p_t^M(y_t)$) in real time and, based on the convergence assessment, adapt the computational effort of the algorithm, i.e., the number of used particles M .

To that end, we run the BF in the usual way with a light addition of computations. At each iteration we generate K ‘‘fictitious observations’’, denoted $\tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$, from the approximate predictive pdf $p_t^M(y_t)$. If the BF is operating with a small enough level of error, then Theorem 1 states that these fictitious observations come approximately from the same distribution as the acquired observation, i.e., $\mu_t^M(dy_t) \approx \mu_t(dy_t)$. In that case, as we explain in Subsection IV-B, a statistic a_t^K can be constructed using $y_t, \tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$, which necessarily has an (approximately) uniform distribution independently of the specific form of the state-space model (1)–(3). By collecting a sequence of such statistics, say $a_{t-W+1}^K, \dots, a_t^K$ for some window size W , one can easily test whether their empirical distribution is close to uniform using standard procedures. The better the approximation $\mu_t^M \approx \mu_t$ generated by the BF, the better fit with the uniform distribution can be expected.

If $K \ll M$ and W is not too large, the cost of the added computations is negligible compared to the cost of running the BF with M particles and, as we numerically show in Section V, the ability to adapt the number of particles online leads to a very significant reduction of the running times without compromising the estimation accuracy.

Below we describe the method, justify its theoretical validity and discuss its computational complexity as well as its extension to the case of multidimensional \mathbf{y}_t 's.

A. Generation of fictitious observations

The proposed method demands at each time t the generation of K fictitious observations (i.e., Monte Carlo samples), denoted $\{\tilde{y}_t^{(k)}\}_{k=1}^K$, from the approximate predictive observation pdf $p_t^M(y_t) = \frac{1}{M} \sum_{m=1}^M p(y_t | \bar{\mathbf{x}}_t^{(m)})$. Since the latter density is a finite mixture, drawing from $p_t^M(y_t)$ is straightforward as long as the conditional density of the observations, $p(y_t | \mathbf{x}_t)$, is itself amenable to sampling. In order to generate $\tilde{y}_t^{(k)}$, it is enough to draw a sample $j^{(k)}$ from the discrete uniform distribution on $\{1, 2, \dots, M\}$ and then generate $\tilde{y}_t^{(k)} \sim p(y_t | \bar{\mathbf{x}}_t^{(j^{(k)})})$.

B. Assessing convergence via invariant statistics

For simplicity, let us assume first that $p_t^M(y_t) = p_t(y_t) = p(y_t | y_{1:t-1})$, i.e., there is no approximation error and, therefore, the fictitious observations $\{\tilde{y}_t^{(k)}\}_{k=1}^K$ have the same distribution as the *true* observation y_t . We define the set $\mathcal{A}_{K,t} := \{y \in \{\tilde{y}_t^{(k)}\}_{k=1}^K : y < y_t\}$ and the r.v. $A_{K,t} := |\mathcal{A}_{K,t}| \in \{0, 1, \dots, K\}$. Note that $\mathcal{A}_{K,t}$ is the set of fictitious observations which are smaller than the actual one, while $A_{K,t}$ is the number of such observations. If we let \mathbb{Q}_K denote the probability mass function (pmf) of A_K , it is not hard to show that \mathbb{Q}_K is uniform independently of the value and distribution of y_t . This is rigorously given by the Proposition below.

Proposition 1. *If $y_t, \tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$ are i.i.d. samples from a common continuous (but otherwise arbitrary) probability distribution, then the pmf of the r.v. $A_{K,t}$ is*

$$\mathbb{Q}_K(n) = \frac{1}{K+1}, \quad n = 0, \dots, K. \quad (8)$$

Proof: Since $y_t, \tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$ are i.i.d., all possible orderings of the $K+1$ samples are a priori equally probable, and the value of the r.v. $A_{K,t}$ depends uniquely on the relative position of y_t after the samples are sorted (e.g., if y_t is the

smallest sample, then $A_{K,t} = 0$, if there is exactly one $\tilde{y}_t^{(i)} < y_t$ then $A_{K,t} = 1$, etc.). There are $(K + 1)!$ different ways in which the samples $y_t, \tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$ can be ordered, but $A_{K,t}$ can only take values from 0 to K . In particular, given the relative position of y_t , there are $K!$ different ways in which the remaining samples $\tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$ can be arranged. Therefore, $\mathbb{Q}_K(A_K = n) = \frac{K!}{(K+1)!} = \frac{1}{K+1}$ for every $n \in \{0, 1, \dots, K\}$. \square

In practice, $p_t^M(y_t)$ is just an approximation of the predictive observation pdf $p_t(y_t)$ and, therefore, the actual and fictitious observations are not i.i.d. However, under the assumptions of Theorem 1, the a.s. convergence of the approximate measure $\mu_t^M(dy_t) = p_t^M(y_t)dy_t$ enables us to obtain an ‘‘approximate version’’ of the uniform distribution in Proposition 1, with the error vanishing as $M \rightarrow \infty$. To be specific, we introduce the set $\mathcal{A}_{K,M,t} := \{y \in \{\tilde{y}_t^{(k)}\}_{k=1}^K : y < y_t\}$, which depends on M because of the mismatch between $p_t^M(y_t)$ and $p_t(y_t)$, and the associated r.v. $A_{K,M,t} = |\mathcal{A}_{K,M,t}|$ with pmf $\mathbb{Q}_{K,M,t}$. We have the following convergence result for $\mathbb{Q}_{K,M,t}$.

Theorem 2. *Let y_t be a sample from $p_t(y_t)$ and let $\{\tilde{y}_t^{(k)}\}_{k=1}^K$ be i.i.d. samples from $p_t^M(y_t)$. If the observations $y_{1:t-1}$ are fixed and Assumptions (\mathfrak{L}) , (\mathfrak{D}) and (\mathfrak{C}) hold, then there exists a sequence of non-negative r.v.'s $\{\varepsilon_t^M\}_{M \in \mathbb{N}}$ such that $\lim_{M \rightarrow \infty} \varepsilon_t^M = 0$ a.s. and*

$$\frac{1}{K+1} - \varepsilon_t^M \leq \mathbb{Q}_{K,M,t}(n) \leq \frac{1}{K+1} + \varepsilon_t^M. \quad (9)$$

In particular, $\lim_{M \rightarrow \infty} \mathbb{Q}_{K,M,t}(n) = \mathbb{Q}_K(n) = \frac{1}{K+1}$ a.s.

See Appendix B for a proof. Proposition 1 states that the statistic $A_{K,t}$ is distribution-invariant, since $\mathbb{Q}_K(n) = \frac{1}{K+1}$ independently of t and the state space model. Similarly, Theorem 2 implies that the statistic $A_{K,M,t}$ is asymptotically distribution-invariant (independently of t and the model) since $\mathbb{Q}_{K,M,t}(n) \rightarrow \frac{1}{K+1}$ when $M \rightarrow \infty$, as the BF converges.³

C. BF algorithm with adaptive number of particles

We propose an algorithm that dynamically adjusts the number of particles of the filter based on the transformed r.v. $A_{K,M,t}$. Table II summarizes the proposed algorithm, that is embedded into a standard BF (see Section II-B) but can be applied to virtually any other particle filter in a straightforward manner. The parameters of the algorithm are shown in Table I.

The BF is initialized in Step 1(a) with M_0 initial particles. At each recursion, in Step 2(a), the filtered distribution of the current state is approximated. In Step 2(b), K fictitious observations $\{\tilde{y}_t^{(k)}\}_{k=1}^K$ are drawn and the statistic $A_{K,M,t} = a_{K,M,t}$ is computed. In Step 2(b), once a set of W consecutive statistics have been acquired, $\mathcal{S}_t = \{a_{K,M,t-W+1}, a_{K,M,t-W+2}, \dots, a_{K,M,t-1}, a_{K,M,t}\}$, a statistical test is performed for checking whether \mathcal{S}_t is a sequence of i.i.d. samples from the pmf given by Eq. (8).

