Submitted to Bernoulli
arXiv: 1308.1883

Nested particle filters for online parameter estimation in discrete–time state–space Markov models

DAN CRISAN[1,*] and JOAQUÍN MÍGUEZ[2,**]

[1] *Department of Mathematics (Huxley Building), Imperial College London. 180 Queens Gate, London SW7 2BZ, UK. E-mail:* [*] d.crisan@imperial.ac.uk

[2] *Department of Signal Theory and Communications, Universidad Carlos III de Madrid. Avenida de la Universidad 30, 28911 Leganés, Madrid, Spain. E-mail:* [**] joaquin.miguez@uc3m.es

We address the problem of approximating the posterior probability distribution of the fixed parameters of a state-space dynamical system using a sequential Monte Carlo method. The proposed approach relies on a nested structure that employs two layers of particle filters to approximate the posterior probability measure of the static parameters and the dynamic state variables of the system of interest, in a vein similar to the recent "sequential Monte Carlo square" (SMC^2) algorithm. However, unlike the SMC^2 scheme, the proposed technique operates in a purely recursive manner. In particular, the computational complexity of the recursive steps of the method introduced herein is constant over time. We analyse the approximation of integrals of real bounded functions with respect to the posterior distribution of the system parameters computed via the proposed scheme. As a result, we prove, under regularity assumptions, that the approximation errors vanish asymptotically in L_p ($p \geq 1$) with convergence rate proportional to $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}$, where N is the number of Monte Carlo samples in the parameter space and $N \times M$ is the number of samples in the state space. This result also holds for the approximation of the joint posterior distribution of the parameters and the state variables. We discuss the relationship between the SMC^2 algorithm and the new recursive method and present a simple example in order to illustrate some of the theoretical findings with computer simulations.

Keywords: particle filtering, parameter estimation, model inference, state space models, recursive algorithms, Monte Carlo, error bounds.

1. Introduction

1.1. Problem statement

The problem of parameter estimation in state-space dynamical systems has received considerable attention, from different viewpoints (Kitagawa, 1998; Liu and West, 2001; Andrieu et al., 2004; Kantas et al., 2015; Carvalho et al., 2010), as it is almost ubiquitous in practical applications. In this paper, we investigate the use of particle filtering methods for the online Bayesian estimation of the static parameters of a state-space system.

In order to ease the discussion, let us consider two (possibly vector-valued) random sequences $\{X_t\}_{t=0,1,...}$ and $\{Y_t\}_{t=1,2,...}$ representing the (hidden) state of a dynamical system and some related observations, respectively, with t denoting discrete time. We assume that the state process is Markov and the observation Y_t is independent of any other observations $\{Y_k; k \neq t\}$, conditional on the state X_t. Both the conditional probability distribution of X_t given the value of the previous state, $X_{t-1} = x_{t-1}$, and the probability density function (pdf) of Y_t given $X_t = x_t$ are assumed to be known up to a vector of static (random) parameters, denoted by Θ. These assumptions are commonly made in the literature and actually hold for many practical systems of interest (see, e.g., (Ristic, Arulampalam and Gordon, 2004; Cappé, Godsill and Moulines, 2007)). Given a sequence of actual observations, $Y_1 = y_1, \ldots, Y_t = y_t, \ldots$, the goal is to track the posterior probability distributions of the state X_t, $t \geq 0$, and the parameter vector Θ over time.

In the sequel, we briefly review various existing approaches to the parameter estimation problem that involve particle filtering in some relevant manner. See (Kantas et al., 2015) for a more detailed survey of the field.

1.2. Particle filters and parameter estimation

When the parameter vector is given, i.e., $\Theta = \theta$ is known, the problem reduces to the standard stochastic filtering setting, which consists in tracking the posterior probability distribution of the state X_t, given the record of observations up to time $t > 0$. In a few special cases (e.g., if the system is linear and Gaussian or the state-space is discrete and finite) there exist closed form solutions for the probability distribution of X_t given $Y_1 = y_1, ..., Y_t = y_t$, which is often termed the *filtering* distribution. However, analytical solutions do not exist for general, possibly nonlinear and non-Gaussian, systems and numerical approximation methods are then needed. One popular class of such methods are the so-called particle filters (Gordon, Salmond and Smith, 1993; Kitagawa, 1996; Liu and Chen, 1998; Doucet, Godsill and Andrieu, 2000). This is a family of recursive Monte Carlo algorithms that generate discrete random approximations of the sequence of probability measures associated to the filtering distributions at discrete time $t \geq 0$.

Particle filters are well suited for solving the standard stochastic filtering problem. However, the design of particle filters that can account for a random vector of parameters in the dynamic system (i.e., a static but unknown Θ) has been an open issue for the past two decades.

When the system of interest is endowed with some structure, there are some elegant techniques to handle the unknown parameters efficiently. For example, there are various conditionally-linear and Gaussian models that admit the analytical integration of Θ using the Kalman filter as an auxiliary tool, see, e.g., (Doucet, Godsill and Andrieu, 2000; Chen, Wang and Liu, 2000). A similar approach can be taken with some non-Gaussian models appearing, e.g., in signal processing (Bruno, 2013). In other cases, the analytical integration may not be feasible but the structure of the model can be such that the conditional probability law of Θ given $X_0 = x_0, \ldots, X_t = x_t$ and $Y_1 = y_1, \ldots, Y_t = y_t$ is tractable. In particular, if Θ depends on $X_{1:t} = \{X_1, ..., X_t\}$ through a low-dimensional

sufficient statistic then it is possible to draw efficiently from the posterior distribution of Θ (given $X_{0:t} = x_{0:t}$ and $Y_{1:t} = y_{1:t}$) (Storvik, 2002; Carvalho et al., 2010) and then integrate the parameters out numerically.

For arbitrary systems, with no particular structure, the more straightforward approach is to augment the state-space by including Θ as a constant-in-time state variable. This has been proposed in a number of forms and in various applications[1] but it can be shown that standard particle filters working on this augmented state-space do not necessarily converge in general because the resulting systems are non-ergodic (Andrieu et al., 2004; Papavasiliou, 2006). Another popular technique to handle static parameters within particle filtering consists in building a suitable kernel estimator of the posterior probability density function (pdf) of Θ given $Y_{1:t} = y_{1:t}$ from where new samples in the parameter space can be drawn (Liu and West, 2001). The latter step is often called "rejuvenation" or "jittering" (we adopt the latter term in the sequel). One key feature of this technique is the "shrinkage" of the density estimator in order to control the variance of the jittered particles. This method has been shown to work in some examples with low-dimensional Θ, but has also been found to deliver poor performance in other simple setups (Miguez, Bugallo and Djuric, 2005). A rigorous analysis of this technique is missing as well.

Finally, there exists a large body of research on maximum likelihood estimation (MLE) for unknown parameters. Instead of handling Θ as a random variable and building an approximation of its posterior distribution, MLE techniques aim at computing a single-point estimate of the parameters. This is typically done by way of gradient optimisation methods, that lend themselves naturally to online implementations. A popular example is the recursive maximum likelihood (RML) algorithm (LeGland and Mevel, 1997; Poyiadjis, Doucet and Singh, 2011; Moral, Doucet and Singh, 2015). As an alternative to gradient search methods, expectation maximization (EM) techniques have also been proposed for the optimisation of the parameter likelihood, both in offline and online versions (Andrieu et al., 2004; Kantas et al., 2015). These techniques use particle filtering as an ancillary tool to approximate either the gradient of the likelihood function (Moral, Doucet and Singh, 2015) or some sufficient statistics (Andrieu et al., 2004) and have been advocated as more robust than those based on state-space augmentation, artificial evolution or kernel density estimation (Andrieu et al., 2004; Kantas et al., 2015).

1.3. Non-recursive methods

A number of new methods related to particle filtering have been proposed in the past few years that tackle the problem of approximating the distribution of the parameter vector Θ given the observations $Y_{1:T} = y_{1:T}$. These techniques include the iterated batch importance sampling (IBIS) algorithm of (Chopin, 2002) and extensions

[1] It has also been proposed to use Markov chain Monte Carlo (MCMC) steps to prevent the collapse of the population representing the parameter posterior, that otherwise occurs due to the resampling steps (Gilks and Berzuini, 2001; Fearnhead, 2002).

of it that rely on the *nesting* of particle methods (such as in (Papavasiliou, 2006) or (Chopin, Jacob and Papaspiliopoulos, 2013)), combinations of Markov chain Monte Carlo (MCMC) and particle filtering (Andrieu, Doucet and Holenstein, 2010), variations of the population Monte Carlo methodology (Koblents and Míguez, 2013) and particle methods for the approximation of the parameter likelihood function (Olsson et al., 2008).

The IBIS method is a sequential Monte Carlo (SMC) algorithm that updates a population of samples $\theta_t^{(i)}$, $i = 1, ..., N$, in the space of Θ, with associated importance weights, at every time step. The technique involves regular MCMC steps, in order to rejuvenate the population of samples, and the ability to compute the pdf of every observation variable Y_t, given the previous observation record $Y_{1:t-1} = y_{1:t-1}$ and a fixed value of the parameters, $\Theta = \theta$. Let us denote such densities as $d(y_t|y_{1:t-1}, \theta)$ for the sake of conciseness. The need to obtain $d(y_t|y_{1:t-1}, \theta)$ in closed-form has two important implications. First, IBIS is not a recursive algorithm, since each time we need to compute $d(y_t|y_{1:t-1}, \theta)$ for a new sample point $\Theta = \theta$ in the parameter space it is necessary to process the entire sequence of observations $Y_{1:t-1} = y_{1:t-1}$. Second, the algorithm can only be applied when the dynamic model has some suitable structure (e.g., the system may be linear and Gaussian conditional on Θ) that enables us to actually find $d(y_t|y_{1:t-1}, \theta)$ in closed form.

In (Papavasiliou, 2006), these difficulties with the IBIS method are addressed by using two layers of Monte Carlo methods. First, a random grid of points in the space of Θ, say $\theta^{(1)}, ..., \theta^{(N)}$, is generated. Then, for each $\Theta = \theta^{(i)}$, $i = 1, ..., N$, a particle filter is employed targeting the signal $\{X_t\}_{t=0,1,...}$. The latter particle filters provide approximations of $d(y_t|y_{1:t-1}, \theta^{(i)})$, $i = 1, ..., N$, and, since the grid in the parameter space is fixed, a single sweep over the observations $Y_{1:T} = y_{1:T}$, $T < \infty$, is sufficient, hence the algorithm is recursive. The practical weakness of this approach is that the random grid over the parameter space is generated a priori (irrespective of the observations $Y_{1:T} = y_{1:T}$) and it is not updated as the observations are processed. Therefore, when the prior distribution of Θ differs from the posterior distribution (of Θ conditional on $Y_{1:T} = y_{1:T}$) significantly, a very large number, N, of samples in the parameter space is needed to guarantee a fair performance.

The methodology proposed in (Chopin, Jacob and Papaspiliopoulos, 2013) is also an extension of the IBIS technique. Similarly to the method in (Papavasiliou, 2006), a random grid is created over the parameter space and a particle filter is run for every node in the grid. However, unlike the technique in (Papavasiliou, 2006), the grid of samples in the space of Θ is updated over time, as the batch of observations $Y_{1:T} = y_{1:T}$ is processed. In particular, if $\{\theta_{t-1}^{(i)}, i = 1, ..., N\}$ is the grid at time $t-1$, a particle filter is used to process y_t and then a new grid $\{\theta_t^{(i)}, i = 1, ..., N\}$ can be generated. This filter involves the computation of weights that depend on the densities $d(y_t|y_{1:t-1}, \theta_t^{(i)})$, $i = 1, ..., N$ (similar to the original IBIS). For each point $\Theta = \theta_t^{(i)}$, a particle filter is run to approximate $d(y_t|y_{1:t-1}, \theta_t^{(i)})$. The resulting method is called SMC^2 in (Chopin, Jacob and Papaspiliopoulos, 2013) because of the two nested layers of particle filters. It is more flexible and general than the original IBIS and its extension in (Papavasiliou, 2006), but it is not a recursive algorithm. New samples in the parameter

space are generated by way of particle MCMC (Andrieu, Doucet and Holenstein, 2010) (see below) moves and resampling steps in order to avoid the degeneracy of the particle filter. However, each time a new point in the parameter space is generated at time t, say θ'_t, a new filter has to be run from time 0 to time t. Therefore, the computationally complexity of the method grows quadratically with time. A major advantage of the SMC^2 algorithm is that the approximation errors vanish asymptotically as the number of samples N on the parameter space increases, independently of the number of particles used to approximate the densities $d(y_t|y_{1:t-1}, \theta_t^{(i)})$ in the second layer of particle filters, which can stay fixed. This is shown in (Chopin, Jacob and Papaspiliopoulos, 2013) resorting to a well known unbiasedness property proved in (Del Moral, 2004).

A technique that has quickly gained popularity for parameter estimation is the particle MCMC method of (Andrieu, Doucet and Holenstein, 2010) (employed as a building block for the SMC^2 method described above). It essentially consists in an MCMC algorithm to approximate the posterior distribution of Θ given $Y_{1:t} = y_{1:t}$. Such construction is intractable if addressed directly because the likelihoods $d(y_{1:t}|\theta)$ cannot be conmputed exactly. To circumvent this difficulty, it was proposed in (Andrieu, Doucet and Holenstein, 2010) to use particle filters in order to approximate them. The same trick has been used in the population Monte Carlo (Cappé et al., 2004) framework to tackle the approximation of the posterior distribution of Θ using particles with nonlinearly transformed weights (Koblents and Míguez, 2013). The latter technique has been reported to be computationally more efficient than particle MCMC methods in some examples. These two types of algorithms, as well as the SMC^2 scheme, revolve around the ability to approximate the factors $d(y_t|y_{1:t-1}, \theta)$ using particle filtering.

An alternative, and conceptually simple, approach to compute the likelihood of Θ given $Y_{1:t}$ has been proposed in (Olsson et al., 2008). The problem is addressed by generating a random grid over the parameter space (either random or deterministic, but fixed), then using particle filters to compute the value of the likelihood at each node and finally obtaining an approximation of the whole function by interpolating the nodes. If a point estimate of the parameters is needed, standard optimisation techniques can be applied to the interpolated approximation. Convergence of the L_p error norms is proved in (Olsson et al., 2008) for problems where both the parameter space and the state space are compact.

1.4. Contributions

We introduce a particle filtering method for the approximation of the joint posterior distribution of the signal and the unknown parameters, X_t and Θ, respectively, given the data $Y_{1:t} = y_{1:t}$. Similar to (Papavasiliou, 2006) and (Chopin, Jacob and Papaspiliopoulos, 2013), the algorithm consists of two nested layers of particle filters: an "outer" filter that approximates the probability measure of Θ given the observations and a set of "inner" filters, one per sample generated in the outer filter, that yield approximations of the posterior measures that result for X_t conditional on the observations *and* each specific sample of Θ. The outer filter directly provides an

approximation of the marginal posterior distribution of Θ, whereas a suitable combination of the latter with the outcomes of the inner filters yields an approximation of the joint posterior probability measure of X_t and Θ.

The method is very similar to the SMC^2 scheme of (Chopin, Jacob and Papaspiliopoulos, 2013) in its structure. However, unlike SMC^2, it is a purely recursive procedure and, therefore, it is more suitable for an online implementation. At every time step, all the probability measure approximations (both marginal and joint) are updated recursively, with a fixed computational cost. Also, the jittering of particles in the SMC^2 algorithm is carried out using a particle MCMC kernel (Chopin, Jacob and Papaspiliopoulos, 2013), that leaves the target distribution invariant but cannot be implemented recursively, while the proposed scheme works with simpler Markov kernels easily amenable to online implementations. A detailed comparison between the proposed algorithm and the SMC^2 method of (Chopin, Jacob and Papaspiliopoulos, 2013) is presented in Section 4.3.

The core of the paper is devoted to the analysis of the proposed algorithm. We study the approximation, via the nested particle filtering scheme, of 1-dimensional statistics of the posterior distribution of the system parameters. Under regularity assumptions, we prove that the L_p norms of the approximation errors vanish with rate proportional to $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}$, where N and $N \times M$ are the number of samples in the parameter space and the number of particles in the state space, respectively. This result also holds for the approximation of the joint posterior distribution of the parameters and the state variables.

The analysis builds upon two basic assumptions, which determine the applicability of the algorithm. The most important one is that the optimal filter for the state space model of interest is continuous with respect to (w.r.t.) the parameter θ, i.e., that small changes to the parameter lead to small changes to the posterior probability measure of the state given the available observations. It is this continuity property that makes the implementation of the proposed recursive algorithm feasible and determines some key practical elements of the algorithm, including the magnitude of the jittering of the particles. Non-recursive methods, such as particle MCMC or SMC^2, are not subject to this constraint. The second basic assumption is that the parameter space is a compact set and the the conditional pdf of the observations is well behaved (positive and upper bounded) uniformly over that set. The proposed technique is not guaranteed to work if the parameters have to be searched over an infinite support or, most importantly, if the conditional pdf of the observations has some singularity (e.g., it becomes unbounded) for some parameter values.

To complement the analysis, we also provide a numerical example, where we apply the proposed algorithm to jointly track the state variables and estimate the fixed parameters of a (stochastic version of the) Lorenz 63 system. The length of the observation periods for this example ($\sim 40,000$ discrete time steps) is large enough to make the application of the non-recursive SMC^2 method impractical, while the proposed technique attains accurate estimates of the unknown parameters and tracks the state variables closely.

1.5. Organisation of the paper

We present a general description of the random state-space Markov models of interest in this paper in Section 2, including a brief review of the standard particle filter with known parameters. The recursive nested particle filter scheme is introduced in Section 3. In Section 4 we provide a summary of the main theoretical properties of the proposed algorithm and discuss how it compares to the (non recursive) SMC^2 method of (Chopin, Jacob and Papaspiliopoulos, 2013). The analysis of the approximation errors in L_p is contained in Section 5, together with a brief discussion on the computation of an effective sample size for the proposed algorithm. Section 6 presents some illustrative numerical results for a simple example and, finally, Section 7 is devoted to the conclusions.

2. Background

2.1. Notation and preliminaries

We first introduce some common notation to be used through the paper, broadly classified by topics. Below, $\mathbb{R}$ denotes the real line, while for an integer $d \geq 1$, $\mathbb{R}^d = \overbrace{\mathbb{R} \times \ldots \times \mathbb{R}}^{d \text{ times}}$.

- Functions. Let $S \subseteq \mathbb{R}^d$ be a subset of $\mathbb{R}^d$.
  - The supremum norm of a real function $f : S \to \mathbb{R}$ is denoted as $\|f\|_\infty = \sup_{x \in S} |f(x)|$.
  - $B(S)$ is the set of bounded real functions over S, i.e., $f \in B(S)$ if, and only if, $\|f\|_\infty < \infty$.
- Measures and integrals.
  - $\mathcal{B}(S)$ is the σ-algebra of Borel subsets of S.
  - $\mathcal{P}(S)$ is the set of probability measures over the measurable space $(\mathcal{B}(S), S)$.
  - $(f, \mu) \triangleq \int f(x)\mu(dx)$ is the integral of a real function $f : S \to \mathbb{R}$ w.r.t. a measure $\mu \in \mathcal{P}(S)$.
  - Given a probability measure $\mu \in \mathcal{P}(S)$, a Borel set $A \in \mathcal{B}(S)$ and the indicator function
$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases},$$
$\mu(A) = (I_A, \mu) = \int I_A(x)\mu(dx)$ is the probability of A.
- Sequences, vectors and random variables (r.v.).
  - We use a subscript notation for finite sequences, namely $x_{t_1:t_2} \triangleq \{x_{t_1}, \ldots, x_{t_2}\}$.
  - For an element $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ of an Euclidean space, its norm is denoted as $\|x\| = \sqrt{x_1^2 + \ldots + x_d^2}$.

– Let Z be a r.v. taking values on $\mathbb{R}^d$, with associated probability measure $P \in \mathcal{P}(\mathbb{R}^d)$. The L_p norm of Z, with $p \geq 1$, is $\|Z\|_p \triangleq E[|Z|^p]^{1/p} = \left(\int |z|^p P(dz)\right)^{\frac{1}{p}}$ (where $E[\cdot]$ denotes expectation).