There are several approaches that can be used to exploit the information contained in \mathcal{S}_t . Here we perform a Pearson's chi-squared test [27], where the χ_t^2 statistic is computed according to Eq. (10) (see Table II). Then, a p-value $p_{K,t}^*$ for testing the hypothesis that the empirical distribution of \mathcal{S}_t is uniform is computed. The value $p_{K,t}^*$ is obtained by comparing the χ_t^2 statistic with the χ^2 distribution with K degrees of freedom. Intuitively, a large $p_{K,t}^*$ suggests a good match of the sequence \mathcal{S}_t with an i.i.d. sample from the uniform distribution on $\{0, 1, \dots, K\}$, while a small $p_{K,t}^*$ indicates a mismatch. Therefore, the p-value $p_{K,t}^*$ is compared with two different significance levels: a low threshold p_ℓ and a high threshold p_h . If $p_{K,t}^* \leq p_\ell$, the number of particles is increased according to the rule $M_t = f_{\text{up}}(M_{t-1})$ whereas, if $p_{K,t}^* \geq p_h$, the number of particles is decreased according to the rule $M_t = f_{\text{down}}(M_{t-1})$. When $p_\ell < p_{K,t}^* < p_h$, the number of particles remains fixed. These

³Specifically note that, under assumptions (\mathfrak{L}) , (\mathfrak{D}) and (\mathfrak{C}) , the convergence of the continuous random measure μ_t^M computed via the BF (which is sufficient to obtain (9); see Appendix B) is guaranteed by Theorem 1.

TABLE I: Parameters of the algorithm

- M_0 , initial number of particles
- M_{\min} , minimum number of particles
- M_{\max} , maximum number of particles
- K , number of fictitious samples per iteration
- W , window length
- p_ℓ , lower significance level of p-values
- p_h , higher significance level of p-values
- $f_{\text{up}}(\cdot)$, rule for increasing M
- $f_{\text{down}}(\cdot)$, rule for decreasing M

two significance levels allow the practitioner to select the operation range by considering a performance-to-computational-cost tradeoff. Note that we set M_{\min} and M_{\max} , maximum and minimum values for the number of particles, respectively.

A large window W yields a more accurate convergence assessment but increases the latency (or decreases the responsiveness) of the algorithm. If the algorithm must be run online, this latency can be critical for detecting a malfunction of the filter and adapting consequently the number of particles. Therefore there is a tradeoff between the accuracy of the convergence assessment procedure and latency of the algorithm.

D. Computational cost

Compared to the BF, the additional computational cost of the method is mainly driven by the generation of the K fictitious observations at each iteration as shown in Subsection IV-A. The generation of these fictitious observations is a two-step procedure, where in the first step, we draw K discrete indices, say j_1, \dots, j_K , from the set $\{1, \dots, M_n\}$ with uniform probabilities, and in the second step, we draw K samples from $p(y_t | \bar{\mathbf{x}}_t^{(j_1)}), \dots, p(y_t | \bar{\mathbf{x}}_t^{(j_K)})$, respectively.

In the proposed algorithm, a Pearson's χ^2 test is performed with a sequence \mathcal{S}_t of W samples, that is, it is carried out only once every W consecutive time steps. Therefore, the computational cost will depend on the parameters K and W . We will show in Section V that the algorithm can work very well with a low number of fictitious observations, which imposes a very light extra computational load.

E. Multidimensional observations

Through this section, we have assumed scalar observations. In the multidimensional case, with $\mathbf{y}_t = [y_{1,t}, \dots, y_{d_y,t}]^\top$, the same assessment scheme can be applied over each marginal $p(y_{i,t} | \mathbf{y}_{1:t-1})$ of the predictive observation pdf. Theoretical guarantees readily follow from the convergence of the marginal measures $\mu_{i,t}^M(dy_{i,t}) = p^M(y_{i,t} | \mathbf{y}_{1:t-1}) dy_{i,t}$ under the same assumptions as the joint measure μ_t^M (see Appendix A).

Note that the convergence of the marginals does not imply the convergence of the joint approximation μ_t^M . However, it can be reasonably expected that when all marginals are approximated well over a period of time, the joint distribution is accurately approximated as well.

TABLE II: Algorithm for adapting the number of particles

1) **[Initialization]**

a) Initialize the particles and the weights of the filter as

$$\begin{aligned} \mathbf{x}_0^{(m)} &\sim p(\mathbf{x}_0), & m = 1, \dots, M_0, \\ w_0^{(m)} &= 1/M_0, & m = 1, \dots, M_0, \end{aligned}$$

and set $n = 1$.

2) **[For $t = 1 : T$]**a) **Bootstrap particle filter:**

- Resample M_n samples of $\bar{\mathbf{x}}_{t-1}^{(m)}$ with weights $w_{t-1}^{(m)}$ to obtain $\mathbf{x}_{t-1}^{(m)}$.
- Propagate $\bar{\mathbf{x}}_t^{(m)} \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(m)})$, $m = 1, \dots, M_n$.
- Compute the non-normalized weights $\bar{w}_t^{(m)} = p(y_t | \bar{\mathbf{x}}_t^{(m)})$, $m = 1, \dots, M_n$.
- Normalize the weights $\bar{w}_t^{(m)}$ to obtain $w_t^{(m)}$, $m = 1, \dots, M_n$.

b) **Fictitious observations:**

- Draw $\tilde{y}_t^{(k)} \sim p^M(y_t | y_{t-1})$, $k = 1, \dots, K$.
- Compute $a_{K,M,t} = A_{K,M}$, i.e., the position of y_t within the set of ordered fictitious observations $\{\tilde{y}_t^{(k)}\}_{k=1}^K$.

c) If $t = nW$, (**assessment of convergence**):

- Compute the χ_t^2 statistic over the empirical distribution of $\mathcal{S}_t = \{a_{K,M,t}, a_{K,M,t-1}, \dots, a_{K,M,t-W+1}\}$ as

$$\chi_t^2 = \sum_{j=0}^K \frac{(O_j - E_j)^2}{E_j}, \quad (10)$$

where O_j is the frequency of the observations in the window being in the j th relative position, i.e., $O_j = |\{a_{K,M,\tau} \in \mathcal{S}_t : a_{K,M,\tau} = j\}|$, and E_j is the expected frequency under the null hypothesis, i.e., $E_j = W \cdot \mathbb{Q}_K(j) = \frac{W}{K+1}$ (see Eq. (8)).

- Calculate the p-value $p_{K,t}^*$ by comparing the statistic χ_t^2 to the χ^2 -distribution with K degrees of freedom.

- If $p_{K,t}^* \leq p_\ell$

increase $M_n = \min\{f_{\text{up}}(M_{n-1}), M_{\text{max}}\}$.

- Else, if $p_{K,t}^* \geq p_h$,

decrease $M_n = \max\{f_{\text{down}}(M_{n-1}), M_{\text{min}}\}$.

- Else,

$M_n = M_{n-1}$.

- Set $n = n + 1$.

d) If $t < nW$, set $t = t + 1$ and go to 2. Otherwise, end.