Remark 1. Let $\alpha, \beta, \bar{\alpha}, \bar{\beta} \in \mathcal{P}(S)$ be probability measures and let $f, h \in B(S)$ be two real bounded functions on S such that $(h, \bar{\alpha}) > 0$ and $(h, \bar{\beta}) > 0$. If the identities

$$(f, \alpha) = \frac{(fh, \bar{\alpha})}{(h, \bar{\alpha})} \quad \text{and} \quad (f, \beta) = \frac{(fh, \bar{\beta})}{(h, \bar{\beta})}$$

hold, then it is straightforward to show (see, e.g., (Crisan, 2001)) that

$$|(f, \alpha) - (f, \beta)| \leq \frac{1}{(h, \bar{\alpha})} \left|(fh, \bar{\alpha}) - (fh, \bar{\beta})\right| + \frac{\|f\|_\infty}{(h, \bar{\alpha})} \left|(h, \bar{\alpha}) - (h, \bar{\beta})\right|. \tag{2.1}$$

2.2. State-space Markov models in discrete time

Consider two random sequences, $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 1}$ taking values in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively, and a r.v. Θ taking values on a compact set $D_\theta \subset \mathbb{R}^{d_\theta}$. Let $\mathbb{P}_t$ be the joint probability measure for the triple $(\{X_k\}_{0\leq k\leq t}, \{Y_k\}_{0<k<t}, \Theta)$, that we assume to be absolutely continuous w.r.t. the Lebesgue measure on $\mathcal{B}(\mathbb{R}^{d_x(t+1)} \times \mathbb{R}^{d_y t} \times D_\theta)$.

We refer to the sequence $\{X_t\}_{t\geq 0}$ as the state (or signal) process and we assume that it is an inhomogeneous Markov chain governed by an initial probability measure $\tau_0 \in \mathcal{P}(\mathbb{R}^{d_x})$ and a sequence of transition kernels $\tau_{t,\theta} : \mathcal{B}(\mathbb{R}^{d_x}) \times \mathbb{R}^{d_x} \to [0, 1]$ indexed by a realisation of the r.v. $\Theta = \theta$. To be specific, we define

$$\tau_0(A) \triangleq \mathbb{P}_0\{X_0 \in A\}, \tag{2.2}$$

$$\tau_{t,\theta}(A|x_{t-1}) \triangleq \mathbb{P}_t\{X_t \in A | X_{t-1} = x_{t-1}, \Theta = \theta\}, \quad t \geq 1, \tag{2.3}$$

where $A \in \mathcal{B}(\mathbb{R}^{d_x})$ is a Borel set. The sequence $\{Y_t\}_{t\geq 1}$ is termed the observation process. Each r.v. Y_t is assumed to be conditionally independent of other observations given X_t and Θ, namely

$$\mathbb{P}_t\{Y_t \in A | X_{0:t} = x_{0:t}, \Theta = \theta, \{Y_k = y_k\}_{k\neq t}\} = \mathbb{P}_t\{Y_t \in A | X_t = x_t, \Theta = \theta\}$$

for any $A \in \mathcal{B}(\mathbb{R}^{d_y})$. Additionally, we assume that every probability measure $\gamma_{t,\theta} \in \mathcal{P}(\mathbb{R}^{d_y})$ in the family

$$\gamma_{t,\theta}(A|x_t) \triangleq \mathbb{P}_t\{Y_t \in A | X_t = x_t, \Theta = \theta\}, \quad A \in \mathcal{B}(\mathbb{R}^{d_x}), \quad \theta \in D_\theta, \quad t \geq 1, \tag{2.4}$$

has a nonnegative density w.r.t. the Lebesgue measure. The function $g_{t,\theta}(y|x) \geq 0$ is proportional to this density, hence we write

$$\gamma_{t,\theta}(A|x_t) = \int cI_A(y) g_{t,\theta}(y|x_t) dy, \tag{2.5}$$

where c is a (possibly unknown) normalisation constant, assumed independent of y, x and θ.

The prior τ_0, the kernels $\{\tau_{t,\theta}\}_{t\geq 1}$, and the functions $\{g_{t,\theta}\}_{t\geq 1}$, describe a stochastic Markov state-space model in discrete time. Note that the model is indexed by $\theta \in D_\theta$, which is henceforth termed the system parameter. The a priori probability measure of the r.v. Θ is denoted μ_0, i.e., for any $A \in \mathcal{B}(D_\theta)$, $\mu_0(A) \triangleq \mathbb{P}_0\{\theta \in A\}$.

If $\Theta = \theta$ (the parameter is given), then the stochastic filtering problem consists in the computation of the posterior probability measure of the state X_t given the parameter and a sequence of observations up to time t. Specifically, for a given observation record $\{y_t\}_{t\geq 1}$, we seek the measures

$$\phi_{t,\theta}(A) \triangleq \mathbb{P}_t\left\{X_t \in A | Y_{1:t} = y_{1:t}, \Theta = \theta\right\}, \quad t = 0, 1, 2, ...$$

where $A \in \mathcal{B}(\mathbb{R}^{d_x})$. For many practical problems, the interest actually lies in the computation of statistics of the form $(f, \phi_{t,\theta})$ for some integrable function $f : \mathbb{R}^{d_x} \to \mathbb{R}$. Note that, for $t = 0$, we recover the prior signal measure, i.e., $\phi_{0,\theta} = \tau_0$ independently of θ.

There are many applications in which the parameter Θ is unknown and the goal is to fit the model using a given sequence of observations. In that case, the sequence of probability measures of interest is

$$\mu_t(A) \triangleq \mathbb{P}_t\left\{\Theta \in A | Y_{1:t} = y_{1:t}\right\}, \quad t = 0, 1, 2, ..., \text{ where } A \in \mathcal{B}(D_\theta).$$

If both the fitting of the model and the tracking of the state variables $\{X_t\}_{t\geq 0}$ are sought, then we need to approximate the joint probability measures

$$\pi_t(A \times A') \triangleq \mathbb{P}_t\left\{X_t \in A, \Theta \in A' | Y_{1:t} = y_{1:t}\right\}, \quad t = 0, 1, 2, ...,$$

where $A \in \mathcal{B}(\mathbb{R}^{d_x})$ and $A' \in \mathcal{B}(D_\theta)$. Note that we can write the joint measure π_t as a function of the marginals $\phi_{t,\theta}$ and μ_t. Indeed, if given $A \in \mathcal{B}(\mathbb{R}^{d_x})$ we introduce the real function $\mathsf{f}_t^A : D_\theta \to [0, 1]$, where $\mathsf{f}_t^A(\theta) = \phi_{t,\theta}(A)$, then

$$\pi_t(A \times A') = (I_{A'}\mathsf{f}_t^A, \mu_t) = \int I_{A'}(\theta)\mathsf{f}_t^A(\theta)\mu_t(d\theta) = \int\int I_{A'}(\theta)I_A(x)\phi_{t,\theta}(dx)\mu_t(d\theta). \tag{2.6}$$

2.3. Standard particle filter

Assume that both the parameter $\Theta = \theta$ and a sequence of observations $Y_{1:T} = y_{1:T}$, $T < \infty$, are fixed. Then, the sequence of measures $\{\phi_{t,\theta}\}_{t\geq 1}$ can be numerically approximated using particle filtering. Particle filters are numerical methods based on the recursive relationship between $\phi_{t,\theta}$ and $\phi_{t-1,\theta}$. In particular, let us introduce the predictive measure $\xi_{t,\theta} \triangleq \tau_{t,\theta}\phi_{t-1,\theta}$ such that, for any integrable function $f : \mathbb{R}^{d_x} \to \mathbb{R}$, we obtain

$$(f, \xi_{t,\theta}) = \int\int f(x)\tau_{t,\theta}(dx|x')\phi_{t-1,\theta}(dx') = ((f, \tau_{t,\theta}), \phi_{t-1,\theta}), \tag{2.7}$$

where we note that $\int f(x)\tau_{t,\theta}(dx|x')$ is itself a map $\mathbb{R}^{d_x} \to \mathbb{R}$. Integrals w.r.t. the filter measure $\phi_{t,\theta}$ can be rewritten by way of $\xi_{t,\theta}$ as

$$(f, \phi_{t,\theta}) = \frac{(f g_{t,\theta}^{y_t}, \xi_{t,\theta})}{(g_{t,\theta}^{y_t}, \xi_{t,\theta})}, \tag{2.8}$$

where $g_{t,\theta}^{y_t}(x) \triangleq g_{t,\theta}(y_t|x)$ is the likelihood of $x \in \mathbb{R}^{d_x}$. Eqs. (2.7) and (2.8) are used extensively through the paper. They are instances of the Chapman-Kolmogorov equation and the Bayes theorem, respectively.

The simplest particle filter, often called 'standard particle filter' or 'bootstrap filter' (Gordon, Salmond and Smith, 1993) (see also (Doucet, de Freitas and Gordon, 2001)), can be described as follows.

Algorithm 1. Bootstrap filter conditional on $\Theta = \theta$.

1. Initialisation. At time $t = 0$, draw N i.i.d. samples, $x_0^{(i)}$, $n = 1, \ldots, N$, from the prior τ_0.
2. Recursive step. Let $\{x_{t-1}^{(n)}\}_{1 \leq n \leq N}$ be the particles (Monte Carlo samples) generated at time $t-1$. At time t, proceed with the two steps below.
   (a) For $n = 1, ..., N$, draw a sample $\bar{x}_t^{(n)}$ from the probability distribution $\tau_{t,\theta}(\cdot|x_{t-1}^{(n)})$ and compute the normalised weight

$$w_t^{(n)} = \frac{g_{t,\theta}^{y_t}(\bar{x}_t^{(n)})}{\sum_{k=1}^{N} g_{t,\theta}^{y_t}(\bar{x}_t^{(k)})}. \tag{2.9}$$

   (b) For $n = 1, ..., N$, let $x_t^{(n)} = \bar{x}_t^{(k)}$ with probability $w_t^{(k)}$, $k \in \{1, ..., N\}$.

Step 2.(b) is referred to as resampling or selection. In the form stated here, it reduces to the so-called multinomial resampling algorithm (Doucet, Godsill and Andrieu, 2000; Douc, Cappé and Moulines, 2005) but convergence of the filter can be easily proved for various other schemes (see, e.g., the treatment of the resampling step in (Crisan, 2001)). Using the set $\{x_t^{(n)}\}_{1 \leq n \leq N}$, we construct random approximations of $\xi_{t,\theta}$ and $\phi_{t,\theta}$, namely

$$\xi_{t,\theta}^N(dx_t) = \frac{1}{N}\sum_{n=1}^{N} \delta_{\bar{x}_t^{(n)}}(dx_t) \quad \text{and} \quad \phi_{t,\theta}^N(dx_t) = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_t^{(n)}}(dx_t), \tag{2.10}$$

where $\delta_{x_t^{(n)}}$ is the Dirac delta measure located at $X_t = x_t^{(n)}$. For any integrable function f in the state space, it is straightforward to approximate the integrals $(f, \xi_{t,\theta})$ and $(f, \phi_{t,\theta})$ as

$$(f, \xi_{t,\theta}) \approx (f, \xi_{t,\theta}^N) = \frac{1}{N}\sum_{n=1}^{N} f(\bar{x}_t^{(n)}) \quad \text{and} \quad (f, \phi_{t,\theta}) \approx (f, \phi_{t,\theta}^N) = \frac{1}{N}\sum_{n=1}^{N} f(x_t^{(n)}), \tag{2.11}$$

respectively.

The convergence of particle filters has been analysed in a number of different ways. Here we use results for the convergence of the L_p norms ($p \geq 1$) of the approximation errors.

Theorem 1. Assume that both the system parameter $\Theta = \theta$ and the sequence of observations $Y_{1:T} = y_{1:T}$ are fixed (with $T < \infty$), $g_{t,\theta}^{y_t} \in B(\mathbb{R}^{d_x})$ and $g_{t,\theta}^{y_t} > 0$ (in particular, $(g_{t,\theta}^{y_t}, \xi_{t,\theta}) > 0$) for every $t = 1, 2, ..., T$. Then for any $f \in B(\mathbb{R}^{d_x})$, any $p \geq 1$ and every $t = 1, \ldots, T$,

$$\left\|(f, \xi_{t,\theta}^N) - (f, \xi_{t,\theta})\right\|_p \leq \frac{\bar{c}_{t,\theta}\|f\|_\infty}{\sqrt{N}} \quad \text{and} \quad \left\|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p \leq \frac{c_{t,\theta}\|f\|_\infty}{\sqrt{N}},$$

where $\bar{c}_{t,\theta}, c_{t,\theta} < \infty$ are constants independent of N, $\|f\|_\infty = \sup_{x \in \mathbb{R}^{d_x}} |f(x)|$ and the expectations are taken over the distributions of the random measures $\xi_{t,\theta}^N$ and $\phi_{t,\theta}^N$, respectively.

Proof: This result is a special case of, e.g., Lemma 1 in (Míguez, Crisan and Djurić, 2013). □

Theorem 1 is fairly standard. A similar proposition was already proved in (Del Moral and Miclo, 2000), albeit under additional assumptions on the state-space model, and bounds for $p = 2$ and $p = 4$ can also be found in a number of references (see, e.g., (Crisan, 2001; Crisan and Doucet, 2002; Del Moral, 2004)). It is also possible to establish conditions that make the convergence result of Theorem 1 uniform over the parameter space. Recall that the r.v. Θ has compact support $D_\theta \subset \mathbb{R}^{d_\theta}$ and denote

$$\|g_t^{y_t}\|_\infty \triangleq \sup_{\theta \in D_\theta} \|g_{t,\theta}^{y_t}\|_\infty, \tag{2.12}$$

$$u_t(\theta) \triangleq (g_{t,\theta}^{y_t}, \xi_{t,\theta}) \quad \text{and} \tag{2.13}$$

$$u_{t,\inf} \triangleq \inf_{\theta \in D_\theta} u_t(\theta). \tag{2.14}$$

We can state a result very similar to Theorem 1, but with the constant in the upper bound of the approximation error being independent of the parameter θ. For convenience in the exposition of the rest of the paper, we first establish the convergence, uniform over the parameter space D_θ, of the recursive step in the particle filter.

Lemma 1. Choose any $\theta \in D_\theta$ and any $f \in B(\mathbb{R}^{d_x})$. Assume that the sequence of observations $Y_{1:t} = y_{1:t}$ is fixed (for some $t < \infty$) and a discrete random measure $\phi_{t-1,\theta}^N(dx_{t-1}) = \frac{1}{N}\sum_{n=1}^N \delta_{x_{t-1}^{(n)}}(dx_{t-1})$ is available such that, for any $p \geq 1$,

$$\|(f, \phi_{t-1,\theta}^N) - (f, \phi_{t-1,\theta})\|_p \leq \frac{c_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{t-1}\|f\|_\infty}{\sqrt{M}}, \tag{2.15}$$

where $M \geq 1$ is an integer and $c_{t-1}, \bar{c}_{t-1} < \infty$ are constants independent of N, M and θ.

If $g_{t,\theta}^{y_t} > 0$, $\|g_t^{y_t}\| < \infty$ and $u_{t,\inf} > 0$, then, for any $p \geq 1$,

$$\begin{aligned} \left\|(f, \xi_{t,\theta}^N) - (f, \xi_{t,\theta})\right\|_p &\leq \frac{\tilde{c}_t \|f\|_\infty}{\sqrt{N}} + \frac{\bar{\tilde{c}}_t \|f\|_\infty}{\sqrt{M}} \quad \text{and} \\ \left\|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p &\leq \frac{c_t \|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t \|f\|_\infty}{\sqrt{M}}, \end{aligned}$$

where $\xi_{t,\theta}^N$ and $\phi_{t,\theta}^N$ are computed as in the recursive step of the standard particle filter, $\tilde{c}_t$, $\bar{\tilde{c}}_t$, c_t and $\bar{c}_t$ are finite constants independent of N, M and θ, and the expectations are taken over the distributions of the random measures $\xi_{t,\theta}^N$ and $\phi_{t,\theta}^N$. If $\bar{c}_{t-1} = 0$ then $\bar{c}_t = \bar{\tilde{c}}_t = 0$.

Proof: See Appendix A. □

The (arbitrary) integer M introduced for notational convenience and the error term $\propto \frac{1}{\sqrt{M}}$ plays no role in the proof of Lemma 2 below. It is included exclusively to ease the exposition of some proofs in Section 5. Given Lemma 1, it is straightforward to establish the convergence, uniform over D_θ, of the standard particle filter.

Lemma 2. Assume that the sequence of observations $Y_{1:T} = y_{1:T}$ is fixed (for some $T < \infty$), $g_{t,\theta}^{y_t} > 0$, $\|g_t^{y_t}\| < \infty$ and $u_{t,\inf} > 0$ for every $t = 1, 2, ..., T$. Then, for any $f \in B(\mathbb{R}^{d_x})$, any $\theta \in D_\theta$ and any $p \geq 1$,

$$\left\|(f, \xi_{t,\theta}^N) - (f, \xi_{t,\theta})\right\|_p \leq \frac{\tilde{c}_t \|f\|_\infty}{\sqrt{N}} \quad \text{and} \quad \left\|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p \leq \frac{c_t \|f\|_\infty}{\sqrt{N}}$$

for $t = 0, 1, \ldots, T$, where $\tilde{c}_t(f)$ and $c_t(f)$ are finite constants, independent of both N and θ, and the expectations are taken over the distributions of the random measures $\xi_{t,\theta}^N$ and $\phi_{t,\theta}^N$.

Proof: See Appendix B. □

Remark 2. Lemmas 1 and 2 also hold for any test function $f^\theta : \mathbb{R}^{d_x} \to \mathbb{R}$ (i.e., dependent on θ) as long as the upper bounds

$$\|f\|_\infty = \sup_{\theta \in D_\theta} \|f^\theta\|_\infty, \quad \text{and} \quad \|g_t^{y_t}\|_\infty = \sup_{\theta \in D_\theta} \|g_{t,\theta}^{y_t}\|_\infty$$

are finite and the lower bound $\inf_{\theta \in D_\theta} g_{t,\theta}^{y_t}(x)$ is positive for every $x \in \mathbb{R}^{d_x}$ and every $t = 1, ..., T$. Note that $\inf_{\theta \in D_\theta} g_{t,\theta}^{y_t}(x) > 0$ implies that $u_{t,inf} = \inf_{\theta \in D_\theta} u_t(\theta) > 0$. Under these assumptions the constants c_t and $\bar{c}_t$ in the statement of Lemma 1 are independent of θ (they depend on $u_{t,\inf}$ and $\|g_t^{y_t}\|_\infty$, though).

3. Nested particle filter

3.1. Sequential importance sampling in the parameter space

We aim at devising a recursive algorithm that generates approximations of the posterior probability measures $\mu_t(d\theta)$, $t = 1, 2, ...$, using a sequential importance sampling scheme. The key object needed to attain this goal is the marginal likelihood of the parameter Θ at time t, i.e., the conditional probability density of the observation Y_t given a parameter value $\Theta = \theta$ and a record of observations $Y_{1:t-1} = y_{1:t-1}$.

To be specific, assume that the observations $Y_{1:t-1} = y_{1:t-1}$ are fixed and let

$$\upsilon_{t,\theta}(A) \triangleq \mathbb{P}_t \left\{ Y_t \in A | Y_{1:t-1} = y_{1:t-1}, \Theta = \theta \right\}, \quad A \in \mathcal{B}(\mathbb{R}^{d_y}),$$

be the probability measure associated to the (random) observation Y_t conditional on $Y_{1:t-1} = y_{1:t-1}$ and the parameter vector $\Theta = \theta$. Let us assume that $\upsilon_{t,\theta}$ has a density $u_{t,\theta} : \mathbb{R}^{d_y} \to [0, +\infty)$ w.r.t. the Lebesgue measure, i.e.,

$$\upsilon_{t,\theta}(A) = \int I_A(y) u_{t,\theta}(y) dy, \quad \text{for any } A \in \mathcal{B}(\mathbb{R}^{d_y}).$$

When the actual obsevation $Y_t = y_t$ is collected, the density $u_{t,\theta}(y_t)$ can be evaluated as an integral, namely $u_{t,\theta}(y_t) = (g_{t,\theta}^{y_t}, \xi_{t,\theta})$, and it yields the marginal likelihood of the parameter value θ, denoted as

$$u_t(\theta) \triangleq u_{t,\theta}(y_t) = (g_{t,\theta}^{y_t}, \xi_{t,\theta}).$$

A straightforward Monte Carlo approximation of μ_t could be obtained in two steps, namely,

- drawing N i.i.d. samples $\{\bar{\theta}_t^{(i)}\}_{1 \le i \le N}$ from the posterior measure at time $t-1$, μ_{t-1},
- and then computing normalised importance weights proportional to the marginal likelihoods $u_t(\bar{\theta}_t^{(i)})$.

Unfortunately, neither sampling from μ_{t-1} nor the computation of the likelihood $u_t(\theta)$ can be carried out exactly, hence some approximations are in order.