V. NUMERICAL EXAMPLE

A. The three-dimensional Lorenz system

In this section we show computer simulation results that demonstrate the performance of the proposed method. We consider the problem of tracking the state of a three-dimensional Lorenz system [28] with additive dynamical noise, partial observations and additive measurement noise [29]. Namely, we consider a three-dimensional stochastic process $\{\mathbf{X}(s)\}_{s \in (0, \infty)}$ taking values on \mathbb{R}^3 , whose dynamics are described by the system of stochastic differential equations

$$\begin{aligned} dX_1 &= -s(X_1 - Y_1) + dW_1, \\ dX_2 &= rX_1 - X_2 - X_1X_3 + dW_2, \\ dX_3 &= X_1X_2 - bX_3 + dW_3, \end{aligned}$$

where $\{W_i(s)\}_{s \in (0, \infty)}$, $i = 1, 2, 3$, are independent one-dimensional Wiener processes and

$$(s, r, b) = \left(10, 28, \frac{8}{3}\right)$$

are static model parameters broadly used in the literature since they lead to a chaotic behavior [28]. Here we use a discrete-time version of the latter system using an Euler-Maruyama scheme with integration step $\Delta = 10^{-3}$, which yields the model

$$X_{1,n} = X_{1,n-1} - \Delta s(X_{1,n-1} - X_{2,n-1}) + \sqrt{\Delta} U_{1,n}, \quad (11)$$

$$X_{2,n} = X_{2,n-1} + \Delta(rX_{1,n-1} - X_{2,n-1} - X_{1,n-1}X_{3,n-1}) + \sqrt{\Delta} U_{2,n}, \quad (12)$$

$$X_{3,n} = X_{3,n-1} + \Delta(X_{1,n-1}X_{2,n-1} - bX_{3,n-1}) + \sqrt{\Delta} U_{3,n}, \quad (13)$$

where $\{U_{i,n}\}_{n=0,1,\dots}$, $i = 1, 2, 3$, are independent sequences of i.i.d. normal random variables with zero mean and unit variance. The system (11)-(13) is partially observed every 200 discrete-time steps. Specifically, we collect a sequence of scalar observations $\{Y_t\}_{t=1,2,\dots}$, of the form

$$Y_t = X_{1,200t} + V_t, \quad (14)$$

where the observation noise $\{V_t\}_{t=1,2,\dots}$ is a sequence of i.i.d. normal random variables with zero mean and variance $\sigma^2 = \frac{1}{2}$.

Let $\mathbf{X}_n = (X_{1,n}, X_{2,n}, X_{3,n}) \in \mathbb{R}^3$ be the state vector. The dynamic model given by Eqs. (11)–(13) defines the transition kernel $p(\mathbf{x}_n | \mathbf{x}_{n-1})$ and the observation model of Eq. (14) is the likelihood function

$$p(y_t | x_{1,200t}) \propto \exp \left\{ -\frac{1}{2\sigma^2} (y_t - x_{1,200t})^2 \right\}.$$

The goal is on tracking the sequence of joint posterior probability measures π_t , $t = 1, 2, \dots$, for $\{\hat{\mathbf{X}}_t\}_{t=1,\dots}$, where $\hat{\mathbf{X}}_t = \mathbf{X}_{200t}$. Note that one can draw a sample $\hat{\mathbf{X}}_t = \hat{\mathbf{x}}_t$ conditional on $\hat{\mathbf{X}}_{t-1} = \hat{\mathbf{x}}_{t-1}$ by successively simulating

$$\tilde{\mathbf{x}}_n \sim p(\mathbf{x}_n | \tilde{\mathbf{x}}_{n-1}), \quad n = 200(t-1) + 1, \dots, 200t,$$

where $\tilde{\mathbf{x}}_{200(t-1)} = \hat{\mathbf{x}}_{t-1}$ and $\hat{\mathbf{x}}_t = \tilde{\mathbf{x}}_{200t}$. The prior measure for the state variables is normal, namely

$$\mathbf{X}_0 \sim \mathcal{N}(\mathbf{x}_*, v_0^2 \mathcal{I}_3),$$

where $\mathbf{x}_* = (-5.9165; -5.5233; 24.5723)$ is the mean and $v_0^2 \mathcal{I}_3$ is the covariance matrix of \mathbf{X}_0 , with $v_0^2 = 10$ and \mathcal{I}_3 being the three-dimensional identity matrix.

B. Simulation setup

With this example, we aim at showing how the proposed algorithm allows to operate the particle filter with a prescribed performance-to-computational-budget tradeoff. With this purpose, we applied a standard BF for tracking the sequence of posterior probability measures of the system system (11)-(13) generated by the three-dimensional Lorenz model described in Section V-A. We generated a sequence of $T = 2000$ synthetic observations, $\{y_t; t = 1, \dots, 2000\}$, spread over an interval of 400 seconds (in continuous time), corresponding to 4×10^5 discrete time steps in the Euler-Maruyama scheme (hence, one observation every 200 steps). Since the time scale of the discrete time approximation of Eqs. (11)–(13) is $n = 200t$, a resampling step is taken every 200 steps of the underlying discrete-time system.

We started running the PF with a sufficiently large number of particles, namely $N = 5000$, and then let the proposed algorithm decrease the number of particles to attain a prescribed point in the performance-to-computation-cost range. This point is controlled by the operation range of the p-value, which is in turn driven by the pair of significance

levels $[p_\ell - p_h]$. We tested the algorithm for different ranges of p-values, namely, $p_\ell \in \{0.5, 0.4, 0.3, 0.2, 0.1, 0.05\}$ and $p_h \in \{0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$. When the p-value is below p_ℓ , the algorithm doubles the number of particles $M_{n+1} = f_{\text{up}}(M_n) = 2M_n$, and when the p-value is over p_h , the number of particles is halved, $M_{n+1} = f_{\text{down}}(M_n) = M_n/2$. We used $K = 7$ fictitious observations and a window of size $W = 20$.

In order to assess the approximation errors, we computed the empirical MSEs of the approximation of the posterior mean, $E[\hat{\mathbf{X}}_t | Y_{1:t} = y_{1:t}]$, by averaging the MSEs for the whole sequences. Note that, since the actual expectation cannot be computed in closed form for this system, we used the true underlying sequence $\{\mathbf{X}_{200t}\}_{t=1,2,\dots}$ as the ground truth.

C. Numerical results

Table III shows results of the MSE of the approximation of the posterior mean, the average number of particles

$$\bar{M} = \frac{2}{T} \sum_{k=\frac{T}{2}+1}^T M_k, \quad (15)$$

the p-values of the χ^2 test, and the Hellinger distance [30] between the empirical distribution of \mathcal{S}_t and the uniform distribution. They were obtained by averaging over 100 runs and averaging over time steps for each run. The initial number of particles $M_0 = 2^{15}$, and the minimum and maximum number of particles are $M_{\min} = 2^5$ and $M_{\max} = 2^{15}$, respectively. The first half of time steps were discarded for obtaining the displayed results in order to test the behavior of the algorithm for different sets of parameters (see Eq. (15)). Regarding the relation between the MSE and \bar{M} and the p-values, it can be seen that selecting a high operation range yields good performance (low MSE) at the cost of using a large number of particles (high \bar{M}). When we decrease the range of p-values, the algorithm decreases the number of particles, increasing also the approximation error. Table III shows that this conclusion holds for any pair of $[p_\ell - p_h]$.

Figure 1 shows the MSE, the number of particles \bar{M} , and the execution time for the different operation ranges (solid blue line) compared to the particle filter with a fixed number of particles $M = 2^{15}$ (dashed red line). It can be seen that with a moderate operation range ($[p_\ell - p_h] = [0.3 - 0.7]$), the algorithm can perform (in terms of MSE) similarly to the case with fixed M , while reducing the execution time approximately by a factor of four. The execution time can be further reduced by decreasing the operation range, although this worsens the performance.

Figure 2 displays the evolution of the number of particles over time (averaged over 100 runs) for $[p_\ell - p_h] = [0.3 - 0.7]$ both when $M_0 = 5000$ and $M_0 = 10$. In this case, the minimum and maximum number of particles are $M_{\min} = 10$ and $M_{\max} = 5000$, respectively. We see that, after some time, the number of particles adjusted by the algorithm does not depend on M_0 .