3.2. Jittering

Let us consider the problem of sampling first. Assume that a particle approximation $\mu_{t-1}^N = \frac{1}{N} \sum_{i=1}^N \delta_{\theta_{t-1}^{(i)}}$ of μ_{t-1} is available. In order to track the variations in μ_t, it is convenient to have a procedure to generate a new set $\{\bar{\theta}_{t-1}^{(i)}\}_{1 \le i \le N}$ which still yields an approximation of μ_{t-1} similar to μ_{t-1}^N. A simple and practically appealing way to

generate the new samples is to mutate the particles $\theta_{t-1}^{(1)}, ..., \theta_{t-1}^{(N)}$ independently using a *jittering* kernel $\kappa_N : \mathcal{B}(D_\theta) \times D_\theta \to [0, 1]$, that we denote as

$$\kappa_N(d\theta|\theta_{t-1}^{(i)}) = \kappa_N^{\theta_{t-1}^{(i)}}(d\theta), \quad i = 1, 2, \ldots, N. \tag{3.1}$$

The subscript N in κ_N indicates that the kernel may depend on the sample size N. This is a key feature in order to keep the distortion introduced by this mutation step sufficiently small, as will be made explicit in Section 5 (see also Section 4.2).

3.3. Conditional bootstrap filter and marginal likelihoods

Let $\bar{\theta}_t^{(i)}$ be a Monte Carlo sample from $\kappa_N(d\theta|\theta_{t-1}^{(i)})$, i.e., a random mutation of $\theta_{t-1}^{(i)}$ as described above. The likelihood $u_t(\bar{\theta}_t^{(i)})$ can be approximated using Algorithm 1 (the standard particle filter), conditional on $\Theta = \bar{\theta}_t^{(i)}$. For notational convenience, we introduce two random transformations of discrete sample sets on $\mathbb{R}^{d_x}$, that will later be used to write down the conditional bootstrap filter.

Definition 1. Let $\{x^{(j)}\}_{1\leq j\leq M}$ be a set of M points on the state space $\mathbb{R}^{d_x}$. The set

$$\{\bar{x}^{(j)}\}_{1\leq j\leq M} = \Upsilon_{n,\theta}\left(\{x^{(j)}\}_{1\leq j\leq M}\right)$$

is obtained by sampling each $\bar{x}^{(j)}$ from the corresponding transition kernel $\tau_{n,\theta}(dx|x^{(j)})$, for $j = 1, ..., M$.

Definition 2. Let $\{\bar{x}^{(j)}\}_{1\leq j\leq M}$ be a set of M points on the state space $\mathbb{R}^{d_x}$. The set

$$\{x^{(j)}\}_{1\leq j\leq M} = \Upsilon_{n,\theta}^{y_n}\left(\{\bar{x}^{(j)}\}_{1\leq j\leq M}\right)$$

is obtained by

- computing normalised weights proportional to the likelihoods,

$$v_n^{(j)} = \frac{g_{n,\theta}^{y_n}(\bar{x}_n^{(j)})}{\sum_{k=1}^M g_{n,\theta}^{y_n}(\bar{x}_n^{(k)})}, \quad j = 1, ..., M.$$

- and then resampling with replacement the set $\{\bar{x}^{(j)}\}_{1\leq j\leq M}$ according to the weights $\{v_n^{(j)}\}_{1\leq j\leq M}$, i.e., assigning $x^{(j)} = \bar{x}^{(k)}$ with probability $v^{(k)}$, for $j = 1, ..., M$ and $k \in \{1, ..., M\}$.

Let us now rewrite the bootstrap filter algorithm using this new notation.

Algorithm 2. Bootstrap filter conditional on $\Theta = \theta_t^{(i)}$.

1. Initialisation. Draw M i.i.d. samples $x_0^{(i,j)}$, $j = 1, ..., M$, from the prior distribution τ_0.
2. Recursive step. Let $\{x_{n-1}^{(i,j)}\}_{1\leq j\leq M}$ be the set of available samples at time $n-1$, with $n \leq t$. The particle set is updated at time n in two steps:
   (a) Compute $\{\bar{x}_n^{(i,j)}\}_{1\leq j\leq M} = \Upsilon_{n,\theta_t^{(i)}}\left(\{x_{n-1}^{(i,j)}\}_{1\leq j\leq M}\right)$.
   (b) Compute $\{x_n^{(i,j)}\}_{1\leq j\leq M} = \Upsilon^{y_n}_{n,\theta_t^{(i)}}\left(\{\bar{x}_n^{(i,j)}\}_{1\leq j\leq M}\right)$.

For $n = t$, we obtain approximations of the posterior measures $\xi_{t,\bar{\theta}_t^{(i)}}(dx_t)$ and $\phi_{t,\bar{\theta}_t^{(i)}}(dx_t)$ of the form

$$\xi^M_{t,\bar{\theta}_t^{(i)}}(dx_t) = \frac{1}{M}\sum_{j=1}^{M}\delta_{\bar{x}_t^{(i,j)}}(dx_t) \quad \text{and} \quad \phi^M_{t,\bar{\theta}_t^{(i)}}(dx_t) = \frac{1}{M}\sum_{j=1}^{M}\delta_{x_t^{(i,j)}}(dx_t), \tag{3.2}$$

respectively, hence the likelihood $u_t(\bar{\theta}_t^{(i)})$ can be approximated as

$$u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}}) = \frac{1}{M}\sum_{j=1}^{M} g^{y_t}_{t,\bar{\theta}_t^{(i)}}(\bar{x}_t^{(i,j)}). \tag{3.3}$$

3.4. Recursive algorithm

If a new sample $\theta_t^{(i)} \in D_\theta$ is produced at time t, one can approximate the likelihood $u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}})$ by running a standard particle filter from time 0 to time t, as shown in Section 3.3. However, the computational cost of this procedure obviously increases with time. We need to avoid this limitation in order to design a recursive algorithm.

Let us assume that the optimal filters $\phi_{t,\theta}(dx)$ are continuous w.r.t the parameter θ, i.e., that if we have two candidate parameters θ and $\tilde{\theta}$ such that $\theta \approx \tilde{\theta}$, then $\phi_{t-1,\theta} \approx \phi_{t-1,\tilde{\theta}}$. If the latter approximation holds, then we can naturally expect that the predictive measure at time t for the parameter $\tilde{\theta}$, namely $\xi_{t,\tilde{\theta}}$, can also be approximated using $\phi_{t-1,\theta}$ instead of $\phi_{t-1,\tilde{\theta}}$. To be specific, we can expect that

$$\xi_{t,\tilde{\theta}} = \tau_{t,\tilde{\theta}}\phi_{t-1,\tilde{\theta}} \approx \tau_{t,\tilde{\theta}}\phi_{t-1,\theta}$$

and, hence, the likelihood of the parameter $u_t(\tilde{\theta}) = (g^{y_t}_{t,\tilde{\theta}}, \xi_{t,\tilde{\theta}})$, can be approximated from the filter at time $t-1$ computed for the *mismatched* parameter value θ (instead of the *actual* $\tilde{\theta}$), i.e.,

$$u_t(\tilde{\theta}) = (g^{y_t}_{t,\tilde{\theta}}, \xi_{t,\tilde{\theta}}) \approx (g^{y_t}_{t,\tilde{\theta}}, \tau_{t,\tilde{\theta}}\phi_{t-1,\theta}). \tag{3.4}$$

If we accept the approximation in Eq. (3.4), then it is possible to devise a truly recursive particle filter for the approximation of the posterior probability measures $\mu_t(d\theta)$.

Assume that, at time $t-1$, we have been able to generate a set of particles in the parameter space $\{\theta_{t-1}^{(i)}\}_{1\leq i\leq N}$ and, for each $\theta_{t-1}^{(i)}$, we have the set of particles in the state space $\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}$. The latter set yields an approximation of the optimal filter conditional on $\theta_{t-1}^{(i)}$, i.e., we have

$$\phi_{t-1,\theta_{t-1}^{(i)}} \approx \phi^M_{t-1,\theta_{t-1}^{(i)}} = \frac{1}{M}\sum_{j=1}^{M} \delta_{x_{t-1}^{(i,j)}}.$$

Now we generate a new parameter sample $\bar{\theta}_t^{(i)}$ by jittering the previous sample $\theta_{t-1}^{(i)}$ in a *controlled* manner (as suggested in Section 3.2). If the modulus of the difference, $\|\bar{\theta}_t^{(i)} - \theta_{t-1}^{(i)}\|$, is small enough, then we can expect that

$$\phi_{t-1,\bar{\theta}_t^{(i)}} \approx \phi_{t-1,\theta_{t-1}^{(i)}} \approx \phi^M_{t-1,\theta_{t-1}^{(i)}} = \frac{1}{M}\sum_{j=1}^{M} \delta_{x_{t-1}^{(i,j)}}, \tag{3.5}$$

i.e., we can use the particle approximation of the filter computed for $\theta_{t-1}^{(i)}$ as a particle approximation of the filter for the new sample $\bar{\theta}_t^{(i)}$. Once we have this approximation, it is straightforward to sample from the Markov kernels $\tau_{t,\bar{\theta}_t^{(i)}}(dx_t|x_{t-1}^{(i,j)})$ (this is the transformation $\Upsilon_{n,\bar{\theta}_t^{(i)}}$ applied to the set $\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}$ from which $\phi^M_{t-1,\theta_{t-1}^{(i)}}$ is constructed) in order to obtain the new predictive measure $\xi^M_{t,\bar{\theta}_t^{(i)}}$ and then approximate the likelihood of $\bar{\theta}_t^{(i)}$ as $u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}})$. In this process, we do not need to run a new particle filter from scratch, but simply to take a recursive step at time t. The price to pay is the introduction of an additional approximation error, that arises from (3.5) and needs to be quantified.

The complete recursive algorithm for the particle approximation of the sequence of measures μ_t is described below.

Algorithm 3. Nested particle filtering for the approximation of μ_t, $t = 0, 1, 2, ...$

1. Initialisation. Draw N i.i.d. samples $\{\theta_0^{(i)}\}_{1\leq i\leq N}$ from the prior distribution $\pi_0(d\theta)$ and $N \times M$ i.i.d. samples $\{x_0^{(i,j)}\}_{1\leq i\leq N; 1\leq j\leq M}$ from the prior distribution τ_0.
2. Recursive step. For $t \geq 1$, assume the particle set $\left\{\theta_{t-1}^{(i)}, \{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}\right\}_{1\leq i\leq N}$ is available and update it taking the following steps.
   (a) For each $i = 1, ..., N$
   - draw $\bar{\theta}_t^{(i)}$ from $\kappa_N^{\theta_{t-1}^{(i)}}(d\theta)$,
   - update $\{\bar{x}_t^{(i,j)}\}_{1\leq j\leq M} = \Upsilon_{t,\bar{\theta}_t^{(i)}}\left(\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}\right)$ and construct $\xi^M_{t,\bar{\theta}_t^{(i)}} = \frac{1}{M}\sum_{j=1}^{M} \delta_{\bar{x}_t^{(i,j)}}$,

– compute the approximate likelihood $u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}})$, and

– update the particle set $\{\tilde{x}_t^{(i,j)}\}_{1\leq j\leq M} = \Upsilon^{y_t}_{t,\bar{\theta}_t^{(i)}}\left(\{\bar{x}_t^{(i,j)}\}_{1\leq j\leq M}\right)$.

(b) Compute normalised weights $w_t^{(i)} \propto u_t^M(\bar{\theta}_t^{(i)})$, $i = 1, ..., N$.

(c) Resample: for each $i = 1, ..., N$, set $\left\{\theta_t^{(i)}, x_t^{(i,j)}\right\}_{1\leq j\leq M} = \left\{\bar{\theta}_t^{(l)}, \tilde{x}_t^{(l,j)}\right\}_{1\leq j\leq M}$ with probability $w_t^{(l)}$, where $l \in \{1, ..., N\}$.

Step 2(a) in Algorithm 3 involves jittering the samples in the parameter space and then taking a single recursive step of a bank of N standard particle filters. In particular, for each $\bar{\theta}_t^{(i)}$, $1 \leq i \leq N$, we have to propagate the particles $\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}$ so as to obtain a new set $\{\tilde{x}_t^{(i,j)}\}_{1\leq j\leq M}$.

Remark 3. The cost of the recursive step in Algorithm 3 is independent of t. We only have to carry out regular 'prediction' and 'update' operations in a bank of standard particle filters. Hence, Algorithm 3 is sequential, purely recursive and can be implemented online.

Remark 4. Algorithm 3 yields several approximations. While $\mu_t^{N,M} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_t^{(i)}}$ is an estimate of μ_t, the joint posterior measure π_t is approximated as $\pi_t^{N,M} = \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M \delta_{\theta_t^{(i)},x_t^{(i,j)}}$. Conditional predictive and filter measures on the state space are also computed by the inner filters, namely $\xi^M_{t,\bar{\theta}_t^{(i)}} = \frac{1}{M}\sum_{j=1}^M \delta_{\bar{x}_t^{(i,j)}}$ and $\phi^M_{t,\theta_t^{(i)}} = \frac{1}{M}\sum_{j=1}^M \delta_{x_t^{(i,j)}}$.

4. Summary of results

4.1. Convergence of the approximation errors in L_p

We pursue a characterisation of the L_p norms of the approximation errors for $\mu_t^{N,M}$, $\phi^M_{t,\theta_t^{(i)}}$ $(i = 1, ..., N)$ and $\pi_t^{N,M}$ which can be stated in a form similar to Lemma 2. Towards this aim, we prove in Section 5 that, under regularity assumptions on the state-space model and the jittering kernel κ_N^θ, the L_p norms of the errors asymptotically decrease toward 0, and provide explicit convergence rates. To be specific, our analysis relies on the following basic assumptions (to be stated in a precise manner in Section 5):

- The optimal filters $\phi_{t,\theta}$ are continuous w.r.t. the parameter θ.
- The jittering steps are "small enough" and, in particular, the variance of the jittering kernel is a decreasing function of the number of particles N.
- The parameter θ is restricted to take values on a compact set D_θ, and the conditional pdf of the observations, $g^{y_t}_{t,\theta}(x_t)$ is positive and uniformly bounded over D_θ.

The continuity of the optimal filters and the constraint on the variance of the jittering kernel are at the core of Algorithm 3. If these conditions are not satisfied, it cannot be expected to converge, as the errors due to the jittering steps may grow without bound. Under the assumptions above, we have proved the results below, that hold true for an arbitrary-but-fixed sequence of observations $y_{1:T}$, with $T < \infty$, and arbitrary test functions $h \in B(D_\theta)$ and $f \in B(D_\theta \times \mathbb{R}^{d_x})$.

Result 1. (Theorem 2, Section 5). There exist constants $c_t, \bar{c}_t < \infty$, independent of N and M, such that

$$\|(h, \mu_t^{N,M}) - (h, \mu_t)\|_p \leq \frac{c_t \|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t \|h\|_\infty}{\sqrt{M}}$$

for any $p \geq 1$ and every $t = 0, \ldots, T$.

Result 2. (Theorem 3, Section 5). There exist constants $c_t, \bar{c}_t < \infty$, independent of N and M, such that

$$\|(f, \pi_t^{N,M}) - (f, \pi_t)\|_p \leq \frac{c_t \|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t \|h\|_\infty}{\sqrt{M}}$$

for any $p \geq 1$ and every $t = 0, \ldots, T$.

Additionally, Algorithm 3 yields explicit approximations of the conditional filter measures (for $\Theta = \theta_t^{(i)}$, $i = 1, ..., N$). In particular, we will show that the statement below also holds under mild assumptions.

Result 3. (Remark 10, Section 5). For any $l \in B(\mathbb{R}^{d_x})$ there exist constants $k_t, \bar{k}_t < \infty$, independent of M and N, such that

$$\sup_{1 \leq i \leq N} \|(l, \phi^M_{t,\theta_t^{(i)}}) - (l, \phi_{t,\theta_t^{(i)}})\|_p \leq \frac{k_t \|l\|_\infty}{\sqrt{N}} + \frac{\bar{k}_t \|l\|_\infty}{\sqrt{M}}$$

for any $p \geq 1$ and every $t = 0, \ldots, T$.

Remark 5. In most practical applications we can expect constraints on the computational effort that can be invested at each time step. Typically, this occurs because a full sequential step of the algorithm must be completed before a new observation is received. This is likely to impose a limitation on the overall number of samples that can be generated, namely the product $K = MN$. For a given value of K (say with integer $\sqrt{K}$), Results 1 and 2 above indicate that the choice of M and N that minimises the error rate is $M = N = \sqrt{K}$. In this case, we obtain approximate measures

$$\hat{\mu}_t^K \triangleq \frac{1}{\sqrt{K}} \sum_{i=1}^{\sqrt{K}} \delta_{\theta_t^{(i)}} \quad \text{and} \quad \hat{\pi}_t^K \triangleq \frac{1}{K} \sum_{i=1}^{\sqrt{K}} \sum_{j=1}^{\sqrt{K}} \delta_{\theta_t^{(i)}, x_t^{(i,j)}}$$

such that

$$\|(h, \hat{\mu}_t^K) - (h, \mu_t)\|_p \leq \frac{\hat{c}_t \|h\|_\infty}{K^{\frac{1}{4}}} \quad \text{and} \quad \|(f, \hat{\pi}_t^K) - (f, \pi_t)\|_p \leq \frac{\hat{c}_t \|f\|_\infty}{K^{\frac{1}{4}}},$$

for any test functions $h \in B(D_\theta)$ and $f \in B(D_\theta \times \mathbb{R}^{d_x})$, and some finite constants $\hat{c}_t$ and $\hat{c}_t$.

4.2. Jittering

The main choice to be made when implementing the algorithm is the type of jittering kernel, as in Eq. (3.1), to be used. This can actually be very simple. Assume for instance a standard Gaussian kernel $\hat{\kappa}^{\theta'}$, with mean θ' and covariance matrix $C = \mathcal{I}_{d_\theta}$, where $\mathcal{I}_{d_\theta}$ is the $d_\theta \times d_\theta$ identity matrix, and let $\kappa^{\theta'}$ the corresponding kernel truncated within the parameter support set D_θ. Any kernel of the form

$$\kappa_N^{\theta'} = (1 - \epsilon_N)\delta_{\theta'} + \epsilon_N \kappa^{\theta'}, \tag{4.1}$$

with $\epsilon_N \leq \frac{1}{N^{\frac{p}{2}}}$ is sufficient to make Results 1 and 2 hold with a prescribed value of p. Note that the choice of κ_N in (4.1) amounts to perturbing each particle with probability ϵ_N (or leave it unchanged with probability $1 - \epsilon_N$). The perturbations applied can be large, but not many particles are actually perturbed.

Alternatively, we can choose a standard Gaussian kernel $\hat{\kappa}_N^{\theta'}$, with mean θ' and covariance matrix $C_N \propto \frac{1}{N^{\frac{p+2}{p}}} \mathcal{I}_{d_\theta}$. The jittering kernel $\kappa_N^{\theta'}$ is then obtained by truncating $\hat{\kappa}_N^{\theta'}$ within the parameter support set D_θ. In this case we perturb every particle, but each single perturbation is small. This choice of κ_N is also sufficient for Results 1 and 2 to hold. See Section 5.1 and Appendix C for a detailed description.

In practice, the magnitude of the jittering introduced by the kernel κ_N is relevant for the performance of the algorithm, because it determines how fast the support of the approximating measure $\mu_t^{N,M}$ can be adapted over time to track changes[2]. If the jittering variance is too small, it may turn out hard to track large changes in the posterior measure μ_t. Such large changes can be expected for small t (when the amount of accumulated data is still limited), in the presence of outliers, due to change-points not accounted for by the model, etc. Some specific techniques can be adapted from (Maíz et al., 2012) to deal with outliers, and we show a simple numerical example at the end of Section 6 to illustrate the effect of change-points. On the other hand, if the jittering variance is made too large, the adaptivity of the algorithm can be improved but its converge rate can be compromised (see Remark 9 in Section 5.2).

[2]The jittering step enables the adaptation of the support set $\{\theta_t^{(i)}\}_{1 \leq N}$. The *shape* of the posterior distribution is tracked by computing the importance weights.

4.3. Comparison with the SMC^2 method

The natural benchmark for the algorithm introduced in this paper is the SMC^2 method of (Chopin, Jacob and Papaspiliopoulos, 2013). This technique is similar in structure to Algorithm 3 and, in particular, it generates and maintains over time N particles in the parameter space and, for each one of them, M particles in the state space. However, it displays two key differences w.r.t. Algorithm 3:

- The particles in the parameter space are jittered using a particle MCMC kernel, with the aim of leaving the approximate posterior distribution of the parameters invariant.
- The weights for the particles in the parameter space at time t are computed using the complete sequence of observations $y_{1:t}$.