Figure 3 shows the same behavior for $[p_\ell - p_h] = [0.2 - 0.6]$. After some time, the filter uses less particles than the filter with results in Fig. 2 because the selected range of thresholds employs smaller p-values.

Figure 4 shows histograms of averaged MSE and \bar{M} for simulations performed with two different sets of thresholds: $[p_\ell - p_h] = [0.3 - 0.5]$ and $[p_\ell - p_h] = [0.5 - 0.7]$. In both cases, the initial number of particles is $M_0 = 5000$. It can be seen that a more demanding pair of thresholds ($[p_\ell - p_h] = [0.5 - 0.7]$) leads to better performance and a larger average number of particles. This behavior can also be seen in Figure 5, where the MSE w.r.t. the number of particles is displayed for three different sets of thresholds. Note that a filter with a too relaxed set of thresholds ($[p_\ell - p_h] = [0.05 - 0.4]$) uses very few particles but obtains a poor performance, while a filter with the most stringent set of thresholds ($[p_\ell - p_h] = [0.5 - 0.9]$) consistently yields a low MSE, at the expense of using a larger number of particles.

| $[p_l - p_h]$ | Fixed $M = 2^{15}$ | [0.4 - 0.8] | [0.35 - 0.7] | [0.3 - 0.7] | [0.25 - 0.65] | [0.2 - 0.6] |
|----------------|--------------------|-------------|--------------|-------------|---------------|-------------|
| MSE | 1.5193 | 1.5234 | 1.5240 | 1.5287 | 3.7552 | 4.6540 |
| \bar{M} | 32768 | 24951 | 14840 | 8729 | 2197 | 451 |
| p-val | 0.5108 | 0.5089 | 0.4902 | 0.4815 | 0.4872 | 0.4785 |
| Hell. distance | 0.2312 | 0.2355 | 0.2493 | 0.2462 | 0.2476 | 0.2521 |
| exec. time (s) | 6201 | 5617 | 3014 | 1532 | 131 | 67 |
| time ratio | 1 | 1.10 | 2.1 | 4.05 | 47.43 | 92.36 |

TABLE III: Lorenz Model (Section V-A): $\Delta = 10^{-3}$, $T_{obs} = 200\Delta$, $\sigma^2 = 0.5$. Algorithm details: $W = 20$, $K = 7$, $M_{max} = 2^{15}$, $M_{min} = 2^7$. MSE in the approximation of the posterior mean, averaged number of particles \bar{M} , averaged p-value, and averaged Hellinger distance.

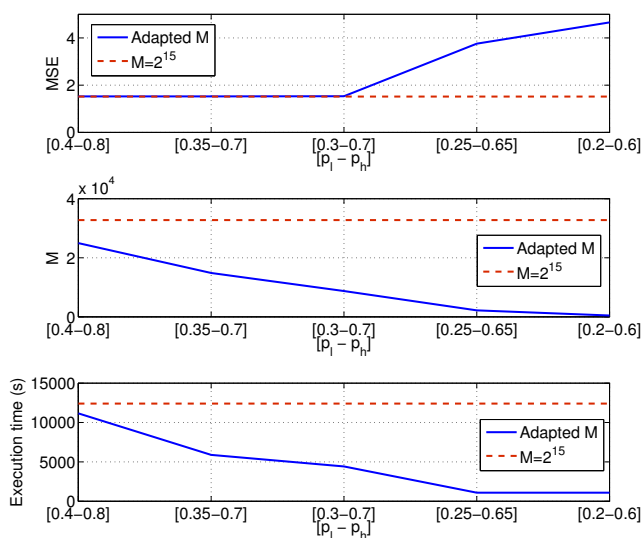


Fig. 1: Lorenz Model (Section V-A). MSE, number of particles M and execution time for different pairs of significance levels $[p_l - p_h]$ in solid blue line, and with a fixed number of particles $M = 2^{15}$ in dashed red line.

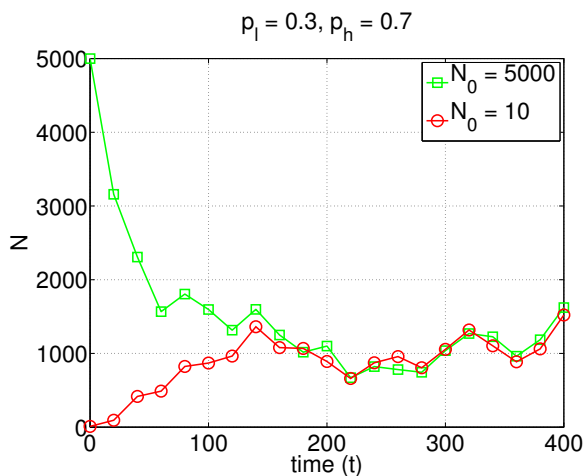


Fig. 2: Lorenz Model (Section V-A). Evolution of the number of particles adapted by the proposed algorithm when the initial number of particles $M_0 \in \{10, 5000\}$. The significance levels were set to $p_l = 0.3$ and $p_h = 0.7$.

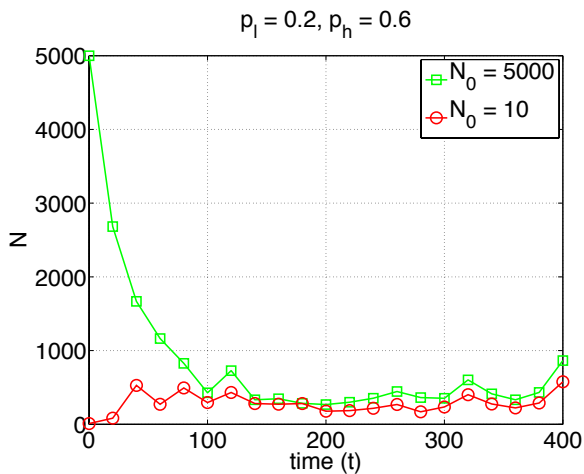


Fig. 3: Lorenz Model (Section V-A). Evolution of the number of particles adapted by the proposed algorithm when the initial number of particles $M_0 \in \{10, 5000\}$. The significance levels were set to $p_\ell = 0.2$ and $p_h = 0.6$.

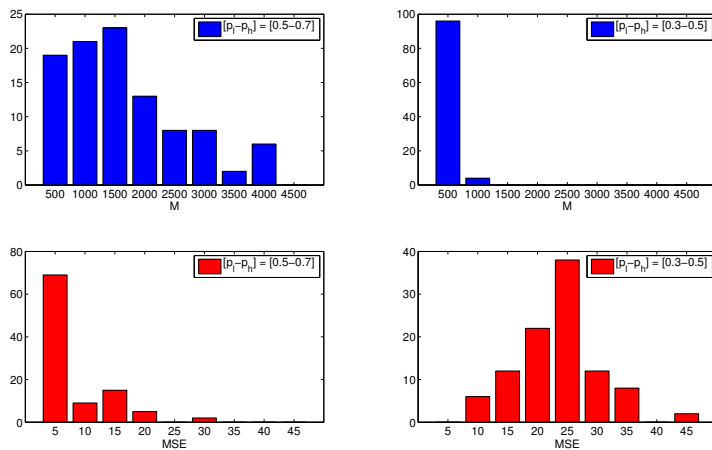


Fig. 4: Lorenz Model (Section V-A). Histograms of averaged MSE and M with $[p_\ell - p_h] = [0.3 - 0.5]$ and $[p_\ell - p_h] = [0.5 - 0.7]$. In both cases, the initial number of particles $M_0 = 5000$.