The SMC^2 algorithm is consistent (Chopin, Jacob and Papaspiliopoulos, 2013, Proposition 1), as it targets a sequence of probability measures (of increasing dimension) that have the parameter posterior measures, $\{\mu_t\}_{t\geq 0}$, as marginals. Although this is not expicitly proved in (Chopin, Jacob and Papaspiliopoulos, 2013), under adequate assumptions it can be shown that the SMC^2 method produces approximate measures $\mu^{N,M}_{t,SMC}$ such that the L_p norms of the approximation errors can be bounded as

$$\|(h, \mu^{N,M}_{t,SMC} - (h, \mu_t)\|_p \leq \frac{C_t}{\sqrt{N}} \tag{4.2}$$

for some constant C_t, independent of N and M. This implies that the approximation errors vanish asymptotically as $N \to \infty$, even if $M < \infty$ is kept fixed. Also, if $K = NM$ is the total number of particles in the state space generated by the SMC^2 algorithm, and M is assumed to be constant, the the inequality (4.2) implies that the approximation errors converge as $K^{-\frac{1}{2}}$.

The obvious drawback of the SMC^2 method is that it is not recursive: *both* the use of a particle MCMC kernel[3] *and* the computation of the particle weights at time t involve the processing of the whole sequence of observations $y_{1:t}$. In particular, a straightforward implementation of the SMC^2 algorithm with periodic resampling steps and a sequence of T observations, $y_{1:T}$, yields complexity $O(NMT^2)$. In comparison, Algorithm 3 is purely recursive, hence for a sequence of observations $y_{1:T}$ the computational cost is $O(NMT)$, i.e., linear in T versus the quadratic complexity of the original SMC^2 approach.

The linear complexity $O(NMT)$ of Algorithm 3, however, comes at the expense of some limitations compared to the SMC^2 technique. The most important one is that the approximation errors converge with $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}$ (see Result 1), hence we need to let $N \to \infty$ and $M \to \infty$ for the errors to vanish, while in the SMC^2 method it is enough to have $N \to \infty$ (and keep M fixed). If $K = NM$ is the total number of particles in the state space, the optimal allocation for Algorithm 3 is $N = M = \sqrt{K}$ and the convergence rate is $K^{-\frac{1}{4}}$ (see Remark 5) while the SMC^2 attains a rate $K^{-\frac{1}{2}}$.

[3]Particularly note that if we replace the jittering kernel in the proposed Algorithm 3 by a particle MCMC kernel, the resulting procedure is not recursive anymore.

We finally remark that the conditional optimal filters $\phi_{t,\theta}$ need to be continuous w.r.t. $\theta \in D_\theta$ in order to ensure the convergence of Algorithm 3, while this is not necessary for the SMC^2, the particle MCMC (Andrieu, Doucet and Holenstein, 2010) or the nonlinear population Monte Carlo (Koblents and Míguez, 2015) methods. This limitation of the proposed scheme is a direct consequence of not using the full sequence of observations to compute the weights.

5. Convergence analysis

We split the analysis of the recursive Algorithm 3 in three steps: jittering, weight computation and resampling. At the beginning of time step t, the approximation $\mu_{t-1}^{N,M}$ of μ_{t-1} is available. After the jittering step we have a new approximation,

$$\bar{\mu}_{t-1}^{N,M} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{\theta}_t^{(i)}},$$

and we need to prove that it converges to μ_{t-1}. After the computation of the weights, the measure

$$\tilde{\mu}_t^{N,M} = \sum_{i=1}^{N} w_t^{(i)} \delta_{\bar{\theta}_t^{(i)}}$$

is obtained (note that the weights $w_t^{(i)} \propto \left(g_{t,\bar{\theta}_t^{(i)}}^{y_t}, \xi_{t,\bar{\theta}_t^{(i)}}^{M}\right)$ depend on M, although we skip this dependence for notational simplicity) and its convergence toward μ_t must be established. Finally, after the resampling step, we need to prove that

$$\mu_t^{N,M} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_t^{(i)}}$$

converges to μ_t in an appropriate manner. We prove the convergence of $\bar{\mu}_{t-1}^{N}$, $\tilde{\mu}_t^{N}$ and μ_t^{N} in three corresponding lemmas and then combine them to prove the asymptotic convergence of Algorithm 3. Splitting the proof has the advantage that we can "reuse" these partial lemmas easily in order to prove different statements. For example, it is straightforward to show that $\pi_t^{N,M} \to \pi_t$, when $N, M \to \infty$, as well (see Section 5.5).

5.1. Jittering step

In the jittering step, a rejuvenated cloud of particles is generated by propagating the existing samples across the kernels $\kappa_N^{\theta_{t-1}^{(i)}}$, $i = 1, ..., N$. For the analysis, we abide by the following assumption.

A. 1. The family of kernels $\kappa_N^{\theta'}$, $\theta' \in D_\theta$, used in the jittering step satisfy the inequality

$$\sup_{\theta' \in D_\theta} \int |h(\theta) - h(\theta')| \kappa_N^{\theta'}(d\theta) \leq \frac{c_\kappa \|h\|_\infty}{\sqrt{N}} \tag{5.1}$$

for any $h \in B(D_\theta)$ and some constant $c_\kappa < \infty$.

Remark 6. One simple class of kernels that complies with A.1 has the form

$$\kappa_N^{\theta'}(d\theta) = (1 - \epsilon_N)\delta_{\theta'}(d\theta) + \epsilon_N \bar{\kappa}_N^{\theta'}(d\theta), \tag{5.2}$$

where $0 \leq \epsilon_N \leq \frac{1}{\sqrt{N}}$ and $\bar{\kappa}_N^{\theta'} \in \mathcal{P}(D_\theta)$ for every $\theta' \in D_\theta$. Note that substituting (5.2) into (5.1) yields

$$\sup_{\theta' \in D_\theta} \int |h(\theta) - h(\theta')| \kappa_N^{\theta'}(d\theta) \leq 2\epsilon_N \|h\|_\infty \leq \frac{2\|h\|_\infty}{\sqrt{N}},$$

hence A.1 is satisfied with $c_\kappa = 2$.

When using a kernel of the form in (5.2) only a small fraction of particles are actually changed in the jittering step. However, when a particle is actually jittered, the move can be large. Note that the variance of $\bar{\kappa}_N^{\theta'}(d\theta)$ can be independent of N and possibly large, since the variance of $\kappa_N^{\theta'}(d\theta)$ is controlled by the choice of $\epsilon_N \leq \frac{1}{\sqrt{N}}$ alone.

Remark 7. Assume that $h \in B(D_\theta)$ is Lipschitz, i.e., there is a constant $c_L < \infty$ such that

$$|h(\theta) - h(\theta')| \leq c_L \|h\|_\infty \|\theta - \theta'\|$$

for any $\theta, \theta' \in D_\theta$. If there exists a constant $\breve{c} < \infty$ independent of N such that the inequality

$$\sigma_{\kappa,N}^2 = \sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^2 \kappa_N^{\theta'}(d\theta) \leq \frac{\breve{c}}{\epsilon_N^3 N^{\frac{3}{2}}} \tag{5.3}$$

is satisfied, then Eq. (5.1) in A.1 holds with $c_\kappa = c_L \left(1 + \breve{c} \sup_{\theta_1, \theta_2 \in D_\theta} \|\theta_1 - \theta_2\|\right) < \infty$. A generalization of this statement is proved in Appendix C. Note that with this class of kernels every particle is jittered at each time step, but the moves are very small.

Lemma 3. Let $Y_{1:T} = y_{1:T}$ be arbitrary but fixed and choose any $0 < t \leq T$. If $h \in B(D_\theta)$, A.1 holds and

$$\|(h, \mu_{t-1}^{N,M}) - (h, \mu_{t-1})\|_p \leq \frac{c_{t-1}\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{t-1}\|h\|_\infty}{\sqrt{M}} \tag{5.4}$$

for some $p \geq 1$ and some constants $c_{t-1}, \bar{c}_{t-1} < \infty$ independent of N and M, then

$$\|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \mu_{t-1})\|_p \leq \frac{c_{1,t}\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{1,t}\|h\|_\infty}{\sqrt{M}}, \tag{5.5}$$

where the constants $c_{1,t}, \bar{c}_{1,t} < \infty$ are also independent of N and M.

Proof: Recall that we draw the particles $\bar{\theta}_t^{(i)}$, $i = 1, \ldots, N$, independently from the kernels $\kappa_N^{\theta_{t-1}^{(i)}}$, $i = 1, \ldots, N$, respectively. In order to prove that (5.5) holds, we start from the iterated triangle inequality

$$\begin{aligned} \|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \mu_{t-1})\|_p \quad \leq \quad & \|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \kappa_N \mu_{t-1}^{N,M})\|_p \\ & + \|(h, \kappa_N \mu_{t-1}^{N,M}) - (h, \mu_{t-1}^{N,M})\|_p \\ & + \|(h, \mu_{t-1}^{N,M}) - (h, \mu_{t-1})\|_p, \end{aligned} \tag{5.6}$$

where

$$(h, \kappa_N \mu_{t-1}^{N,M}) = \frac{1}{N} \sum_{i=1}^{N} \int h(\theta) \kappa_N^{\theta_{t-1}^{(i)}}(d\theta),$$

and then analyse each of the terms on the right hand side of (5.6) separately. Note that the last term, in particular, is straightforward: its bound follows directly from the assumption in Eq. (5.4).

Let $\mathcal{G}_{t-1}$ be the σ-algebra generated by the random particles $\{\bar{\theta}_{1:t-1}^{(i)}, \theta_{0:t-1}^{(i)}\}_{1 \leq i \leq N}$. Then

$$E\left[(h, \bar{\mu}_{t-1}^{N,M}) | \mathcal{G}_{t-1}\right] = \frac{1}{N} \sum_{i=1}^{N} \int h(\theta) \kappa_N^{\theta_{t-1}^{(i)}}(d\theta) = (h, \kappa_N \mu_{t-1}^{N,M})$$

and the difference $(h, \bar{\mu}_{t-1}^{N,M}) - (h, \kappa_N \mu_{t-1}^{N,M})$ can be written as

$$(h, \bar{\mu}_{t-1}^{N,M}) - (h, \kappa_N \mu_{t-1}^{N,M}) = \frac{1}{N} \sum_{i=1}^{N} \bar{Z}_{t-1}^{(i)},$$

where the random variables $\bar{Z}_{t-1}^{(i)} = h(\bar{\theta}_t^{(i)}) - E[h(\bar{\theta}_t^{(i)}) | \mathcal{G}_{t-1}]$, $i = 1, ..., N$, are conditionally independent (given $\mathcal{G}_{t-1}$), have zero mean and can be bounded as $|\bar{Z}_{t-1}^{(i)}| \leq 2\|h\|_\infty$. It is an exercise in combinatorics to show that the number of non-zero terms in

$$E\left[\left(\sum_{i=1}^{N} \bar{Z}_{t-1}^{(i)}\right)^p | \mathcal{G}_{t-1}\right] = \sum_{i_1} \cdots \sum_{i_p} E\left[\bar{Z}_{t-1}^{(i_1)} \ldots \bar{Z}_{t-1}^{(i_p)} | \mathcal{G}_{t-1}\right]$$

is a polynomial of order no greater than $N^{\frac{p}{2}}$ with coefficients independent of N. As a consequence, there exists a constant $\tilde{c}_1$, independent of N, M and h (actually independent of the distribution of the $\bar{Z}_{t-1}^{(i)}$'s) such that

$$E\left[\left|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \kappa_N \mu_{t-1}^{N,M})\right|^p | \mathcal{G}_{t-1}\right] = E\left[\left|\frac{1}{N} \sum_{i=1}^{N} \bar{Z}_{t-1}^{(i)}\right|^p | \mathcal{G}_{t-1}\right] \leq \frac{\tilde{c}_1^p \|h\|_\infty^p}{N^{\frac{p}{2}}}. \tag{5.7}$$

From (5.7) we readily obtain that

$$\|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \kappa_N \mu_{t-1}^{N,M})\|_p \leq \frac{\tilde{c}_1 \|h\|_\infty}{\sqrt{N}}. \tag{5.8}$$

For the remaining term in (5.6), namely, $\|(h, \kappa_N \mu_{t-1}^{N,M}) - (h, \mu_{t-1}^{N,M})\|_p$, we simply note that

$$\begin{aligned} \left|(h, \kappa_N \mu_{t-1}^{N,M}) - (h, \mu_{t-1}^{N,M})\right| &= \left|\frac{1}{N}\sum_{i=1}^{N} \int \left(h(\theta) - h(\theta_{t-1}^{(i)})\right) \kappa_N^{\theta_{t-1}^{(i)}}(d\theta)\right| \\ &\leq \frac{1}{N}\sum_{i=1}^{N} \int \left|h(\theta) - h(\theta_{t-1}^{(i)})\right| \kappa_N^{\theta_{t-1}^{(i)}}(d\theta) \leq \frac{c_\kappa \|h\|_\infty}{\sqrt{N}}, \end{aligned} \tag{5.9}$$

where the last inequality follows from assumption A.1, with the constant $c_\kappa < \infty$ independent of N and M.

Substituting the inequalities (5.4), (5.8) and (5.9) into Eq. (5.6) yields the desired conclusion, viz., Eq. (5.5), with constants $c_{1,t} = c_{t-1} + c_\kappa + \tilde{c}_1$ and $\bar{c}_{1,t} = \bar{c}_{t-1}$ independent of N and M. □

5.2. Computation of the weights

Since the integral $u_t(\theta) = (g_{t,\theta}^{y_t}, \xi_{t,\theta})$ is intractable, the importance weights are computed as

$$w_t^{(i)} \propto (g_{t,\bar{\theta}_t^{(i)}}^{y_t}, \xi_{t,\bar{\theta}_t^{(i)}}^{M}) = u_t^M(\bar{\theta}_t^{(i)}), \quad i = 1, ..., N.$$

We also recall that the particles in the set $\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}$, which yield the approximate filter $\phi_{t-1,\theta_{t-1}^{(i)}}^{M} = \frac{1}{M}\sum_{j=1}^{M} \delta_{x_{t-1}^{(i,j)}}$, are propagated through the transition kernels as

$$\bar{x}_t^{(i,j)} \sim \tau_{t,\bar{\theta}_t^{(i)}}(dx_t | x_{t-1}^{(i)}), \quad j = 1, \ldots, M, \quad \text{to obtain} \quad \xi_{t,\bar{\theta}_t^{(i)}}^{M} = \frac{1}{M}\sum_{j=1}^{M} \delta_{\bar{x}_t^{(i,j)}}.$$

This means that we are using $\phi_{t-1,\theta_{t-1}^{(i)}}^{M}$ as an estimate of $\phi_{t-1,\bar{\theta}_t^{(i)}}$ in order to compute the predictive measure $\xi_{t,\bar{\theta}_t^{(i)}}^{M}$ and, as a consequence, it is necessary to prove that the error introduced at this step can be bounded in the same way as the approximation errors in Lemma 3. To attain that result, we need to strengthen slightly our assumptions on the structure of the kernel κ_N.

A. 2. The family of kernels $\kappa_N^{\theta'}$, $\theta' \in D_\theta$, used in the jittering step satisfies the inequality

$$\sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^p \kappa_N^{\theta'}(d\theta) \leq \frac{c_\kappa^p}{N^{\frac{p}{2}}} \tag{5.10}$$

for some prescribed $p \geq 1$ and some constant $c_\kappa < \infty$.

Remark 8. It is simple to prove that kernels of the class

$$\kappa_N^{\theta'} = (1-\epsilon_N)\delta_{\theta'} + \epsilon_N \bar{\kappa}_N^{\theta'}, \tag{5.11}$$

with $0 < \epsilon_N \leq \frac{1}{N^{\frac{p}{2}}}$ and $\bar{\kappa}_N^{\theta'} \in \mathcal{P}(D_\theta)$, satisfy assumption A.2 for every $p \geq 1$. Simply note that

$$\sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^p \kappa_N^{\theta'}(d\theta) \leq \epsilon_N \hat{C}^p \leq \frac{\hat{C}^p}{N^{\frac{p}{2}}},$$

where $\hat{C}^p = \sup_{\theta_1,\theta_2 \in D_\theta} \|\theta_1 - \theta_2\|^p < \infty$, since D_θ is compact. The inequality (5.10) also holds for any kernel $\kappa_N^{\theta'}$ that satisfies the inequality

$$\sigma^2_{\kappa,N} = \sup_{\theta,\theta' \in D_\theta} \int \|\theta - \theta'\|^2 \kappa_N^{\theta'}(d\theta) \leq \frac{\breve{c}}{N^{\frac{p+2}{2}}} \tag{5.12}$$

for some constant $\breve{c} < \infty$ (see Appendix C for a generalisation of this result).

In the first case, $\epsilon_N \leq \frac{1}{\sqrt{N}}$, we control the number of particles that are jittered. However, those which are actually jittered may experience large perturbations. In the second case, we allow for the jittering of all particles but, in exchange, the second order moment of the perturbation is controlled. Kernels of the class in (5.11) with $\epsilon_N \leq \frac{1}{\sqrt{N}}$ trivially satisfy A.1. The inequality (5.1) in A.1 also holds for any kernel $\kappa_N^{\theta'}$ that satisfies (5.12) for the prescribed value of p.

Remark 9. It is possible to replace the factor $N^{-\frac{1}{2}}$ in assumptions A.1 and A.2 by some strictly decreasing function of N, say $r(N)$, and still prove the convergence of the nested particle filtering scheme (Algorithm 3). However, the error rates would depend directly on the choice of $r(N)$, so that if $r(N) > N^{-\frac{1}{2}}$, then convergence would be attained at a slower pace (relative to N). If $r(N)$ were chosen to be constant, convergence would not be guaranteed.

Using $\phi^M_{t-1,\theta^{(i)}_{t-1}}$ as an estimate of $\phi_{t-1,\bar{\theta}^{(i)}_t}$ can only work consistently if the filter measure $\phi_{t-1,\theta}$ is continuous in the parameter θ. Here we assume that $\phi_{t-1,\theta}$ is Lipschitz, as stated below.

A. 3. The measures $\phi_{t,\theta}$, $t \geq 1$, are Lipschitz in the parameter $\theta \in D_\theta$. Specifically, for every function $f \in B(\mathbb{R}^{d_x})$ there exists a constant $b_t < \infty$ such that

$$|(f, \phi_{t,\theta'}) - (f, \phi_{t,\theta''})| \leq b_t \|f\|_\infty \|\theta' - \theta''\| \quad \text{for any } \theta', \theta'' \in D_\theta.$$

Assumptions A.2 and A.3 enable us to quantify the error $\|(f, \phi_{t-1,\bar{\theta}^{(i)}_t}) - (f, \phi^M_{t-1,\theta^{(i)}_{t-1}})\|_p$, as made explicit by the following lemma.

Lemma 4. Assume that:

(a) A.3 holds (i.e., $\phi_{t-1,\theta}$ is Lipschitz in θ);
(b) for any $\theta' \in D_\theta$ and $f \in B(\mathbb{R}^{d_x})$, $\phi^M_{t-1,\theta'}$ is a random measure that satisfies the inequality

$$\|(f, \phi^M_{t-1,\theta'}) - (f, \phi_{t-1,\theta'})\|_p \leq \frac{c_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{t-1}\|f\|_\infty}{\sqrt{M}},$$

for some constants $c_{t-1}, \bar{c}_{t-1} < \infty$ independent of N, M and θ'; and
(c) the random parameter θ'' is distributed according to a probability measure $\kappa_N^{\theta'}(d\theta)$ that complies with A.2 for some prescribed $p \geq 1$.

Then, for every $f \in B(\mathbb{R}^{d_x})$ and every $\theta' \in D_\theta$, there exist constants $\tilde{c}_{t-1}, \bar{\tilde{c}}_{t-1} < \infty$, independent of N, M and θ', such that

$$\|(f, \phi^M_{t-1,\theta'}) - (f, \phi_{t-1,\theta''})\|_p \leq \frac{\tilde{c}_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{\tilde{c}}_{t-1}\|f\|_\infty}{\sqrt{M}}.$$

Proof: Consider the triangle inequality

$$\|(f, \phi^M_{t-1,\theta'}) - (f, \phi_{t-1,\theta''})\|_p \leq \|(f, \phi^M_{t-1,\theta'}) - (f, \phi_{t-1,\theta'})\|_p + \|(f, \phi_{t-1,\theta'}) - (f, \phi_{t-1,\theta''})\|_p. \tag{5.13}$$

We aim at bounding the two terms on the right hand side of (5.13).

For the first term, we simply apply assumption (b) in the statement of Lemma 4, which yields

$$\|(f, \phi^M_{t-1,\theta'} - (f, \phi_{t-1,\theta'})\|_p \leq \frac{c_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{t-1}\|f\|_\infty}{\sqrt{M}}, \tag{5.14}$$

where $c_{t-1}, \bar{c}_{t-1} < \infty$ are constants independent of N, M and θ'.

To control the second term on the right hand side of (5.13) we resort to assumption A.3. In particular, note that for any $\theta', \theta'' \in D_\theta$ and any $f \in B(\mathbb{R}^{d_x})$, we have

$$|(f, \phi_{t-1,\theta'}) - (f, \phi_{t-1,\theta''})| \leq b_{t-1}\|f\|_\infty \|\theta' - \theta''\| \tag{5.15}$$

where the constant $b_{t-1} < \infty$ is independent of θ' and θ''. Moreover, if θ'' is random with probability distribution given by $\kappa_N^{\theta'}$, from assumption A.2 we obtain that

$$E\left[\|\theta' - \theta''\|^p\right] \leq \sup_{\theta' \in D_\theta} \int \|\theta' - \theta\|^p \, \kappa_N^{\theta'}(d\theta) \leq \frac{c_\kappa^p}{N^{\frac{p}{2}}}. \tag{5.16}$$

Combining the inequalities (5.15) and (5.16) yields

$$\|(f, \phi_{t-1,\theta'}) - (f, \phi_{t-1,\theta''})\|_p \leq \frac{b_{t-1}c_\kappa\|f\|_\infty}{\sqrt{N}}. \tag{5.17}$$

Finally, substituting (5.17) and (5.14) into the triangle inequality (5.13) completes the proof, with constants $\tilde{c}_{t-1} = c_{t-1} + b_{t-1}c_\kappa$ and $\bar{\tilde{c}}_{t-1} = \bar{c}_{t-1}$. □

Lemma 4 implies that we can "leap" from $\theta^{(i)}_{t-1}$ to $\bar{\theta}^{(i)}_t$ and still keep the associated particle filter in the inner layer running recursively, i.e., we do not have to start it over

every time the particle position in the parameter space changes. If we incorporate some regularity assumptions on the likelihoods $g_{t,\theta}^{y_t}$, $t \geq 1$ (in such a way that we can resort to Lemma 2), then we arrive at an upper bound for the error $\|(h, \tilde{\mu}_t^{N,M}) - (h, \mu_t)\|_p$ after the weight update step. These assumptions are made explicit below.

A. 4. Given a fixed sequence $Y_{1:T} = y_{1:T}$, the family of functions $\{g_{t,\theta}^{y_t}; 1 \leq t \leq T, \theta \in D_\theta\}$ satisfies the following inequalities:

1. $\|g_t^{y_t}\|_\infty = \sup_{\theta \in D_\theta} \|g_{t,\theta}^{y_t}\|_\infty < \infty$ (which implies $\sup_{\theta \in D_\theta} u_t(\theta) = \sup_{\theta \in D_\theta}(g_{t,\theta}^{y_t}, \xi_{t,\theta}) \leq \|g_t^{y_t}\|_\infty$), and
2. $\inf_{\theta \in D_\theta} g_{t,\theta}^{y_t}(x) > 0$ (which implies $u_{t,\inf} = \inf_{\theta \in D_\theta} u_t(\theta) = \inf_{\theta \in D_\theta}(g_{t,\theta}^{y_t}, \xi_{t,\theta}) > 0$)

for every $0 < t \leq T$.

Lemma 5. Let $Y_{1:T} = y_{1:T}$ be fixed and choose any $0 < t \leq T$, any $h \in B(D_\theta)$ and any $f \in B(\mathbb{R}^{d_x})$. Let $p \geq 1$ and assume that A.2, A.3 and A.4 hold. In Algorithm 3, if

$$\|(h, \bar{\mu}_{t-1}^{N,M}) - (h, \mu_{t-1})\|_p \leq \frac{c_{1,t}\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{1,t}\|h\|_\infty}{\sqrt{M}} \tag{5.18}$$

for some constants $c_{1,t}, \bar{c}_{1,t} < \infty$ independent of N and M, and the random measures $\{\phi^M_{t-1,\theta_{t-1}^{(i)}}\}_{1 \leq i \leq N}$ satisfy

$$\sup_{1 \leq i \leq N} \|(f, \phi^M_{t-1,\theta_{t-1}^{(i)}}) - (f, \phi_{t-1,\theta_{t-1}^{(i)}})\|_p \leq \frac{k_{1,t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{k}_{1,t-1}\|f\|_\infty}{\sqrt{M}}, \tag{5.19}$$

for some constants $k_{1,t-1}, \bar{k}_{1,t-1} < \infty$ independent of N and M, then

$$\|(h, \tilde{\mu}_t^{N,M}) - (h, \mu_t)\|_p \leq \frac{c_{2,t}\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{2,t}\|h\|_\infty}{\sqrt{M}}, \tag{5.20}$$

$$\sup_{1 \leq i \leq N} \|(f, \xi^M_{t,\bar{\theta}_t^{(i)}}) - (f, \xi_{t,\bar{\theta}_t^{(i)}})\|_p \leq \frac{\tilde{k}_{2,t}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{\tilde{k}}_{2,t}\|f\|_\infty}{\sqrt{M}}, \tag{5.21}$$

$$\sup_{1 \leq i \leq N} \|(f, \phi^M_{t,\theta_t^{(i)}}) - (f, \phi_{t,\theta_t^{(i)}})\|_p \leq \frac{k_{2,t}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{k}_{2,t}\|f\|_\infty}{\sqrt{M}} \tag{5.22}$$

where the constants $c_{2,t}, \bar{c}_{2,t}, \tilde{k}_{2,t}, \bar{\tilde{k}}_{2,t}, k_{2,t}, \bar{k}_{2,t} < \infty$ are independent of N and M.

Proof: Recall that the particle $\bar{\theta}_t^{(i)}$ is drawn from the kernel $\kappa_N^{\theta_{t-1}^{(i)}}(d\theta)$. Therefore, the inequality (5.19) together with Lemma 4 yields

$$\sup_{1 \leq i \leq N} \|(f, \phi^M_{t-1,\theta_{t-1}^{(i)}}) - (f, \phi_{t-1,\bar{\theta}_t^{(i)}})\|_p \leq \frac{\tilde{c}_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{\tilde{c}}_{t-1}\|f\|_\infty}{\sqrt{M}}, \tag{5.23}$$

where the constants $\tilde{c}_{t-1}, \bar{\tilde{c}}_{t-1} < \infty$ are independent of N, M. However, the key feature of Algorithm 3 is to set the approximation

$$\phi^M_{t-1,\bar{\theta}^{(i)}_t} \triangleq \phi^M_{t-1,\theta^{(i)}_{t-1}} = \frac{1}{M}\sum_{j=1}^{M} \delta_{x^{(i,j)}_{t-1}}, \quad i = 1, ..., N.$$

This choice of $\phi^M_{t-1,\bar{\theta}^{(i)}_t}$, together with the inequality (5.23) and Lemma 1, yields the inequalities (5.21) and (5.22) in the statement of Lemma 5.

Now we address the characterisation of the weights and, therefore, of the approximate measure $\tilde{\mu}^{N,M}_t = \sum_{i=1}^N w^{(i)}_t \delta_{\bar{\theta}^{(i)}_t}$. From the Bayes' theorem, the integral of h w.r.t. μ_t can be written as

$$(h, \mu_t) = \frac{(u_t h, \mu_{t-1})}{(u_t, \mu_{t-1})}, \quad \text{while} \quad (h, \tilde{\mu}^{N,M}_t) = \frac{(u^M_t h, \bar{\mu}^{N,M}_{t-1})}{(u^M_t, \bar{\mu}^{N,M}_{t-1})}. \tag{5.24}$$

Therefore, from the inequality (2.1) we readily obtain

$$\begin{aligned} |(h, \tilde{\mu}^{N,M}_t) - (h, \mu_{t-1})| \quad \leq \quad & \frac{1}{(u_t, \mu_{t-1})}\Big[\|h\|_\infty |(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})| \\ & + |(hu^M_t, \bar{\mu}^{N,M}_{t-1}) - (hu_t, \mu_{t-1})|\Big], \end{aligned} \tag{5.25}$$

and (5.25), together with Minkowski's inequality, yields

$$\begin{aligned} \|(h, \tilde{\mu}^{N,M}_t) - (h, \mu_{t-1})\|_p \quad \leq \quad & \frac{1}{(u_t, \mu_{t-1})}\Big[\|h\|_\infty \|(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})\|_p \\ & + \|(hu^M_t, \bar{\mu}^{N,M}_{t-1}) - (hu_t, \mu_{t-1})\|_p,\Big] \end{aligned} \tag{5.26}$$

where $(u_t, \mu_{t-1}) > 0$ from assumption A.4-2

We need to find upper bounds for the two terms on the right hand side of (5.26). Consider first the term $\|(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})\|_p$. A simple triangle inequality yields

$$\|(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})\|_p \leq \|(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \bar{\mu}^{N,M}_{t-1})\|_p + \|(u_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})\|_p. \tag{5.27}$$

On one hand, since $\sup_{\theta \in D_\theta} |u_t(\theta)| \leq \|g^{y_t}_t\|_\infty < \infty$ (see A.4), it follows from the assumption in Eq. (5.18) that

$$\|(u_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \mu_{t-1})\|_p \leq \frac{c_{1,t}\|g^{y_t}_t\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{1,t}\|g^{y_t}_t\|_\infty}{\sqrt{M}}. \tag{5.28}$$

On the other hand, we may note that

$$\begin{aligned} |(u^M_t, \bar{\mu}^{N,M}_{t-1}) - (u_t, \bar{\mu}^{N,M}_{t-1})|^p \quad &= \quad \left|\frac{1}{N}\sum_{i=1}^N \left(u^M_t(\bar{\theta}^{(i)}_t) - u_t(\bar{\theta}^{(i)}_t)\right)\right|^p \\ &\leq \quad \frac{1}{N}\sum_{i=1}^N |u^M_t(\bar{\theta}^{(i)}_t) - u_t(\bar{\theta}^{(i)}_t)|^p, \end{aligned} \tag{5.29}$$

which is readily obtained from Jensen's inequality. However, the i-th term of the summation above is simply the (p-th power of the) approximation error of the integral $u_t(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi_{t,\bar{\theta}_t^{(i)}})$. Indeed, taking expectations on both sides of the inequality (5.29) yields

$$\begin{aligned} E\left[\left|(u_t^M, \bar{\mu}_{t-1}^{N,M}) - (u_t, \bar{\mu}_{t-1}^{N,M})\right|^p\right] &\leq \frac{1}{N}\sum_{i=1}^{N} E\left[\left|(g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}}) - (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi_{t,\bar{\theta}_t^{(i)}})\right|^p\right] \\ &\leq \frac{1}{N}\sum_{i=1}^{N} \sup_{\theta\in D_\theta}\sup_{i\leq 1\leq N} E\left[\left|(g^{y_t}_{t,\theta}, \xi^M_{t,\bar{\theta}_t^{(i)}}) - (g^{y_t}_{t,\theta}, \xi_{t,\bar{\theta}_t^{(i)}})\right|^p\right] \end{aligned} \tag{5.30}$$

From assumption A.4 we have $\sup_{\theta\in D_\theta}\|g^{y_t}_{t,\theta}\|_\infty \leq \|g^{y_t}_t\|_\infty$ and $\inf_{\theta\in D_\theta} g^{y_t}_{t,\theta}(x) > 0$ for every $t = 1, ..., T$ and every $x \in \mathbb{R}^{d_x}$, hence Lemma 1 (see also Remark 2) readily yields

$$\sup_{\theta\in D_\theta}\sup_{1\leq i\leq N} E\left[\left|(g^{y_t}_{t,\theta}, \xi^M_{t,\bar{\theta}_t^{(i)}}) - (g^{y_t}_{t,\theta}, \xi_{t,\bar{\theta}_t^{(i)}})\right|^p\right] \leq \frac{\hat{k}^p_{2,t}\|g^{y_t}_t\|^p_\infty}{N^{\frac{p}{2}}} + \frac{\bar{\hat{k}}^p_{2,t}\|g^{y_t}_t\|^p_\infty}{M^{\frac{p}{2}}} \tag{5.31}$$

for some finite constants $\hat{k}_{2,t}$ and $\bar{\hat{k}}_{2,t}$ independent of N and M. Substituting (5.31) into (5.30) yields

$$E\left[\left|(u_t^M, \bar{\mu}_{t-1}^{N,M}) - (u_t, \bar{\mu}_{t-1}^{N,M})\right|^p\right] \leq \frac{\hat{k}^p_{2,t}\|g^{y_t}_t\|^p_\infty}{N^{\frac{p}{2}}} + \frac{\bar{\hat{k}}^p_{2,t}\|g^{y_t}_t\|^p_\infty}{M^{\frac{p}{2}}}$$

or, equivalently,

$$\|(u_t^M, \bar{\mu}_{t-1}^{N,M}) - (u_t, \bar{\mu}_{t-1}^{N,M})\|_p \leq \frac{\hat{k}_{2,t}\|g^{y_t}_t\|_\infty}{\sqrt{N}} + \frac{\bar{\hat{k}}_{2,t}\|g^{y_t}_t\|_\infty}{\sqrt{M}}. \tag{5.32}$$

Substituting (5.32) and (5.28) into (5.27) yields

$$\|(u_t^M, \bar{\mu}_{t-1}^{N,M}) - (u_t, \mu_{t-1})\|_p \leq \frac{c'_t\|g^{y_t}_t\|_\infty}{\sqrt{N}} + \frac{\bar{c}'_t\|g^{y_t}_t\|_\infty}{\sqrt{M}}, \tag{5.33}$$

where $c'_t = c_{1,t} + \hat{k}_{2,t}$ and $\bar{c}'_t = \bar{c}_{1,t} + \bar{\hat{k}}_{2,t}$ are constants independent of N and M.

Since $\|hu_t\|_\infty \leq \|h\|_\infty\|g^{y_t}_t\|_\infty$ (the bound is independent of θ), the same argument leading to the bound in (5.33) can be repeated, step by step, on the norm $\|(hu_t^N, \bar{\mu}_{t-1}^N) - (hu_t, \mu_{t-1})\|_p$, to arrive at

$$\|(hu_t^M, \bar{\mu}_{t-1}^{N,M}) - (hu_t, \mu_{t-1})\|_p \leq \frac{c''_t\|h\|_\infty\|g^{y_t}_t\|_\infty}{\sqrt{N}} + \frac{\bar{c}''_t\|h\|_\infty\|g^{y_t}_t\|_\infty}{\sqrt{M}}, \tag{5.34}$$

where $c''_t, \bar{c}''_t < \infty$ are constants independent of N and M.

To complete the proof, we substitute (5.33) and (5.34) back into (5.26) and so obtain

$$\|(h,\tilde{\mu}_t^{N,M})-(h,\mu_{t-1})\|_p \leq \frac{c_{2,t}\|h\|_\infty}{\sqrt{N}}+\frac{\bar{c}_{2,t}\|h\|_\infty}{\sqrt{M}},$$

where the constants $c_{2,t} = \|g_t^{y_t}\|_\infty (c_t' + c_t'')/(u_t, \mu_{t-1}) < \infty$ and $\bar{c}_{2,t} = \|g_t^{y_t}\|_\infty (\bar{c}_t' + \bar{c}_t'')/(u_t, \mu_{t-1}) < \infty$ are independent of N and M. □

5.3. Resampling

We quantify the error in the resampling step 2(c) of Algorithm 3.

Lemma 6. Let the sequence $Y_{1:T} = y_{1:T}$ be fixed and choose any $0 < t \leq T$. If $h \in B(\mathbb{R}^{d_\theta})$ and

$$\|(h,\tilde{\mu}_t^{N,M})-(h,\mu_t)\|_p \leq \frac{c_{2,t}\|h\|_\infty}{\sqrt{N}}+\frac{\bar{c}_{2,t}\|h\|_\infty}{\sqrt{M}} \tag{5.35}$$

for some constants $c_{2,t}, \bar{c}_{2,t} < \infty$ independent of N and M, then

$$\|(h,\mu_t^{N,M})-(h,\mu_t)\|_p \leq \frac{c_{3,t}\|h\|_\infty}{\sqrt{N}}+\frac{\bar{c}_{3,t}\|h\|_\infty}{\sqrt{M}},$$

where the constants $c_{3,t}, \bar{c}_{3,t} < \infty$ are independent of N and M as well.

Proof: The proof of this Lemma is straightforward. The resampling step is the same as in a standard particle filter. See, e.g., the proof of (Míguez, Crisan and Djurić, 2013, Lemma 1) or simply the argument leading from Eq. (A.16) to Eq. (A.19) in Appendix A. □

5.4. Asymptotic convergence of the errors in L_p

Finally, we can put Lemmas 3, 5 and 6 together in order to prove the convergence of the recursive Algorithm 3.

Theorem 2. Let the sequence $Y_{1:T} = y_{1:T}$ be fixed ($T < \infty$), take an arbitrary test function $h \in B(\mathbb{R}^{d_\theta})$, and assume that A.1–A.4 hold. Then, for Algorithm 3,

$$\|(h,\mu_t^{N,M})-(h,\mu_t)\|_p \leq \frac{c_t\|h\|_\infty}{\sqrt{N}}+\frac{\bar{c}_t\|h\|_\infty}{\sqrt{M}}, \quad 1 \leq t \leq T, \tag{5.36}$$

where $\{c_t, \bar{c}_t\}_{0\leq t\leq T}$ is a sequence of constants independent of N and M.

Proof: We prove (5.36) by induction in t. At time $t = 0$, we draw $\theta_0^{(i)}$, $i = 1, ..., N$, independently from the prior μ_0 and it is straightforward to show that $\|(h, \mu_0^{N,M}) -$

$(h,\mu_0)\|_p \leq \frac{c_0\|h\|_\infty}{\sqrt{N}}$, where c_0 does not depend on N. Similarly, for each $i=1,...,N$ we draw M i.i.d. samples $\{x_0^{(i,j)}\}_{1\leq j\leq M}$ from the distribution with measure τ_0 and it is not difficult to check that the random measures $\phi^M_{0,\theta_0^{(i)}} = \frac{1}{M}\sum_{j=1}^M \delta_{x_0^{(i,j)}}$ satisfy

$$\|(f,\phi^M_{0,\theta_0^{(i)}}) - (f,\phi_{0,\theta_0^{(i)}})\|_1 \leq \frac{\bar{k}_0\|f\|_\infty}{\sqrt{M}}$$

for every $i\in\{1,...,N\}$ and any $f\in B(\mathbb{R}^{d_x})$. The constant k_0 is independent of M and $\{\theta_0^{(i)}\}_{1\leq i\leq N}$ (note that $\tau_0=\phi_{0,\theta}$ is actually independent of θ).

Assume that, at time $t-1$,

$$\|(h,\mu^{N,M}_{t-1}) - (h,\mu_{t-1})\|_p \leq \frac{c_{t-1}\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{t-1}\|h\|_\infty}{\sqrt{M}},$$

where $c_{t-1},\bar{c}_{t-1}<\infty$ are independent of N and M, and, for any $f\in B(\mathbb{R}^{d_x})$,

$$\sup_{1\leq i\leq N}\|(f,\phi^M_{t-1,\theta_{t-1}^{(i)}}) - (f,\phi_{t-1,\theta_{t-1}^{(i)}})\|_p \leq \frac{k_{t-1}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{k}_{t-1}\|f\|_\infty}{\sqrt{M}},$$

where $k_{t-1},\bar{k}_{t-1}<\infty$ are constants independent of N and M. Then, we simply "concatenate" Lemmas 3, 5 and 6 (in that order) to obtain

$$\begin{aligned}\|(h,\mu_t^{N,M}) - (h,\mu_t)\|_p &\leq \frac{c_t\|h\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t\|h\|_\infty}{\sqrt{M}},\\ \sup_{1\leq i\leq N}\|(f,\phi^M_{t,\theta_t^{(i)}}) - (f,\phi_{t,\theta_t^{(i)}})\|_p &\leq \frac{k_t\|f\|_\infty}{\sqrt{N}} + \frac{\bar{k}_t\|f\|_\infty}{\sqrt{M}},\end{aligned} \tag{5.37}$$

for some constants $c_t,\bar{c}_t,k_t,\bar{k}_t<\infty$ independent of N and M. □

Remark 10. The argument of the proof of Theorem 2 also yields, as a by-product, error rates for the (approximate) conditional filters $\phi^M_{t,\theta_t^{(i)}}$ computed for each particle in the parameter space, as shown by the inequality in (5.37). These rates are uniform for any $\theta\in D_\theta$.