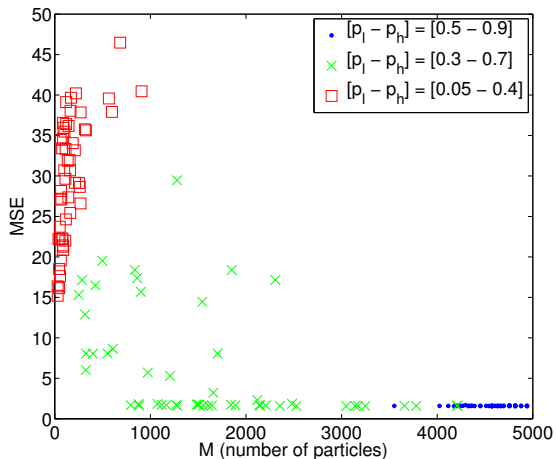


Fig. 5: Lorenz Model (Section V-A). MSE w.r.t. the averaged number of particles M for runs with different sets of thresholds.

VI. CONCLUSIONS

In practice, the number of particles needed in a particle filter is usually determined in an ad hoc manner. Furthermore, this number is typically kept constant throughout tracking. In this paper, we have proposed a methodology for the online determination of the number of particles needed by the filter. The approach is based on assessing the convergence of the predictive distribution of the observations online. First we have proved, under standard assumptions, a novel convergence result on the approximation of this distribution. Then, we have proposed a method for adapting the number of particles based on the online assessment of the filter convergence. The proposed procedure is simple but not unique. One can develop a range of algorithms for adapting the number of particles using the proposed methodology. We have illustrated the performance of the suggested algorithm by computer simulations.

APPENDIX A

PROOF OF THEOREM 1

Recall that the likelihood of $\mathbf{X}_t = \mathbf{x}_t$ given the observation $\mathbf{Y}_t = \mathbf{y}_t$ is denoted $g_t^{\mathbf{y}_t}(\mathbf{x}_t)$, i.e., $g_t^{\mathbf{y}_t}(\mathbf{x}_t) = p(\mathbf{y}_t|\mathbf{x}_t)$. For the sake of notational accuracy, we introduce the Markov transition kernel $\tau_t(d\mathbf{x}_t|\mathbf{x}_{t-1})$ that determines the dynamics of the state process. The correspondence with the notation in Section II is $\tau_t(d\mathbf{x}_t|\mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})d\mathbf{x}_t$, however all the results in this appendix (including Theorem 1) are proved for the general case in which τ_t does not necessarily have a density w.r.t. the Lebesgue measure. For notational coherence, we denote $\tau_0(d\mathbf{x}_0) = p(\mathbf{x}_0)d\mathbf{x}_0$.

The same as in Section II, the integral of a function $f : \mathcal{Z} \rightarrow \mathbb{R}$ w.r.t. a measure α on the measurable space $(\mathcal{B}(\mathcal{Z}), \mathcal{Z})$ is denoted (f, α) and the absolute supremum of f is written $\|f\|_\infty = \sup_{z \in \mathcal{Z}} |f(z)|$. The class of bounded functions is denoted $B(\mathcal{Z}) = \{f : \mathcal{Z} \rightarrow \mathbb{R} : \|f\|_\infty < \infty\}$. For $p \geq 1$, the L_p norm of a r.v. Z with associated probability measure $\gamma(dz)$ is denoted

$$\|Z\|_p := E[|Z|^p]^{\frac{1}{p}} = \left(\int |z|^p \gamma(dz) \right)^{\frac{1}{p}},$$

where $E[\cdot]$ denotes expectation.

We start introducing some auxiliary results on the convergence of the approximate measure ξ_t^M and integrals of the form $(D_y^1 g_t^{\mathbf{y}_t}, \xi_t^M)$. This leads to the core result, which is the uniform convergence of $p_t^M(\mathbf{y}_t) \rightarrow p_t(\mathbf{y}_t)$ on a sequence of compact sets. The proof of Theorem 1 follows readily from the latter result.

The analysis in this Appendix draws from methods developed in [31] for the estimation of the filter pdf $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ using kernel functions, which herein are suitably adapted to the problem of approximating the predictive density $p_t(\mathbf{y}_t)$.

Lemma 1. *Assume that the sequence $\mathbf{y}_{0:T}$, for $T < \infty$, is arbitrary but fixed, and, for each $t = 1, 2, \dots, T$, $g_t^{\mathbf{y}_t} \in B(\mathcal{X})$ and $g_t^{\mathbf{y}_t} > 0$. Then, there exist constants $c_t < \infty$, $t = 0, 1, \dots, T$, independent of M such that*

$$\|(f, \xi_t^M) - (f, \xi_t)\|_p \leq \frac{c_t \|f\|_\infty}{\sqrt{M}}, \quad t = 0, 1, 2, \dots$$

for every $f \in B(\mathcal{X})$.

Proof: This is a particular case of [32, Lemma 1]. □

Lemma 2. *Assume that the sequence $\mathbf{y}_{0:t-1}$, for $t < \infty$, is arbitrary but fixed. If assumptions (\mathfrak{L}) and (\mathfrak{D}) hold, then for each $p \geq 1$ there exists a constant $\bar{c}_t < \infty$ independent of M such that*

$$E \left[\left| D_y^1 p_t^M(\mathbf{y}_t) - D_y^1 p_t(\mathbf{y}_t) \right|^p \right] \leq \frac{\bar{c}_t^p}{M^{\frac{p}{2}}}. \quad (16)$$

Proof: We first note that $D_y^1 p_t(\mathbf{y}_t) = (D_y^1 g_t^{y_t}, \xi_t)$ and $D_y^1 p_t^M(\mathbf{y}_t) = (D_y^1 g_t^{y_t}, \xi_t^M)$, where function $D_y^1 g_t^{y_t}(\mathbf{x}_t)$ is bounded (for any \mathbf{y}_t) because of assumption (\mathfrak{D}) . Since (\mathfrak{L}) also holds, Lemma 1 yields

$$\begin{aligned} \|D_y^1 p_t^M(\mathbf{y}_t) - D_y^1 p_t(\mathbf{y}_t)\|_p &= \|(D_y^1 g_t^{y_t}, \xi_t^M) - (D_y^1 g_t^{y_t}, \xi_t)\|_p \\ &\leq \frac{\bar{c}_t}{\sqrt{M}}, \end{aligned} \quad (17)$$

where the constant $\bar{c}_t = c_t \|D_y^1 g_t\|_\infty$ is finite (see (\mathfrak{D})) and independent of M . If we raise both sides of (17) to power p , then we obtain the desired result (16). \square

Lemma 3. Let $\{\theta^M\}_{M \geq 1}$ be a sequence of non-negative r.v.'s such that, for every $p \geq 4$,

$$E[(\theta^M)^p] \leq \frac{c}{M^{\frac{p}{2}-\nu}} \quad (18)$$

where $c < \infty$ and $0 \leq \nu < 1$ are constants independent of M . Then, for every $\epsilon \in (0, \frac{1}{2})$ there exists an a.s. finite r.v. U^ϵ independent of M such that

$$\theta^M \leq \frac{U^\epsilon}{M^{\frac{1}{2}-\epsilon}}.$$

Proof: Let us choose an arbitrary constant $\psi \in (\nu, 1)$ and define the r.v. $U^{\psi,p} = \sum_{M=1}^{\infty} M^{\frac{p}{2}-1-\psi} (\theta^M)^p$. If (18) holds, then the expectation $E[U^{\psi,p}]$ is finite, as we prove in the sequel. Indeed, from Fatou's lemma,

$$E[U^{\psi,p}] \leq \sum_{M=1}^{\infty} M^{\frac{p}{2}-1-\psi} E[(\theta^M)^p] \quad (19)$$

$$\leq \sum_{M=1}^{\infty} M^{\nu-\psi-1}, \quad (20)$$

where (20) follows from substituting (18) into (19). Since we have chosen $\psi \in (\nu, 1)$, then it follows that $-1 < \nu - \psi < 0$ and $\nu - \psi - 1 < -1$, which ensures that $\sum_{M=1}^{\infty} M^{\nu-\psi-1} < \infty$ and, therefore, $E[U^{\psi,p}] < \infty$. Since $E[U^{\psi,p}] < \infty$, then $U^{\psi,p} < \infty$ a.s.