5.5. Approximation of the joint measure π_t

Integrals w.r.t. the joint measure π_t introduced in (2.6) can be written naturally in terms of the marginal measures $\phi_{t,\theta}$ and μ_t. To be specific, choose any integrable function $f: D_\theta\times\mathbb{R}^{d_x}\to\mathbb{R}$ and define $f^\theta:\mathbb{R}^{d_x}\to\mathbb{R}$, where $f^\theta(x_t)\triangleq f(\theta,x_t)$, and $\mathsf{f}_t: D_\theta\to\mathbb{R}$, where $\mathsf{f}_t(\theta)\triangleq\int f^\theta(x_t)\phi_{t,\theta}(dx_t)=(f^\theta,\phi_{t,\theta})$. Then we can write

$$(f,\pi_t) = \int\int f(\theta,x_t)\pi_t(d\theta,dx_t) = \int \mathsf{f}_t(\theta)\mu_t(d\theta) = (\mathsf{f}_t,\mu_t). \tag{5.38}$$

It is straightforward to approximate π_t as

$$\pi_t^{N,M}(d\theta \times dx_t) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{\theta_t^{(i)},x_t^{(i,j)}}(d\theta \times dx_t),$$

which yields

$$(f, \pi_t^{N,M}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f(\theta_t^{(i)}, x_t^{(i,j)}) = (\mathsf{f}_t^M, \mu_t^N), \tag{5.39}$$

where $\mathsf{f}_t^M(\theta_t^{(i)}) = (f^{\theta_t^{(i)}}, \phi^M_{t,\theta_t^{(i)}})$.

It is relatively easy to use the results obtained earlier in this Section in order to show that, for any $f \in B(D_\theta \times \mathbb{R}^{d_x})$, the L_p error norm $\|(f, \pi_t^{N,M}) - (f, \pi_t)\|_p$ has an upper bound of order $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}$.

Theorem 3. Let the sequence $Y_{1:T} = y_{1:T}$ be fixed, take an arbitrary test function $f \in B(D_\theta \times \mathbb{R}^{d_\theta})$ and assume that A.1–A.4 hold. Then, for any $p \geq 1$, Algorithm 3 yields

$$\|(f, \pi_t^{N,M}) - (f, \pi_t)\|_p \leq \frac{c_t\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t\|f\|_\infty}{\sqrt{M}}, \quad 1 \leq t \leq T, \tag{5.40}$$

where $\{c_t, \bar{c}_t\}_{1\leq t\leq T}$ is a sequence of finite constants independent of N and M.

Proof: From Eqs. (5.38) and (5.39), $(f, \pi_t^{N,M}) - (f, \pi_t) = (\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t)$ and a triangle inequality yields

$$\|(\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t)\|_p \leq \|(\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t^{N,M})\|_p + \|(\mathsf{f}_t, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t)\|_p. \tag{5.41}$$

Since $\mathsf{f}_t \in B(D_\theta)$ (namely, $\|\mathsf{f}_t\|_\infty \leq \|f\|_\infty$), Theorem 2 yields a bound for the second term on the right hand side of (5.41), i.e.,

$$\|(\mathsf{f}_t, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t)\|_p \leq \frac{\hat{c}_t\|f\|_\infty}{\sqrt{N}} + \frac{\bar{\hat{c}}_t\|f\|_\infty}{\sqrt{M}}, \tag{5.42}$$

where $\hat{c}_t, \bar{\hat{c}}_t < \infty$ are constants independent of N and M.

In order to control the first term on the right hand side of (5.41), we note that

$$\begin{aligned} E\left[\left|(\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t^{N,M})\right|^p\right] &\leq \frac{1}{N}\sum_{i=1}^{N} E\left[\left|(f^{\theta_t^{(i)}}, \phi^M_{t,\theta_t^{(i)}}) - (f^{\theta_t^{(i)}}, \phi_{t,\theta_t^{(i)}})\right|^p\right] \quad (5.43)\\ &\leq \sup_{\theta\in D_\theta}\sup_{1\leq i\leq N} E\left[\left|(f^\theta, \phi^M_{t,\theta_t^{(i)}}) - (f^\theta, \phi_{t,\theta_t^{(i)}})\right|^p\right], \end{aligned}$$

where (5.43) follows from Jensen's inequality. However, since $f^\theta \leq \|f\|_\infty < \infty$, we can resort to Remark 10 in order to obtain

$$\sup_{1\leq i\leq N} E\left[\left|(f^\theta, \phi^N_{t,\theta_t^{(i)}}) - (f^\theta, \phi_{t,\theta_t^{(i)}})\right|^p\right] \leq \frac{k_t^p\|f\|_\infty^p}{N^{\frac{p}{2}}} + \frac{\bar{k}_t^p\|f\|_\infty^p}{M^{\frac{p}{2}}},$$

where the constants $k_t, \bar{k}_t < \infty$ are independent of N and M. Since the latter upper bound is uniform over D_θ (recall Remark 2), it follows that

$$\begin{aligned} E\left[\left|(\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t^{N,M})\right|^p\right] &\leq \sup_{\theta\in D_\theta} \sup_{1\leq i\leq N} E\left[\left|(f^\theta, \phi^N_{t,\theta_t^{(i)}}) - (f^\theta, \phi_{t,\theta_t^{(i)}})\right|^p\right] \\ &\leq \frac{k_t^p \|f\|_\infty^p}{N^{\frac{p}{2}}} + \frac{\bar{k}_t^p \|f\|_\infty^p}{M^{\frac{p}{2}}} \end{aligned}$$

as well or, equivalently,

$$\|(\mathsf{f}_t^M, \mu_t^{N,M}) - (\mathsf{f}_t, \mu_t^{N,M})\|_p \leq \frac{k_t \|f\|_\infty}{\sqrt{N}} + \frac{\bar{k}_t \|f\|_\infty}{\sqrt{M}}. \tag{5.44}$$

Substituting (5.44) and (5.42) into the triangle inequality (5.41) yields the desired result, with constants $c_t = \hat{c}_t + k_t$ and $\bar{c}_t = \bar{\hat{c}}_t + \bar{k}_t$, $1 \leq t \leq T$, independent of N and M. □

5.6. Effective sample size

After completing all operations at time $t-1$, Algorithm 3 produces a system of particles $\{\theta_{t-1}^{(i)}\}_{1\leq i\leq N}$, where many of its elements may be located at the same position in the parameter space because of the resampling step. At time t, the first operation of Algorithm 3 is the jittering of the particles in order to restore their diversity. After jittering, the new system $\{\bar{\theta}_t^{(i)}\}_{1\leq i\leq N}$ is available. However, depending on the choice of kernel κ_N, it is possible that *not* every particle in $\{\theta_{t-1}^{(i)}\}_{1\leq i\leq N}$ has actually been changed, hence the jittered system $\{\bar{\theta}_t^{(i)}\}_{1\leq i\leq N}$ may still contain replicated elements, i.e., particles with different indices that correspond to the same position in the parameter space D_θ.

Let $\hat{N}_t$ denote the number of distinct particles in the system $\{\bar{\theta}_t^{(i)}\}_{1\leq i\leq N}$ and let $\{\tilde{\theta}_t^{(i)}\}_{1\leq i\leq \hat{N}_t}$ be the set of those distinct particles. Obviously, $1 \leq \hat{N}_t \leq N$. We use $n_t^{(i)}$ to denote the number of replicas of $\tilde{\theta}_t^{(i)}$ included in the original system $\{\bar{\theta}_t^{(i)}\}_{1\leq i\leq N}$. It is straightforward to check that, for every $i = 1, ..., \hat{N}_t$,

$$1 \leq n_t^{(i)} \leq N - \hat{N}_t + 1,$$

while $\sum_{i=1}^{\hat{N}_t} n_t^{(i)} = N$.

The size of the set $\{\tilde{\theta}_t^{(i)}\}_{1\leq i\leq \hat{N}_t}$ is particularly relevant to the computation of the so-called effective sample size (ESS) (Kong, Liu and Wong, 1994) (see also (Doucet, Godsill and Andrieu, 2000)) of the particle approximation produced by Algorithm 3. The ESS, which is commonly used to assess the numerical stability of particle filters (Chopin, Jacob and Papaspiliopoulos, 2013; Beskos et al., 2014), was defined in (Kong, Liu and Wong, 1994) as

$$\mathrm{ESS}_t(N) = \frac{N}{1 + V_t^2},$$

where V_t^2 denotes the variance of the non-normalised importance weights (namely, the variance of $u_t^M(\theta)$ in the case of Algorithm 3). Since this variance cannot be computed in closed form, the ESS has to be estimated and the most commonly used estimator takes the form (Kong, Liu and Wong, 1994; Doucet, Godsill and Andrieu, 2000)

$$\begin{aligned}\widehat{\mathrm{ESS}}_t(N) &= \frac{1}{\sum_{i=1}^N w_t^{(i)2}} \\ &= \frac{\left(\sum_{i=1}^N u_t^M(\bar{\theta}_t^{(i)})\right)^2}{\sum_{i=1}^N u_t^M(\bar{\theta}_t^{(i)})^2} \qquad (5.45) \\ &= \frac{\left(\sum_{i=1}^{\hat{N}_t} n_t^{(i)} u_t^M(\tilde{\theta}_t^{(i)})\right)^2}{\sum_{i=1}^{\hat{N}_t} n_t^{(i)} u_t^M(\tilde{\theta}_t^{(i)})^2}, \qquad (5.46)\end{aligned}$$

where (5.45) follows from the construction of the normalised weights in Algorithm 3 and in (5.46) we write the estimator explicitly in terms of the system of distinct particles[4] $\{\tilde{\theta}_t^{(i)}\}_{1\leq i\leq \hat{N}_t}$.

The estimator of the ESS in Eq. (5.46) takes values between 1 and N, with 1 being the worst and N being the best outcome. However, it can become uninformative when we actually have replicated particles, i.e., when $\hat{N}_t < N$. To see the problem, let us consider the extreme case in which $\hat{N}_t = 1$ and, as a consequence, $n_t^{(1)} = N$. If we substitute these values in (5.46) and realise that $\sum_{i=1}^{\hat{N}_t} n_t^{(i)} u_t^M(\tilde{\theta}_t^{(i)}) = N u_t^M(\tilde{\theta}_t^{(1)})$, then we readily obtain that $\widehat{\mathrm{ESS}}_t(N) = N$. This seems to indicate that we have an "optimal" set of particles, as the maximum ESS is attained, when it is actually a fully degenerate set with one single particle replicated N times. This difficulty does not arise in standard particle filtering applications because the ESS is typically estimated after the weight update step, before resampling, when all particles are different with probability 1.

To overcome this problem, we propose to use a different estimator of the ESS. Recall that $w_t^{(i)} = \frac{u_t^M(\bar{\theta}_t^{(i)})}{\sum_{k=1}^N u_t^M(\bar{\theta}_t^{(k)})}$, $1 \leq i \leq N$, are the normalised weights. When there are multiple samples at the same position in D_θ, the resulting probability measure

$$\mu_t^{N,M} = \sum_{i=1}^N w_t^{(i)} \delta_{\bar{\theta}_t^{(i)}}$$

can be rewritten as

$$\mu_t^{N,M} = \sum_{i=1}^{\hat{N}_t} v_t^{(i)} \delta_{\tilde{\theta}_t^{(i)}}, \qquad (5.47)$$

[4] We assume that the algorithm is implemented efficiently, meaning that when a subset of particles is found to correspond to the same position in the parameter space the likelihood of that position is estimated only once. In other words, if we have indices $i_0, i_1, \ldots, i_{n_t^{(i_0)}}$ such that $\bar{\theta}_t^{(i_0)} = \bar{\theta}_t^{(i_1)} = \ldots = \bar{\theta}_t^{\left(i_{n_t^{(i_o)}}\right)}$, then we compute $u_t^M(\tilde{\theta}_t^{(i_0)})$ only once.

where $v_t^{(i)} = n_t^{(i)} w_t^{(i)}$ is the probability mass that $\mu_t^{N,M}$ allocates at position $\tilde{\theta}_t^{(i)}$. If we are given $\mu_t^{N,M}$ in the form of (5.47), a fairly natural estimator the ESS is

$$\overline{\mathrm{ESS}}_t(N) = \frac{1}{\sum_{i=1}^{\hat{N}_t} \left(v_t^{(i)}\right)^2} = \frac{\left(\sum_{k=1}^{N} u_t^M(\bar{\theta}_t^{(k)})\right)^2}{\sum_{i=}^{\hat{N}_t} \left(n_t^{(i)} u_t^M(\tilde{\theta}_t^{(i)})\right)^2} \tag{5.48}$$

where we note that $\sum_{k=1}^{\hat{N}_t} n_t^{(i)} u_t(\tilde{\theta}_t^{(k)}) = \sum_{k=1}^{N} u_t(\bar{\theta}_t^{(k)})$.

When all the particles are distinct, $\hat{N}_t = N$ and $n_t^{(i)} = 1$ for every i, the estimator in (5.48) reduces to the standard one in (5.46). On the other hand, when $\hat{N}_t = 1$ and $n_t^{(1)} = N$, the formula in (5.48) yields $\overline{\mathrm{ESS}}_t(N) = 1$, which is the minimal ESS and the expected result in this fully degenerate case. We recall that $\widehat{\mathrm{ESS}}_t(N) = N$ in the same scenario. Finally, if we divide the expression in (5.48) by N then we obtain an estimate of the normalised ESS (NESS) (Doucet, Godsill and Andrieu, 2000) of the form

$$\overline{\mathrm{NESS}}_t(N) = \frac{\left(\sum_{k=1}^{N} u_t^M(\bar{\theta}_t^{(k)})\right)^2}{N \sum_{i=1}^{\hat{N}_t} \left(n_t^{(i)} u_t^M(\tilde{\theta}_t^{(i)})\right)^2} \tag{5.49}$$

that takes values in the interval $[N^{-1}, 1]$.

6. A numerical example

Let us consider the problem of jointly tracking the dynamic variables and estimating the fixed parameters of a 3-dimensional Lorenz system (Lorenz, 1963) with additive dynamical noise and partial observations (Chorin and Krause, 2004). To be specific, consider a 3-dimensional stochastic process $\{X(s)\}_{s\in(0,\infty)}$ taking values on $\mathbb{R}^3$, whose dynamics is described by the system of stochastic differential equations

$$dX_1 = -S(X_1 - Y_1) + dW_1, \quad dX_2 = RX_1 - X_2 - X_1X_3 + dW_2, \quad dX_3 = X_1X_2 - BX_3 + dW_3,$$

where $\{W_i(s)\}_{s\in(0,\infty)}$, $i = 1, 2, 3$, are independent 1-dimensional Wiener processes and $(S, R, B) \in \mathbb{R}$ are static model parameters. A discrete-time version of the latter system using the Euler-Maruyama method with integration step $\mathsf{T}_e > 0$ is straightforward to obtain and yields the model

$$X_{1,t} = X_{1,t-1} - \mathsf{T}_e S(X_{1,t-1} - X_{2,t-1}) + \sqrt{\mathsf{T}_e} U_{1,t}, \tag{6.1}$$

$$X_{2,t} = X_{2,t-1} + \mathsf{T}_e(RX_{1,t-1} - X_{2,t-1} - X_{1,t-1}X_{3,t-1}) + \sqrt{\mathsf{T}_e} U_{2,t}, \tag{6.2}$$

$$X_{3,t} = X_{3,t-1} + \mathsf{T}_e(X_{1,t-1}X_{2,t-1} - BX_{3,t-1}) + \sqrt{\mathsf{T}_e} U_{3,t}, \tag{6.3}$$

where $\{U_{i,t}\}_{t=0,1,\ldots}$, $i = 1, 2, 3$, are independent sequences of i.i.d. normal random variables with 0 mean and variance 1. System (6.1)-(6.3) is partially observed every

40 discrete-time steps, i.e., we collect a sequence of 2-dimensional observations $\{Y_n = (Y_{1,n}, Y_{3,n})\}_{n=1,2,...}$, of the form

$$Y_{1,n} = k_o X_{1,40n} + V_{1,n}, \quad Y_{3,n} = k_o X_{3,40n} + V_{3,n}, \tag{6.4}$$

where $k_o > 0$ is a fixed scale parameter and $\{V_{i,n}\}_{n=1,2,...}$, $i = 1, 3$, are independent sequences of i.i.d. normal random variables with zero mean and variance $\sigma^2 = \frac{1}{10}$.

Let $X_t = (X_{1,t}, X_{2,t}, X_{3,t})$ be the state vector, let $Y_n = (Y_{1,n}, Y_{3,n})$ be the observation vector and let $\Theta = (S, R, B, k_o)$ be the set of model parameters to be estimated. The dynamic model given by Eqs. (6.1)–(6.3) yields the family of kernels $\tau_{t,\theta}(dx|x_{t-1})$ and the observation model of Eq. (6.4) yields the likelihood function $g_{n,\theta}^{y_n}(x_n)$, both in a straightforward manner. The goal is to track the sequence of joint posterior probability measures π_n, $n = 1, 2, ...$, for $\{\hat{X}_n, \Theta\}_{n=1,...}$, where $\hat{X}_n = X_{40n}$. Note that one can draw a sample $\hat{X}_n = \hat{x}_n$ conditional on some θ and $\hat{X}_{n-1} = \hat{x}_{n-1}$ by successively simulating

$$\tilde{x}_t \sim \tau_{t,\theta}(dx|\tilde{x}_{t-1}), \quad t = 40(n-1) + 1, ..., 40n,$$

where $\tilde{x}_{40(n-1)} = \hat{x}_{n-1}$ and $\hat{x}_n = \tilde{x}_{40n}$. For the sake of the example, the prior probability measure for the parameters, $\mu_0(d\theta)$, is chosen to be uniform, namely

$$S \sim \mathcal{U}(5, 20), \quad R \sim \mathcal{U}(18, 50), \quad R \sim \mathcal{U}(1, 8) \quad \text{and} \quad k_o \in \mathcal{U}(0.5, 3),$$

where $\mathcal{U}(a, b)$ is the uniform probability distribution in the interval (a, b). The prior measure for the state variables is normal, namely $X_0 \sim \mathcal{N}(x_*, v_0^2 \mathcal{I}_3)$, where $x_* = (-5.91652; -5.52332; 24.5723)$ is the mean and $v_0^2 \mathcal{I}_3$ is the covariance matrix, with $v_0^2 = 10$. (The value x_* is taken from a typical run of the deterministic Lorenz 63 model, once in its stationary regime.)

We have applied the nested particle filter (Algorithm 3), with $N = M$ (i.e., the same number of particles in the outer and inner filters, following Remark 5), to estimate the fixed parameters S, R, B and k_o. Besides selecting the total number of particles $K = NM$, the only "tuning" necessary for the algorithm is the choice of the jittering kernel. One of the simplest possible choices is to jitter each parameter independently from the others, using Gaussian distributions truncated to fit the support of each parameter. To be specific, let $\mathsf{TN}(\mu, \sigma^2, A, B)$ denote the Gaussian distribution with mean μ and variance σ^2 truncated to have support on the interval (A, B), i.e., the distribution with pdf

$$p_{\mathsf{TN}}(x; \mu, \sigma^2, A, B) = \frac{\exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}}{\int_A^B \exp\left\{\frac{1}{2\sigma^2}(z-\mu)^2\right\} dz}.$$

We choose the jittering kernel $\kappa_N^{\theta'}$, with $\theta' = (S', R', B', k_o')$, to be the conditional probability distribution with density

$$\begin{aligned} \kappa_N^{S',R',B',k_o'}(S, R, B, k_o) \quad = \quad & p_{\mathsf{TN}}(S; S', \sigma_{N,S}^2, 5, 20) \times p_{\mathsf{TN}}(R; R', \sigma_{N,R}^2, 18, 50) \\ & \times p_{\mathsf{TN}}(B; B', \sigma_{N,B}^2, 1, 8) \times p_{\mathsf{TN}}(k_o; k_o', \sigma_{N,k_o}^2, 0.5, 3). \end{aligned}$$

This choice of kernel is possibly far from optimal (in terms or estimation accuracy) but it is simple and enables us to show that Algorithm 3 works without having to fit a sophisticated kernel.

If we are merely interested in estimating the parameter values, then the test function $h \in B(D_\theta)$ in Theorem 2 is simply the projection of the parameter vector on the desired component, i.e., for $\theta = (\theta_1, ..., \theta_4) = (S, R, B, k_o)$ we are interested in the functions $h_i(\theta) = \theta_i, \quad i = 1, ..., 4$. Therefore, the estimator of the parameter θ_i at time t has the form

$$\theta_{i,t}^{N,N} = (h_i, \mu_t^{N,N}) = \frac{1}{N} \sum_{j=1}^{N} h_i(\theta_t^{(j)}), \quad i = 1, ..., 4.$$

Furthermore, if we aim at the minimising the L_1 errors, $E\left[|\theta_{i,t}^N - \theta_i|\right]$, Proposition 1 in Appendix C shows that it is enough to choose the jittering variances as

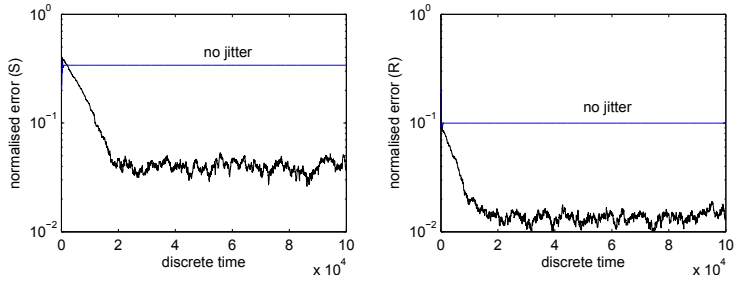
$$(\sigma_{N,S}^2, \sigma_{N,R}^2, \sigma_{N,B}^2, \sigma_{N,k_o}^2) = \frac{1}{N^{\frac{3}{2}}}(c_S, c_R, c_B, c_{k_o})$$

for arbitrary positive constants c_S, c_R, c_B and c_{k_o} in order to satisfy the assumptions A.1 and A.2. For the simulations in this section we have set $(c_S, c_R, c_B, c_{k_o}) = (60, 60, 10, 1)$ (we roughly choose bigger constants for the parameters with bigger support).

Figure 1 shows the average, over 50 independent simulations, of the normalised absolute errors $|\theta_{i,t}^{N,N} - \theta_i|/\theta_i$ versus continuous time when we run Algorithm 3 with $N = M = 300$. The figure shows how the errors converge over time (as μ_t concentrates around the true value $\theta = (10, 28, 8/3, 0.8)$). We have also included the errors attained by a modified version of Algorithm 3 in which the jittering step is removed. It is seen that the particle representation of μ_t soon collapses and the algorithm *without* jittering turns out unable to estimate the parameters. The integration period for all the simulations shown in this section is $\mathsf{T}_e = 10^{-3}$, hence 100×10^3 discrete-time steps amount to 100 continuous time units. Observations are collected every 40 discrete steps. Even for this relatively simple system, running a non-recursive algorithm such as SMC^2 becomes impractical (recall that the computational complexity of the SMC^2 method increases quadratically with the number of discrete-time steps).

In Figure 2 we plot the average of the normalised errors versus the number of particles in Algorithm 3 (namely, for $N = 150, 300, 600$). We have carried out 20 independent simulation trials (per point in the plot). In each simulation, the Lorenz system is run from continous time 0 to 24 (i.e., 24,000 discrete time steps), with the errors computed by averaging $|\theta_{i,t}^{N,N} - \theta_i|/\theta_i$ over the continuous time interval (22,24). As in Figure 1, the performace of Algorithm 3 with the jittering step removed is also displayed, and again we observe how it fails to yield accurate parameter estimates. For the outputs of Algorithm 3 *with* jittering, we also display a least squares fit of the function $e(N) = \frac{c}{\sqrt{N}}$ to the averaged errors (with c constant w.r.t. N), as suggested by Theorem 2.

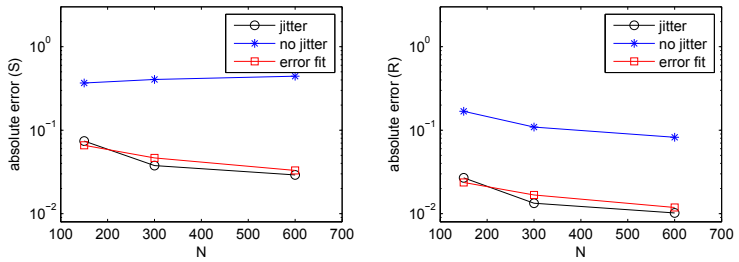
Figure 3 displays the empirical variance for the average errors of Figure 2, with and without jittering. It shows that the variability of the estimators is relatively large for small t and it reduces considerably as a longer observation record is accumulated.



(a) Parameter S. (b) Parameter R.

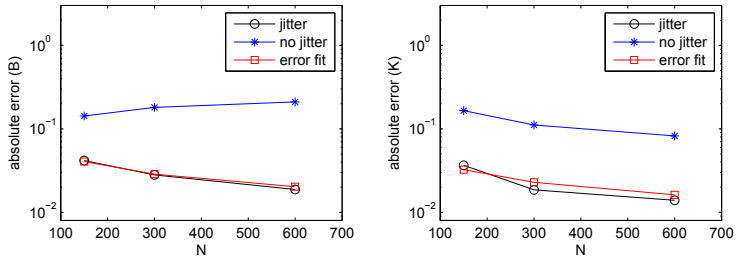
(c) Parameter B. (d) Parameter k_o.

Figure 1: Average of the absolute parameter estimation errors over 50 independent simulation runs using Algorithm 3 with $N = M = 300$ particles ($K = N^2 = 90 \times 10^3$ particles overall). The absolute errors are normalised w.r.t. the true parameter values, $S = 10, R = 28, B = \frac{8}{3}$ and $k_o = \frac{4}{5}$. The results obtained when jittering is suppressed in Algorithm 3 (labeled as *no jitter*) are shown for comparison. The horizontal axis is in discrete-time units. As the integration period is $\mathsf{T}_e = 10^{-3}$, 100,000 discrete-time steps amount 100 continuous time units. Observations are collected every 40 discrete-time steps.



(a) Parameter S. The least squares fit of the errors yields $c \approx 0.807$.

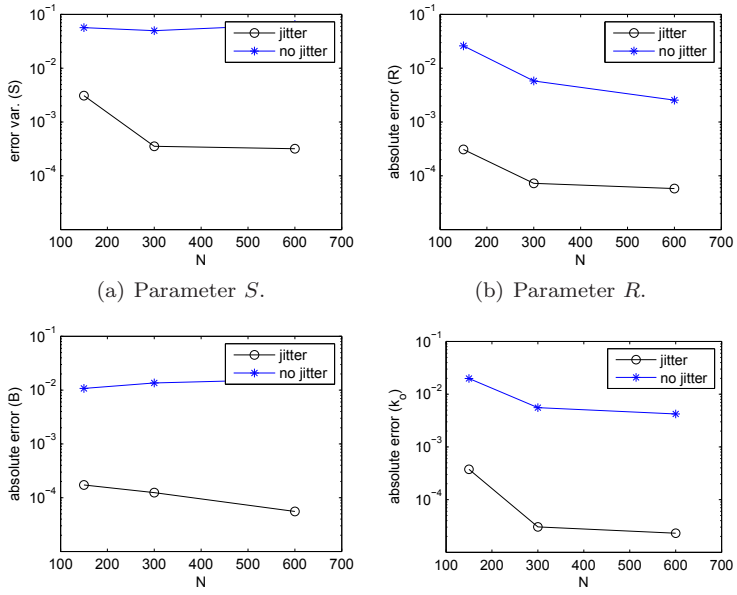
(b) Parameter R. The least squares fit of the errors yields $c \approx 0.290$.



(c) Parameter B. The least squares fit of the errors yields $c \approx 0.496$.

(d) Parameter k_o. The least squares fit of the errors yields $c \approx 0.397$.

Figure 2: Average of the absolute parameter estimation errors over 20 independent simulation runs using Algorithm 3 with $N = M = 150$, $N = M = 300$ and $N = M = 600$ (the total number of particles is N^2). The errors are normalised w.r.t. the true parameter values, $S = 10, R = 28, B = \frac{8}{3}$ and $k_o = \frac{4}{5}$. The curves labeled *error fit* have the form $\frac{c}{\sqrt{N}}$, where the constant c is a least squares estimate computed independently for each parameter. The results obtained when jittering is suppressed in Algorithm 3 (labeled as *no jitter*) are also shown for comparison. In each simulation, the Lorenz system was run for 24,000 discrete-time steps (24 continuous-time steps, for $\mathsf{T}_e = 10^{-3}$), with observations collected every 40 discrete steps.



(a) Parameter S. (b) Parameter R.

(c) Parameter B. (d) Parameter k_o.

Figure 3: Empirical variance of the absolute parameter estimation errors over 20 independent simulation runs using Algorithm 3 with $N = M = 150$, $N = M = 300$ and $N = M = 600$ (the total number of particles is N^2). The errors are normalised w.r.t. the true parameter values, $S = 10, R = 28, B = \frac{8}{3}$ and $k_o = \frac{4}{5}$. The results obtained when jittering is suppressed in Algorithm 3 (labeled as *no jitter*) are also shown for comparison. In each simulation, the Lorenz system was run for 24,000 discrete-time steps (24 continuous-time steps, for $\mathsf{T}_e = 10^{-3}$), with observations collected every 40 discrete steps.

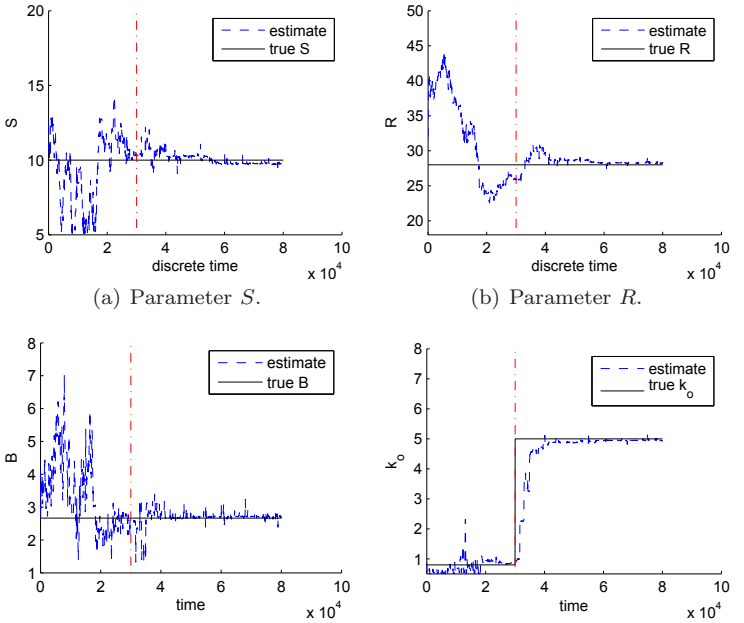
Finally, we have carried out a simple computer experiment to test the effect of a change-point in one of the parameters (the observation scale factor k_o). The simulation setup is the same as in the rest of this Section except that we extend the support of the parameter k_o to be the interval $\left[\frac{1}{2}, 8\right]$, with uniform a priori probability distribution, and artificially introduce a change-point at continuous time instant 30, where k_o changes its value from 0.8 to 5. This change-point is not described by the model, that represents k_o as strictly constant. We have run Algorithm 3 once, with $N = M = 500$ particles, and observed the evolution over time of the posterior-mean estimators for S, B, R and k_o.

Figure 4 shows that the posterior-mean estimates fluctuate considerably for (relatively) small t, as we concluded from observing their empirical variance. The value of k_o is changed at discrete time 3×10^4, which corresponds to continuous time 30 and a sequence of 750 observations. The change is instantaneous, yielding a step function for k_o as plotted in Figure 4(d). Before the change-point, the random support of the posterior distribution of k_o concentrates around the original value $k_o = 0.8$. After the change-point, this support has to be adapted. However, the pace of this adaptation is limited by the variance of the jittering kernel and, hence, we observe a transition in the sequence of estimates that lasts for nearly 10^4 time steps (10 continuous time units, 250 observations). Eventually, the posterior mean settles around the new value of k_o in this simulation; however, further investigation is needed regarding the speed at which the random support of $\mu_t^{N,N}$ can be adapted and its interplay with estimation errors.

7. Conclusions

We have introduced a recursive Monte Carlo scheme, consisting of two (nested) layers of particle filters, for the approximation and tracking of the posterior probability distribution of the unknown parameters of a state-space Markov system. Unlike existing SMC^2 and particle MCMC methods, the proposed algorithm is purely recursive and can be seen as a natural adaptation of the classic bootstrap filter to operate on the space of the static system-parameters.

The main theoretical contribution of the paper is the analysis of the errors in the approximation of integrals of bounded functions w.r.t. the posterior probability measure of the parameters. Using induction arguments, and placing only mild constraints on the state-space model and the parameters, we have proved that the L_p norms of the approximation errors for the proposed algorithm vanish with rate proportional to $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}$, where N is the number of particles in the parameter space and $N \times M$ is the number of particles in the state space. This is achieved with a computational cost that grows only linearly with time. In comparison, the computational load of the SMC^2 method increases quadratically with time. The price to pay for this reduction in computational cost is that in the new scheme we need $N \to \infty$ and $M \to \infty$ in order to make the error converge towards 0, while the SMC^2 algorithm is consistent for fixed M, i.e., $N \to \infty$ is sufficient for the errors to vanish, independently of M. As a consequence, if $K = NM$ is the total number of particles in the state space, then the optimal allocation



(a) Parameter S.

(b) Parameter R.

(c) Parameter B.

(d) Parameter k_o, with a change-point at time 3×10^4.

Figure 4: Evolution over time of the posterior-mean estimates of the parameters S, B, R and k_o for a single run of Algorithm 3 with $N = M = 500$. The actual parameter values of S, R, and B are indicated with a horizontal solid line. The value of k_o is also indicated, however it has a change-point at discrete time 3×10^4 (from 0.8 to 5). The change-point itself is marked by a vertical dashed line in the four plots. The algorithm is capable of tracking the change in k_o, however the adaptation of the estimator is limited by the variance of the jittering kernel and we observe a relatively long transition period of $\approx 10^4$ discrete time steps until the posterior mean settles around the new value.

for the proposed nested particle filter is $N = M = \sqrt{K}$ and the errors converge as $K^{-\frac{1}{4}}$ in L_p, while the SMC^2 scheme, with M fixed, converges as $K^{-\frac{1}{2}}$.

The proposed algorithm can be combined with a SMC^2 scheme for practical convenience. For example, one may run a standard SMC^2 algorithm on the initial part of the observation sequence (possibly a few tens or a few hundreds of observations, depending on the problem and the available computational resources) to take advantage of its faster convergence rate and then switch to a recursive nested particle filter (Algorithm 3) when the computational cost of batch processing becomes too high.

We also note that the continuity argument that leads to the derivation the the recursive nested particle filter, and the theoretical framework for the analysis of the resulting approximations, can be extended to other similar filtering algorithms. For example, it would be relatively straightforward to obtain a recursive version of the original IBIS algorithm of (Chopin, 2002).

Acknowledgements

The work of the D. Crisan has been partially supported by the EPSRC grant no EP/N023781/1. The work of J. Míguez was partially supported by the Office of Naval Research Global (award no. N62909- 15-1-2011), *Ministerio de Economía y Competitividad* of Spain (project TEC2015-69868-C2-1-R ADVENTURE) and *Ministerio de Educación, Cultura y Deporte* of Spain (*Programa Nacional de Movilidad de Recursos Humanos* PRX12/00690).

Part of this work was carried out while J. M. was a visitor at the Department of Mathematics of Imperial College London, with partial support from an EPSRC Mathematics Platform grant. D. C. and J. M. would also like to acknowledge the support of the Isaac Newton Institute through the program "Monte Carlo Inference for High-Dimensional Statistical Models", as well as the constructive comments of an anonymous Reviewer, who helped improving the final version of this manuscript.

Appendix A: Proof of Lemma 1

We consider first the predictive measure

$$\xi_{t,\theta}^N(dx) = \frac{1}{N}\sum_{n=1}^{N} \delta_{\bar{x}_t^{(n)}}(dx)$$

where $\bar{x}_t^{(n)}$, $n = 1, ..., N$, are the state particles drawn from the transition kernels $\tau_{t,\theta}^{x_{t-1}^{(n)}}(dx) \triangleq \tau_{t,\theta}(dx|x_{t-1}^{(n)})$ at the sampling step of the particle filter. Recall that $\xi_{t,\theta} =$

$\tau_{t,\theta}\phi_{t-1,\theta}$ and consider the triangle inequality

$$\begin{aligned}\left\|(f,\xi_{t,\theta}^N)-(f,\xi_{t,\theta})\right\|_p &= \left\|(f,\xi_{t,\theta}^N)-(f,\tau_{t,\theta}\phi_{t-1,\theta})\right\|_p \\ &\leq \left\|(f,\xi_{t,\theta}^N)-(f,\tau_{t,\theta}\phi_{t-1,\theta}^N)\right\|_p \\ &\quad + \left\|(f,\tau_{t,\theta}\phi_{t-1,\theta}^N)-(f,\tau_{t,\theta}\phi_{t-1,\theta})\right\|_p, \end{aligned} \tag{A.1}$$

where

$$(f,\tau_{t,\theta}\phi_{t-1,\theta}^N)=\frac{1}{N}\sum_{n=1}^N\int f(x)\tau_{t,\theta}(dx|x_{t-1}^{(n)})=\frac{1}{N}\sum_{n=1}^N(f,\tau_{t,\theta}^{x_{t-1}^{(n)}}). \tag{A.2}$$

In the sequel we seek upper bounds for the L_p norms in the right hand side of (A.1).

Let us introduce the σ-algebra generated by the random paths $x_{0:t}^{(n)}$ and $\bar{x}_{1:t}^{(n)}$, $n=1,...,N$, denoted $\mathcal{F}_t=\sigma\left(x_{0:t}^{(n)},\bar{x}_{1:t}^{(n)},\quad n=1,...,N\right)$. The conditional expectation of the integral $(f,\xi_{t,\theta}^N)$ given $\mathcal{F}_{t-1}$ is

$$\begin{aligned}E\left[(f,\xi_{t,\theta}^N)|\mathcal{F}_{t-1}\right] &= \frac{1}{N}\sum_{n=1}^N E\left[f(\bar{x}_t^{(n)})|\mathcal{F}_{t-1}\right] \\ &= \frac{1}{N}\sum_{n=1}^N(f,\tau_{t,\theta}^{x_{t-1}^{(n)}})=(f,\tau_{t,\theta}\phi_{t-1,\theta}^N)\end{aligned}$$

and we note that the random variables $S_{t,\theta}^{(n)}=f(\bar{x}_t^{(n)})-(f,\tau_{t,\theta}^{x_{t-1}^{(n)}})$, $n=1,...,N$, are independent and zero-mean conditional on the σ-algebra $\mathcal{F}_{t-1}$. For even p, the approximation error between $\xi_{t,\theta}^N$ and its (conditional) expected value $\tau_{t,\theta}\phi_{t-1,\theta}^N$ can then be written as

$$\begin{aligned}E\left[\left((f,\xi_{t,\theta}^N)-(f,\tau_t\phi_{t-1,\theta}^N)\right)^p|\mathcal{F}_{t-1}\right] &= E\left[\left(\frac{1}{N}\sum_{n=1}^N S_{t,\theta}^{(n)}\right)^p|\mathcal{F}_{t-1}\right] \\ &= \frac{1}{N^p}\sum_{n_1=1}^N\cdots\sum_{n_p=1}^N E\left[S_{t,\theta}^{(n_1)}\ldots S_{t,\theta}^{(n_p)}|\mathcal{F}_{t-1}\right].\end{aligned} \tag{A.3}$$

Since the random variables $S_{t,\theta}^{(n_i)}$ are conditionally independent and zero-mean, every term in the summation of (A.3) involving a moment of order 1 vanishes. It is an exercise in combinatorics to show that the number of terms which *do not* contain any moment of order 1 is a polynomial function of N with degree $\frac{p}{2}$, whose coefficients depend only on p. As a consequence, there exists a constant $\tilde{c}$ independent of N such that the number of non-zero terms in (A.3) is at most $\tilde{c}^p N^{\frac{p}{2}}$. Moreover, for each non-zero term we readily calculate the upper bound $E\left[S_{t,\theta}^{(n_1)}\ldots S_{t,\theta}^{(n_p)}|\mathcal{F}_{t-1}\right]\leq 2^p\|f\|_\infty^p$. Therefore, for even p, we

arrive at the inequality

$$E\left[\left((f,\xi^N_{t,\theta})-(f,\tau_t\phi^N_{t-1,\theta})\right)^p|\mathcal{F}_{t-1}\right]\leq\frac{\tilde{c}^p2^p\|f\|^p_\infty}{N^{\frac{p}{2}}}\tag{A.4}$$

and taking unconditional expectations on both sides of (A.4), we readily find that,

$$\|(f,\xi^N_{t,\theta})-(f,\tau_t\phi^N_{t-1,\theta}\|_p\leq\frac{c_1\|f\|_\infty}{\sqrt{N}},\tag{A.5}$$

where $c_1=2\tilde{c}$ is a constant independent of N and θ. The same inequality (A.5) holds for any real p because of the monotonicity of L_p norms (an application of Jensen's inequality).