For any given value of M , it is apparent from the definition of $U^{\psi,p}$ that

$$M^{\frac{p}{2}-1-\psi} (\theta^M)^p \leq U^{\psi,p}$$

and, as a consequence,

$$\theta^M \leq \frac{(U^{\psi,p})^{\frac{1}{p}}}{M^{\frac{1}{2}-\frac{1+\psi}{p}}} = \frac{U^\epsilon}{M^{\frac{1}{2}-\epsilon}} \quad (21)$$

where the equality in (21) follows from defining $\epsilon := \frac{1+\psi}{p}$ and $U^\epsilon := (U^{\psi,p})^{\frac{1}{p}}$. Since $\psi < 1$, it is sufficient to choose $p \geq 4$ to ensure that $\epsilon = \frac{1+\psi}{p} < \frac{1}{2}$. Also, since p can actually be chosen as large as we wish, it follows that (21) holds for $\epsilon > 0$ as small as needed. \square

Lemma 4. Assume that the sequence $\mathbf{y}_{0:T}$, for $T < \infty$, is arbitrary but fixed, and, for $t = 1, 2, \dots, T$, $g_t^{y_t} \in B(\mathcal{X})$ and $g_t^{y_t} > 0$. Then, for every $0 < \epsilon < \frac{1}{2}$ (arbitrarily small) there exist a.s. finite r.v.'s $U_t^\epsilon < \infty$, $t = 0, 1, \dots, T$, independent of M such that

$$\|(f, \xi_t^M) - (f, \xi_t)\|_p \leq \frac{U_t^\epsilon}{M^{\frac{1}{2}-\epsilon}}, \quad t = 0, 1, 2, \dots \quad (22)$$

for every $f \in B(\mathcal{X})$.

Proof: From Lemma 1, for each $t = 1, \dots, T$, there is a constant c_t independent of M such that

$$E[|(f, \xi_t^M) - (f, \xi_t)|^p] \leq \frac{c_t^p \|f\|_\infty^p}{M^{\frac{p}{2}}}$$

for any $f \in B(\mathcal{X})$. Therefore, we can apply Lemma 3 with $c = c_t^p \|f\|_\infty^p$ and $\nu = 0$, to obtain the desired inequality (22). \square

For the statement of the next result, we need to recall the definition of the sequence of hypercubes

$$C_M := \left[-\frac{M^{\frac{\beta}{d_y p}}}{2}, +\frac{M^{\frac{\beta}{d_y p}}}{2}\right] \times \cdots \times \left[-\frac{M^{\frac{\beta}{d_y p}}}{2}, +\frac{M^{\frac{\beta}{d_y p}}}{2}\right] \subset \mathbb{R}^{d_y}$$

in assumption (C), where $p \geq 4$ and $0 < \beta < 1$ are constants w.r.t. M .

Lemma 5. *Let the sequence $\mathbf{y}_{0:T}$, $T < \infty$, be arbitrary but fixed and assume that (L) and (D) hold. Then, for any $0 < \varepsilon < \frac{1}{2}$ and each $t = 1, 2, \dots, T$ there exists an a.s. finite r.v. V_t^ε independent of M such that*

$$\sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq \frac{V_t^\varepsilon}{M^{\frac{1}{2}-\varepsilon}}. \quad (23)$$

In particular,

$$\lim_{M \rightarrow \infty} \sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| = 0 \quad a.s.$$

Proof: Let $b_M = \frac{1}{2} M^{\frac{\beta}{d_y p}}$, in such a way that the hypercube C_M can be written as $C_M = [-b_M, +b_M]^{d_y} \subset \mathbb{R}^{d_y}$. For any $\mathbf{y} = [y_1, y_2, \dots, y_{d_y}]^\top \in C_M$ and any function $f : \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ continuous, bounded and differentiable, one can write

$$f(\mathbf{y}) - f(\mathbf{0}) = \int_{-b_M}^{y_1} \cdots \int_{-b_M}^{y_{d_y}} D_z^1 f(\mathbf{z}) d\mathbf{z} - \int_{-b_M}^0 \cdots \int_{-b_M}^0 D_z^1 f(\mathbf{z}) d\mathbf{z}.$$

In particular, if $\mathbf{y} \in [-b_M, b_M]^{d_y}$ and assumption (D) holds, then we can write

$$p_t^M(\mathbf{y}) - p_t(\mathbf{y}) = \int_{-b_M}^{b_M} \cdots \int_{-b_M}^{b_M} (D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})) d\mathbf{y} + (p_t^M(\mathbf{0}) - p_t(\mathbf{0}))$$

and, as a consequence,

$$|p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq \int_{-b_M}^{b_M} \cdots \int_{-b_M}^{b_M} |D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})| d\mathbf{y} + |p_t^M(\mathbf{0}) - p_t(\mathbf{0})|$$

which, in turn, yields the inequality

$$\sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq A^M + |p_t^M(\mathbf{0}) - p_t(\mathbf{0})|, \quad (24)$$

where

$$A^M = \int_{-b_M}^{b_M} \cdots \int_{-b_M}^{b_M} |D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})| d\mathbf{y}.$$

An application of Jensen's inequality yields, for $p \geq 1$,

$$\left(\frac{1}{(2b_M)^{d_y}} A^M\right)^p \leq \frac{1}{2^{d_y} b_M^{d_y}} \int_{-b_M}^{b_M} \cdots \int_{-b_M}^{b_M} |D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})|^p d\mathbf{y},$$

which leads to

$$(A^M)^p \leq 2^{d_y(p-1)} b_M^{d_y(p-1)} \times \int_{-b_M}^{b_M} \cdots \int_{-b_M}^{b_M} |D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})|^p d\mathbf{y}. \quad (25)$$

Since, from Lemma 2,

$$E \left[|D_y^1 p_t^M(\mathbf{y}) - D_y^1 p_t(\mathbf{y})|^p \right] \leq \frac{\bar{c}_t^p}{M^{\frac{p}{2}}}, \quad (26)$$

we can combine (26) and (25) to arrive at

$$E [(A^M)^p] \leq \frac{2^{d_y p} b_M^{d_y p} \bar{c}_t^p}{M^{\frac{p}{2}}} = \frac{\bar{c}_t^p}{M^{\frac{p}{2}-\beta}},$$

where the equality follows from the relationship $b_M = \frac{1}{2}M^{\frac{\beta}{d_{y^p}}}$. If we apply Lemma 3 with $\theta^M = A^M$, $p \geq 4$, $\nu = \beta$ and $c = \tilde{c}_t^p$, then we obtain a constant $\varepsilon_1 \in \left(\frac{1+\beta}{p}, \frac{1}{2}\right)$ (see (21)) and a non-negative and a.s. finite random variable V^{A, ε_1} , both of them independent of M , such that

$$A^M \leq \frac{V^{A, \varepsilon_1}}{M^{\frac{1}{2} - \varepsilon_1}}. \quad (27)$$

Moreover, Lemma 2 yields

$$\begin{aligned} E \left[|p_t^M(\mathbf{y}) - p_t(\mathbf{y})|^p \right] &= E \left[|(g_t^{\mathbf{y}}, \xi_t^M) - (g_t^{\mathbf{y}}, \xi_t)|^p \right] \\ &\leq \frac{\tilde{c}_t^p \|g_t^{\mathbf{y}}\|_{\infty}^p}{M^{\frac{p}{2}}}, \end{aligned}$$

where $\tilde{c}_t < \infty$ is a constant independent of M and $\|g_t^{\mathbf{y}}\|_{\infty} < \infty$ independently of \mathbf{y} from assumption (\mathfrak{L}) . Therefore, we can apply Lemma 3 again, with $\theta^M = |p_t^k(0) - p_t(0)|$, $p \geq 4$, $\nu = 0$ and $c = \tilde{c}_t^p$ to obtain the inequality

$$|p_t^M(\mathbf{0}) - p_t(\mathbf{0})| \leq \frac{V^{p_t(\mathbf{0}), \varepsilon_2}}{M^{\frac{1}{2} - \varepsilon_2}}, \quad (28)$$

where $\varepsilon_2 \in \left(\frac{1}{p}, \frac{1}{2}\right)$ is a constant and $V^{p_t(\mathbf{0}), \varepsilon_2}$ is a non-negative and a.s. finite r.v., both of them independent of M .