For the second term in the right hand side of (A.1), we note that $(f,\tau_{t,\theta}\phi_{t-1,\theta})=(\bar{f}_\theta,\phi_{t-1,\theta})$, where $\bar{f}_\theta\in B(\mathbb{R}^{d_x})$ is a bounded[5] function defined as

$$\bar{f}_\theta(x)=\int f(x')\tau^x_{t,\theta}(dx')=(f,\tau^x_{t,\theta})$$

and, similarly, $(f,\tau_{t,\theta}\phi^N_{t-1,\theta})=(\bar{f}_\theta,\phi^N_{t-1,\theta})$. Therefore, assumption (2.15) yields the upper bound

$$\begin{aligned}\left\|(f,\tau_{t,\theta}\phi^N_{t-1,\theta})-(f,\tau_{t,\theta}\phi_{t-1,\theta})\right\|_p&=\left\|(\bar{f}_\theta,\phi^N_{t-1,\theta})-(\bar{f}_\theta,\phi_{t-1,\theta})\right\|_p\\&\leq\frac{c_{t-1}\|f\|_\infty}{\sqrt{N}}+\frac{\bar{c}_{t-1}\|f\|_\infty}{\sqrt{M}},\end{aligned}\tag{A.6}$$

where the constants $c_{t-1},\bar{c}_{t-1}$ are independent of N, M and θ. Substituting (A.5) and (A.6) into (A.1) yields

$$\left\|(f,\xi^N_{t,\theta})-(f,\xi_{t,\theta})\right\|_p\leq\frac{\tilde{c}_t\|f\|_\infty}{\sqrt{N}}+\frac{\bar{\tilde{c}}_t\|f\|_\infty}{\sqrt{M}},\tag{A.7}$$

where $\tilde{c}_t=c_{t-1}+c_1$ and $\bar{\tilde{c}}_t=\bar{c}_{t-1}$ are finite constants independent of N, M and θ.

Next, we use inequality (A.7) to calculate a bound on $\|(f,\phi^N_{t,\theta})-(f,\phi_{t,\theta})\|_p$. Let us first note that, after the computation of the weights, we obtain a random measure of the form

$$\bar{\phi}^N_{t,\theta}(dx)=\sum_{n=1}^N w_t^{(n)}\delta_{\bar{x}_t^{(n)}}(dx),\quad\text{where}\quad w_t^{(n)}=\frac{g^{y_t}_{t,\theta}(\bar{x}_t^{(n)})}{\sum_{k=1}^N g^{y_t}_{t,\theta}(\bar{x}_t^{(n)})}.$$

As a consequence, integrals w.r.t. the measure $\bar{\phi}^N_{t,\theta}$ can be written in terms of $g^{y_t}_{t,\theta}$ and $\xi^N_{t,\theta}$, namely

$$(f,\bar{\phi}^N_{t,\theta})=\frac{(fg^{y_t}_{t,\theta},\xi^N_{t,\theta})}{(g^{y_t}_{t,\theta},\xi^N_{t,\theta})}.\tag{A.8}$$

[5]Trivially note that $\|\bar{f}_\theta\|_\infty\leq\|f\|_\infty$, independently of θ.

This is natural, though, since from the Bayes theorem we readily derive the same relationship between $\phi_{t,\theta}$ and $\xi_{t,\theta}$,

$$(f, \phi_{t,\theta}) = \frac{(f g_{t,\theta}^{y_t}, \xi_{t,\theta})}{(g_{t,\theta}^{y_t}, \xi_{t,\theta})}. \tag{A.9}$$

Given (A.8) and (A.9), we can readily apply the inequality (2.1) to obtain

$$\begin{aligned} \left|(f, \bar{\phi}_{t,\theta}^N) - (f, \phi_{t,\theta})\right| \quad \leq \quad & \frac{1}{(g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta})} \left(\|f\|_\infty \left|(g_{t,\theta}^{y_t}, \xi_{t,\theta}) - (g_{t,\theta}, \xi_{t,\theta}^N)\right| \right. \\ & \left. + \left|(f g_{t,\theta}^{y_t}, \xi_{t,\theta}) - (f g_{t,\theta}, \xi_{t,\theta}^N)\right| \right), \end{aligned} \tag{A.10}$$

where $u_t(\theta) = (g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta}) > 0$ by assumption. From (A.10) and Minkowski's inequality,

$$\begin{aligned} \left\|(f, \bar{\phi}_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p \quad \leq \quad & \frac{1}{(g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta})} \times \left(\|f\|_\infty \left\|(g_{t,\theta}^{y_t}, \xi_{t,\theta}) - (g_{t,\theta}^{y_t}, \xi_{t,\theta}^N)\right\|_p \right. \\ & \left. + \left\|(f g_{t,\theta}^{y_t}, \xi_{t,\theta}) - (f g_{t,\theta}, \xi_{t,\theta}^N)\right\|_p \right) \end{aligned} \tag{A.11}$$

and, since $\|g_{t,\theta}^{y_t}\|_\infty \leq \|g_t^{y_t}\|_\infty < \infty$ by assumption (in particular, $\|g_t^{y_t}\|_\infty$ is independent of θ), the inequalities (A.7) and (A.11) together yield

$$\left\|(f, \bar{\phi}_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p \leq \frac{2\|f\|_\infty \|g_t^{y_t}\|_\infty \tilde{c}_t}{(g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta})} \times \frac{1}{\sqrt{N}} + \frac{2\|f\|_\infty \|g_t^{y_t}\|_\infty \bar{\tilde{c}}_t}{(g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta})} \times \frac{1}{\sqrt{M}}, \tag{A.12}$$

where the finite constants $\tilde{c}_t$ and $\bar{\tilde{c}}_t = \bar{c}_{t-1}$ are independent of N, M and θ. Indeed, the only factor that depends on θ in the right-hand side of (A.12) is the integral $u_t(\theta) = (g_{t,\theta}^{y_{t,\theta}}, \xi_{t,\theta})$. However, we have assumed that

$$u_{t,\inf} = \inf_{\theta \in D_\theta} u_t(\theta) > 0, \tag{A.13}$$

hence the inequality (A.12) leads to

$$\left\|(f, \bar{\phi}_{t,\theta}^N) - (f, \phi_{t,\theta})\right\|_p \leq \frac{c_{2,t}\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_{2,t}\|f\|_\infty}{\sqrt{M}} \tag{A.14}$$

where

$$c_{2,t} = \frac{2\|g_t^{y_t}\|_\infty \tilde{c}_t}{u_{t,\inf}} < \infty \quad \text{and} \quad \bar{c}_{2,t} = \frac{2\|g_t^{y_t}\|_\infty \bar{c}_{t-1}}{u_{t,\inf}} < \infty \tag{A.15}$$

are constants independent of N, M and θ.

Finally, we only need to verify the resampling step, i.e., that the L_p norm $\|(f, \phi_{t,\theta}^N) - (f, \bar{\phi}_{t,\theta}^N)\|_p$ is bounded as well. Let $\bar{\mathcal{F}}_t = \sigma\left(x_{0:t-1}^{(n)}, \bar{x}_{1:t}^{(n)}; n = 1, \ldots, N\right)$ be the σ-algebra

generated by the random sequences $x_{0:t-1}^{(n)}$ and $\bar{x}_{1:t}^{(n)}$, $n = 1, ..., N$. It is straightforward to check that, for every $n = 1, ..., N$,

$$E\left[f(x_t^{(n)})|\bar{\mathcal{F}}_t\right] = (f, \bar{\phi}_{t,\theta}^N), \tag{A.16}$$

hence the random variables $\bar{S}_{t,\theta}^{(n)} = f(x_t^{(n)}) - (f, \bar{\phi}_{t,\theta}^N)$ are independent and zero-mean conditional on the σ-algebra $\bar{\mathcal{F}}_t$. Therefore, the same combinatorial argument that led to Eq. (A.5) now yields

$$\left\|(f, \phi_{t,\theta}^N) - (f, \bar{\phi}_{t,\theta}^N)\right\|_p \leq \frac{c_3\|f\|_\infty}{\sqrt{N}} \tag{A.17}$$

where the constant c_3 is independent of both N and θ (it does not depend on the distribution of the error variables $\bar{S}_{t,\theta}^{(n)}$). Since

$$\|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\|_p \leq \|(f, \phi_{t,\theta}^N) - (f, \bar{\phi}_{t,\theta}^N)\|_p + \|(f, \bar{\phi}_{t,\theta}^N) - (f, \phi_{t,\theta})\|_p, \tag{A.18}$$

substituting Eqs. (A.17) and (A.14) into the inequality (A.18) yields

$$\|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\|_p \leq \frac{c_t\|f\|_\infty}{\sqrt{N}} + \frac{\bar{c}_t\|f\|_\infty}{\sqrt{M}}, \tag{A.19}$$

where $c_t = c_3 + c_{2,t}$ and $\bar{c}_t = \bar{c}_{2,t}$ are finite constants independent of both N, M and θ.

To complete the proof, simply note that $\bar{c}_{t-1} = 0$ implies $\bar{c}_t = \bar{c}_{2,t} = 0$ (see (A.15)). □

Appendix B: Proof of Lemma 2

We proceed by induction in t. For $t = 0$, the measure $\phi_{0,\theta}^N(dx) = \frac{1}{N}\sum_{n=1}^N \delta_{x_0^{(n)}}(dx)$ is constructed from an i.i.d. sample of size N from the prior distribution $\phi_{0,\theta} \equiv \tau_0$. Then, it is straightforward to prove that

$$\|(f, \phi_{0,\theta}^N) - (f, \phi_{0,\theta})\|_p \leq \frac{c_0\|f\|_\infty}{\sqrt{N}},$$

where $c_0 < \infty$ is independent of N. Note that, since $\phi_{0,\theta} \equiv \tau_0$ is actually independent of θ, the constant c_0 is independent of θ as well.

For the induction step, we assume that

$$\|(f, \phi_{t-1,\theta}^N) - (f, \phi_{t-1,\theta})\|_p \leq \frac{c_{t-1}\|f\|_\infty}{\sqrt{N}} \tag{B.1}$$

holds true for some constant $c_{t-1} < \infty$ independent of N and θ. Given (B.1), Lemma 1 yields

$$\|(f, \xi_{t,\theta}^N) - (f, \xi_{t,\theta})\|_p \leq \frac{\tilde{c}_t\|f\|_\infty}{\sqrt{N}} \quad \text{and} \quad \|(f, \phi_{t,\theta}^N) - (f, \phi_{t,\theta})\|_p \leq \frac{c_t\|f\|_\infty}{\sqrt{N}}$$

at time t, where $\tilde{c}_t$ and c_t are finite constants independent of N and θ. □

Appendix C: A family of jittering kernels

Proposition 1. Assume that $h \in B(D_\theta)$ is Lipschitz, with constant $c_L\|h\|_\infty < \infty$, and consider the class of kernels $\kappa_N^{\theta'} = (1-\epsilon_N)\delta_{\theta'} + \epsilon_N \bar{\kappa}_N^{\theta'}$, where $0 \leq \epsilon_N \leq 1$ and $\bar{\kappa}_N^{\theta'} \in \mathcal{P}(D_\theta)$. For any $p \geq 1$, if the kernel $\kappa_N^{\theta'}$ is selected in such a way that

$$\sigma^2_{\kappa,N} = \sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^2 \bar{\kappa}_N^{\theta'}(d\theta) \leq \frac{\breve{c}}{\epsilon_N^{\frac{p+2}{p}} N^{\frac{p+2}{2}}} \tag{C.1}$$

is satisfied for some constant $\breve{c} < \infty$ independent of N, then the inequality

$$\sup_{\theta' \in D_\theta} \int |h(\theta) - h(\theta')|^p \kappa_N^{\theta'}(d\theta) \leq \frac{c_\kappa^p \|h\|_\infty^p}{N^{\frac{p}{2}}}$$

holds for a constant $c_\kappa^p = c_L^p \left(1 + \breve{c} \sup_{\theta_1,\theta_2 \in D_\theta} \|\theta_1 - \theta_2\|^p\right) < \infty$ independent of N.

Proof. Since $\kappa_N^{\theta'} = (1-\epsilon_N)\delta_{\theta'} + \epsilon_N \bar{\kappa}_N^{\theta'}$ and h is Lipschitz with constant $c_L\|h\|_\infty < \infty$, we readily obtain

$$\int |h(\theta) - h(\theta')|^p \kappa_N^{\theta'}(d\theta) \leq \epsilon_N c_L^p \|h\|_\infty^p \int \|\theta - \theta'\|^p \bar{\kappa}_N^{\theta'}(d\theta). \tag{C.2}$$

Let

$$\beta_N = \frac{1}{\epsilon_N^{\frac{1}{p}} \sqrt{N}}. \tag{C.3}$$

We can rewrite (C.2) as

$$\begin{aligned} \int |h(\theta) - h(\theta')|^p \kappa_N^{\theta'}(d\theta) &\leq \epsilon_N c_L^p \|h\|_\infty^p \left[\int I_{\theta \in D_\theta : \|\theta - \theta'\| < \beta_N}(\theta) \|\theta - \theta'\|^p \bar{\kappa}_N^{\theta'}(d\theta) \right. \\ &\quad \left. + \int I_{\theta \in D_\theta : \|\theta - \theta'\| \geq \beta_N}(\theta) \|\theta - \theta'\|^p \bar{\kappa}_N^{\theta'}(d\theta) \right] \\ &\leq \epsilon_N c_L^p \|h\|_\infty^p \left[\beta_N^p + \hat{C}^p \int I_{\theta \in D_\theta : \|\theta - \theta'\| \geq \beta_N}(\theta) \bar{\kappa}_N^{\theta'}(d\theta) \right], \end{aligned} \tag{C.4}$$

where $\hat{C}^p = \sup_{\theta_1,\theta_2 \in D_\theta} \|\theta_1 - \theta_2\|^p < \infty$, since D_θ is compact. Using Chebyshev's inequality on the right hand side of (C.4) yields

$$\int |h(\theta) - h(\theta')|^p \kappa_N^{\theta'}(d\theta) \leq \epsilon_N c_L^p \|h\|_\infty^p \left(\beta_N^p + \hat{C}^p \frac{\sigma^2_{\kappa,N}}{\beta_N^2} \right) \tag{C.5}$$

and substituting (C.1) and (C.3) into (C.5) we arrive at

$$\int |h(\theta) - h(\theta')|^p \kappa_N^{\theta'}(d\theta) \leq \frac{c_L^p \|h\|_\infty^p \left(1 + \breve{c}\hat{C}^p\right)}{N^{\frac{p}{2}}},$$

where all the constants are independent of θ' and N. □

Corollary 1. Consider the same class of kernels $\kappa_N^{\theta'} = (1-\epsilon_N)\delta_{\theta'} + \epsilon_N \bar{\kappa}_N^{\theta'}$, where $0 \leq \epsilon_N \leq 1$ and $\bar{\kappa}_N^{\theta'} \in \mathcal{P}(D_\theta)$. For any $p \geq 1$, if (C.1) holds for some $\check{c} < \infty$ independent of N then

$$\sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^p \kappa_N^{\theta'}(d\theta) \leq \frac{c_\kappa^p}{N^{\frac{p}{2}}}$$

where $c_\kappa^p = 1 + \check{c} \sup_{\theta_1, \theta_2 \in D_\theta} \|\theta_1 - \theta_2\|^p < \infty$ is constant and independent of N.

Proof. Simply note that

$$\int \|\theta - \theta'\|^p \kappa_N^{\theta'}(d\theta) \leq \epsilon_N \int \|\theta - \theta'\|^p \bar{\kappa}_N^{\theta'}(d\theta)$$

and then follow the same argument as in the proof of Proposition 1. □

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B* **72** 269–342.

Andrieu, C., Doucet, A., Singh, S. S. and Tadić, V. B. (2004). Particle Methods for Change Detection, System Identification and Control. *Proceedings of the IEEE* **92** 423-438.

Beskos, A., Crisan, D., Jasra, A. et al. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *The Annals of Applied Probability* **24** 1396–1445.

Bruno, M. G. S. (2013). Sequential Monte Carlo Methods for Nonlinear Discrete-Time Filtering. *Synthesis Lectures on Signal Processing* **6** 1–99.

Cappé, O., Godsill, S. J. and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE* **95** 899–924.

Cappé, O., Gullin, A., Marin, J. M. and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics* **13** 907-929.

Carvalho, C. M., Johannes, M. S., Lopes, H. F. and Polson, N. G. (2010). Particle learning and smoothing. *Statistical Science* **25** 88–106.

Chen, R., Wang, X. and Liu, J. S. (2000). Adaptive Joint Detection and Decoding in Flat-Fading Channels via Mixture Kalman Filtering. *IEEE Transactions Information Theory* **46** 2079-2094.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–552.

Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 397–426.

Chorin, A. J. and Krause, P. (2004). Dimensional Reduction for a Bayesian Filter. *PNAS* **101** 15013-15017.

Crisan, D. (2001). Particle Filters - A Theoretical Perspective. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 2, 17-42. Springer.

Crisan, D. and Doucet, A. (2002). A Survey of Convergence Results on Particle Filtering. *IEEE Transactions Signal Processing* **50** 736-746.

Douc, R., Cappé, O. and Moulines, E. (2005). Comparison of Resampling Schemes for Particle Filtering. In *Proceedings of the 4^{th} International Symposium on Image and Signal Processing and Analysis* 64-69.

Doucet, A., de Freitas, N. and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 1, 4-14. Springer.

Doucet, A., Godsill, S. and Andrieu, C. (2000). On sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing* **10** 197-208.

Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics* **11** 848–862.

Gilks, W. R. and Berzuini, C. (2001). Following a moving targetÑMonte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 127–146.

Gordon, N., Salmond, D. and Smith, A. F. M. (1993). Novel Approach to Nonlinear and Non-Gaussian Bayesian State Estimation. *IEE Proceedings-F* **140** 107-113.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J. M. and Chopin, N. (2015). On Particle Methods for Parameter Estimation in State-Space Models. *Statistical Science* **30** 328-351.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models. *J. Comput. Graph. Statist.* **1** 1-25.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association* 1203–1215.

Koblents, E. and Míguez, J. (2013). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*.

Koblents, E. and Míguez, J. (2015). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing* **25** 407–425.

Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **9** 278-288.

LeGland, F. and Mevel, L. (1997). Recursive estimation in hidden Markov models. In *Proceedings of the 36th IEEE Conference on Decision and Control, 1997* **4** 3468–3473. IEEE.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association* **93** 1032-1044.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.) 10, 197-223. Springer.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences* **20** 130-141.

Maíz, C. S., Molanes-López, E., Míguez, J. and Djurić, P. M. (2012). A Particle Filtering Scheme for Processing Time Series Corrupted by Outliers. *IEEE Transactions*

on Signal Processing **9**.

Miguez, J., Bugallo, M. and Djuric, P. M. (2005). Novel particle filtering algorithms for fixed parameter estimation in dynamic systems. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)* 46–51. IEEE.

Míguez, J., Crisan, D. and Djurić, P. M. (2013). On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing* **23** 91–107.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer.

Moral, P. D., Doucet, A. and Singh, S. (2015). Uniform stability of a particle approximation of the optimal filter derivative. *SIAM Journal on Control and Optimization* **53** 1278–1304.

Del Moral, P. and Miclo, L. (2000). Branching and interacting particle systems. Approximations of Feynman-Kac formulae with applications to non-linear filtering. *Lecture Notes in Mathematics* 1-145.

Olsson, J., Cappé, O., Douc, R. and Moulines, E. (2008). Sequential Monte Carlo smoothing with Application to Parameter Estimation in Nonlinear State Space Models. *Bernoulli* **14** 155-179.

Papavasiliou, A. (2006). Parameter Estimation and Asymptotic Stability in Stochastic Filtering. *Stochastic Processes and Their Applications* **116** 1048-1065.

Poyiadjis, G., Doucet, A. and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika* **98** 65–80.

Ristic, B., Arulampalam, S. and Gordon, N. (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Boston.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions Signal Processing* **50** 281-289.