If we choose $\varepsilon = \varepsilon_1 = \varepsilon_2 \in \left(\frac{1+\beta}{p}, \frac{1}{2}\right)$ and define $V_t^\varepsilon = V^{A, \varepsilon_1} + V^{p_t(\mathbf{0}), \varepsilon_2}$, then the combination of Eqs. (24), (27) and (28) yields

$$\sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq \frac{V_t^\varepsilon}{M^{\frac{1}{2} - \varepsilon}},$$

where V_t^ε is a.s. finite. Note that V_t^ε and ε are independent of M . Moreover, we can choose p as large as we wish and $\beta > 0$ as small as needed, hence we can effectively select $\varepsilon \in (0, \frac{1}{2})$. \square

Before stating the next partial result, let us recall assumption (\mathfrak{C}) again, namely the inequality $\mu_t(\overline{C_M}) \leq bM^{-\eta}$, where $b > 0$ and $\eta < 1$ are constants w.r.t M and $\overline{C_M}$ is the complement of C_M .

Lemma 6. *Let the sequence $\mathbf{y}_{0:T}$, $T < \infty$, be arbitrary but fixed and assume that (\mathfrak{L}) , (\mathfrak{D}) and (\mathfrak{C}) hold. Then, for any $0 < \varepsilon < \frac{1}{2}$ and each $t = 1, 2, \dots, T$ there exists an a.s. finite r.v. W_t^ε independent of M such that*

$$\int |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} \leq \frac{\tilde{W}_t^\varepsilon}{M^{(\frac{1}{2} - \varepsilon) \wedge \eta}}. \quad (29)$$

Proof: We start with a trivial decomposition of the integrated absolute error,

$$\begin{aligned} \int |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} &= \int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} + \int_{\overline{C_M}} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} \\ &\leq \int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} + 2 \int_{\overline{C_M}} p_t(\mathbf{y}) d\mathbf{y} \int_{\overline{C_M}} (p_t^M(\mathbf{y}) - p_t(\mathbf{y})) d\mathbf{y}, \end{aligned}$$

where the equality follows from $C_M \cup \overline{C_M} = \mathbb{R}^{d_y}$ and the inequality is obtained from the fact that p_t and p_t^M are non-negative, hence $|p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq p_t^M(\mathbf{y}) + p_t(\mathbf{y})$. Moreover, if we realize that

$$\begin{aligned} \int_{\overline{C_M}} (p_t^M(\mathbf{y}) - p_t(\mathbf{y})) d\mathbf{y} &= 1 - \int_{C_M} p_t^M(\mathbf{y}) d\mathbf{y} - 1 + \int_{C_M} p_t(\mathbf{y}) d\mathbf{y} \\ &= \int_{C_M} (p_t(\mathbf{y}) - p_t^M(\mathbf{y})) d\mathbf{y} \end{aligned}$$

then it is straightforward to see that

$$\int_{\overline{C_M}} (p_t^M(\mathbf{y}) - p_t(\mathbf{y})) d\mathbf{y} \leq \int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} \quad (30)$$

and, as a consequence, substituting (30) into (30),

$$\int |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} \leq 2 \int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} + 2 \int_{C_M} p_t(\mathbf{y}) d\mathbf{y} \quad (31)$$

The first term on the right-hand side of (31) can be bounded easily because C_M is compact, namely

$$\int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| dx \leq \mathcal{L}(\overline{C_M}) \sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})|, \quad (32)$$

where $\mathcal{L}(C_M) = (2b_M)^{d_y} = M^{\frac{\beta}{p}}$ is the Lebesgue measure of C_M . From Lemma 5, the supremum in (32) can be bounded as $\sup_{\mathbf{y} \in C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| \leq V_t^{\varepsilon_1} / M^{\frac{1}{2} - \varepsilon_1}$, where $V_t^{\varepsilon_1} \geq 0$ is an a.s. finite r.v. and $\frac{1+\beta}{p} < \varepsilon_1 < \frac{1}{2}$ is a constant, both independent of M . Therefore, the inequality (32) can be extended to yield

$$\int_{C_M} |p_t^M(\mathbf{y}) - p_t(\mathbf{y})| d\mathbf{y} \leq \frac{V_t^{\varepsilon_1}}{M^{\frac{1}{2} - \varepsilon_1 - \frac{\beta}{p}}} = \frac{V_t^\varepsilon}{M^{\frac{1}{2} - \varepsilon}}, \quad (33)$$

where $\varepsilon = \varepsilon_1 + \frac{\beta}{p}$ and $V_t^\varepsilon = V_t^{\varepsilon_1}$. If we choose $\varepsilon_1 < \frac{1}{2} - \frac{\beta}{p}$, then $\varepsilon \in \left(\frac{1+2\beta}{p}, \frac{1}{2}\right)$. Note that, for $\beta < 1$ and choosing $p \geq 6$, $\frac{1}{2} - \frac{\beta}{p} - \frac{1+\beta}{p} > \frac{1}{2} - \frac{3}{p} > 0$, hence both ε_1 and ε are well defined. Now, taking p large enough we can effectively select $\varepsilon \in (0, \frac{1}{2})$.

For the second integral in Eq. (31), note that $\int_{\overline{C_M}} p_t(\mathbf{y}) d\mathbf{y} = \mu_t(\overline{C_M})$ and, therefore, it can be bounded directly from the assumptions in the present Lemma, i.e.,

$$2 \int_{\overline{C_M}} p_t(\mathbf{y}) d\mathbf{y} \leq 2bM^{-\eta}, \quad (34)$$

where $b > 0$ and $\eta > 0$ are constant w.r.t. M . Putting together Eqs. (31), (33) and (34) yields the desired result, with $\tilde{W}_t^\varepsilon = 2(V_t^\varepsilon + b) < \infty$ a.s. \square

Finally, the proof of Theorem 1 is a straightforward application of Lemma 6.

Proof of Theorem 1. We first note that, for any bounded function h ,

$$\begin{aligned} (h, \mu_t^M) - (h, \mu_t) &= \int h(y) p_t^M(y) dy - \int h(y) p_t(y) dy \\ &= \int h(y) (p_t^M(y) - p_t(y)) dy, \end{aligned}$$

hence, trivially,

$$|(h, \mu_t^M) - (h, \mu_t)| \leq \|h\|_\infty \int |p_t^M(y) - p_t(y)| dy. \quad (35)$$

If we apply Lemma 6 on the right hand side of (35) then we readily obtain

$$|(h, \mu_t^M) - (h, \mu_t)| \leq \|h\|_\infty \frac{\tilde{W}_t^\varepsilon}{M^{(\frac{1}{2} - \varepsilon) \wedge \eta}}. \quad (36)$$

where $\varepsilon \in (0, \frac{1}{2})$ is an arbitrarily small constant independent of M and $W_t^\varepsilon = \|h\|_\infty \tilde{W}_t^\varepsilon$ is an a.s. finite r.v., also independent of M . \square

APPENDIX B

PROOF OF THEOREM 2

Let Y_t denote the (random) observation at time t . Assume, without loss of generality, that $\mathcal{Y} = \mathbb{R}$. The probability measure associated to $Y_t | Y_{1:t-1} = y_{1:t-1}$ is $\mu_t(dy)$ and, therefore, we can write the cumulative distribution function of $Y_t | Y_{1:t-1} = y_{1:t-1}$ as $F_t(z) = (I_{(-\infty, z]}, \mu_t)$, where

$$I_A(y) = \begin{cases} 1, & \text{if } y \in A \\ 0, & \text{otherwise} \end{cases}$$

is the indicator function. Obviously, $\|I_A\|_\infty = 1 < \infty$ independently of the set A and, therefore, Theorem 1 yields

$$\lim_{M \rightarrow \infty} F_t^M(z) = F_t(z) \quad \text{a.s.}$$

for any $z \in \mathbb{R}$, where $F_t^M(z) = (I_{(-\infty, z]}, \mu_t^M)$ is the approximation of the cdf of $Y_t | Y_{1:t-1} = y_{1:t-1}$ provided by the BF.

Assume the actual observation is $Y_t = y_t$ and we draw K i.i.d. fictitious observations $\tilde{y}_t^{(1)}, \dots, \tilde{y}_t^{(K)}$ from the distribution with cdf F_t^M . Given $Y_t = y_t$ is fixed, the probability that exactly n out of K of these samples are lesser than y_t coincides with the probability to have n successes out of K trials for a binomial r.v. with parameter (i.e., success probability) $F_t^M(y_t)$, which can be written as

$$h_n^M(y_t) = \binom{K}{n} (F_t^M(y_t))^n (1 - F_t^M(y_t))^{K-n}.$$

By integrating $h_n^M(y_t)$ over the predictive distribution of Y_t , we obtain the probability to have exactly n fictitious observations, out of K , which are less than the r.v. Y_t , i.e., the probability that $A_{K,M,t} = n$ is

$$\mathbb{Q}_{K,M,t}(n) = (h_n^M, \mu_t). \quad (37)$$

However, Theorem 1 yields $\lim_{M \rightarrow \infty} (h_n^M, \mu_t^M) = (h_n^M, \mu_t)$ a.s.⁴ and, in particular, there exists a sequence of non-negative r.v.'s $\{\varepsilon_M\}_{M \geq 1}$ such that $\lim_{M \rightarrow \infty} \varepsilon_M = 0$ a.s. and

$$(h_n^M, \mu_t^M) - \varepsilon_M \leq (h_n^M, \mu_t) \leq (h_n^M, \mu_t^M) + \varepsilon_M \quad (38)$$

for each M . Moreover, it is apparent that $(h_n^M, \mu_t^M) = \frac{1}{K+1}$ (see Proposition 1) which, together with (37) and (38) yields the desired relationship

$$\frac{1}{K+1} - \varepsilon_M \leq \mathbb{Q}_{K,M,t}(n) \leq \frac{1}{K+1} + \varepsilon_M$$

for every $n \in \{0, \dots, K\}$. □

REFERENCES

- [1] M. West and J. Harrison, *Bayesian Forecasting, 2nd ed.*, Springer-Verlag, New York, 1996.
- [2] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter*, Artech House, Boston, 2004.
- [3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.
- [4] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Englewood Cliffs, 1979.
- [5] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.
- [6] J. S. Liu, R. Chen, and W. H. Wong, "Rejection control and sequential importance sampling," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1022–1031, September 1998.
- [7] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo Sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.
- [8] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, New York (USA), 2001.
- [9] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez, "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, September 2003.
- [10] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.
- [11] N. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear and non-Gaussian Bayesian state estimation," *IEE Proceedings-F Radar and Signal Processing*, vol. 140, pp. 107–113, 1993.
- [12] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*, Springer, 2008.
- [13] P. Del Moral and L. Miclo, "Branching and interacting particle systems. Approximations of Feynman-Kac formulae with applications to non-linear filtering," *Lecture Notes in Mathematics*, pp. 1–145, 2000.

⁴Note that $\|h_n^M\|_\infty = 1$ independently of n and M . If we recall the proof of Theorem 1, namely inequality (36), we observe that the error rates for the approximation errors of the form $|(h, \mu_t^M) - (h, \mu_t)|$ depend on the test function h only through its supremum $\|h\|_\infty$, i.e., the r.v. \tilde{W}_t^ε in (36) only depends on the observations $y_{1:t-1}$ and the model (specifically the likelihood functions). Therefore, Theorem 1 (the same as, e.g., Lemmas 1 and 2) also holds for any test function that depends on M (even a random one) as long as its supremum is deterministic and independent of M . This is the case of function $h_n^M(y)$.

- [14] D. Fox, "Adapting the sample size in particle filters through kld-sampling," *The International Journal of Robotics Research*, vol. 22, no. 12, pp. 985–1003, 2003.
- [15] A. Soto, "Self adaptive particle filter," in *IJCAI*, 2005, pp. 1398–1406.
- [16] O. Straka and M. Šimandl, "Particle filter adaptation based on efficient sample size," in *14th IFAC Symposium on System Identification*, 2006.
- [17] P. M. Djurić and J. Míguez, "Assessment of nonlinear dynamic models by Kolmogorov–Smirnov statistics," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5069–5079, 2010.
- [18] R. Chen, X. Wang, and J. S. Liu, "Adaptive joint detection and decoding in flat-fading channels via mixture kalman filtering," *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 2079–2094, September 2000.
- [19] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "SMC2: An efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.
- [20] D. Crisan and J. Míguez, "Nested particle filters for online parameter estimation in discrete-time state-space markov models," *arXiv: 1308.1883v1 [stat.CO]*, 2013.
- [21] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society B*, vol. 72, no. 3, pp. 269–342, 2010.
- [22] A. Doucet, N. de Freitas, and N. Gordon, "An introduction to sequential Monte Carlo methods," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds., chapter 1, pp. 4–14. Springer, 2001.
- [23] J. Carpenter, P. Clifford, and P. Fearnhead, "Improved particle filter for nonlinear problems," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 146, no. 1, pp. 2–7, February 1999.
- [24] D. Crisan, "Particle filters - a theoretical perspective," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds., chapter 2, pp. 17–42. Springer, 2001.
- [25] T. Li, M. Bolić, and P. M. Djurić, "Resampling methods for particle filtering," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 70–86, May 2015.
- [26] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*, Springer, 2004.
- [27] R. L. Plackett, "Karl pearson and the chi-squared test," *International Statistical Review/Revue Internationale de Statistique*, pp. 59–72, 1983.
- [28] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of Atmospheric Sciences*, vol. 20, no. 2, pp. 130–141, 1963.
- [29] A. J. Chorin and P. Krause, "Dimensional reduction for a Bayesian filter," *PNAS*, vol. 101, no. 42, pp. 15013–15017, October 2004.
- [30] M. S. Nikulin, "Hellinger distance," *Encyclopedia of Mathematics*, 2001.
- [31] D. Crisan and J. a Míguez, "Particle-kernel estimation of the filter density in state-space models," *Bernoulli*, vol. 20, no. 4, pp. 1879–1929, 2014.
- [32] J. Míguez, D. Crisan, and P. M. Djurić, "On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization," *Statistics and Computing*, vol. 23, no. 1, pp. 91–107, 2013.