Submitted to the Bernoulli
arXiv: 1111.5866

Particle-kernel estimation of the filter density in state-space models

DAN CRISAN[1] and JOAQUíN MíGUEZ[2]

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

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

Sequential Monte Carlo (SMC) methods, also known as particle filters, are simulation-based recursive algorithms for the approximation of the *a posteriori* probability measures generated by state-space dynamical models. At any given time t, a SMC method produces a set of samples over the state space of the system of interest (often termed "particles") that is used to build a discrete and random approximation of the posterior probability distribution of the state variables, conditional on a sequence of available observations. One potential application of the methodology is the estimation of the densities associated to the sequence of *a posteriori* distributions. While practitioners have rather freely applied such density approximations in the past, the issue has received less attention from a theoretical perspective. In this paper, we address the problem of constructing kernel-based estimates of the posterior probability density function and its derivatives, and obtain asymptotic convergence results for the estimation errors. In particular, we find convergence rates for the approximation errors that hold uniformly on the state space and guarantee that the error vanishes almost surely as the number of particles in the filter grows. Based on this uniform convergence result, we first show how to build continuous measures that converge almost surely (with known rate) toward the posterior measure and then address a few applications. The latter include maximum *a posteriori* estimation of the system state using the approximate derivatives of the posterior density and the approximation of functionals of it, e.g., Shannon's entropy.

Keywords: sequential Monte Carlo, particle filtering, density estimation, stochastic filtering, state-space models, Markov systems.

1. Introduction

1.1. Background

Consider two random sequences, $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 1}$, possibly multidimensional, where X_t represents the unobserved state of a system of interest and Y_t is a related observation. Very often, the dependence between the two sequences is given by a Markov state-space model and the posterior probability measure that characterizes the random variable X_t conditional on the observations $\{Y_s, 1 \leq s \leq t\}$ is usually termed the "filtering measure",

denoted as π_t in the sequel. If the model is linear and Gaussian, π_t is also Gaussian and can be computed exactly using a set of recursive equations known as the Kalman filter [33]. However, if X_t takes values in a continuous space and the model is nonlinear or non-Gaussian, the exact filter is intractable and numerical approximation techniques are necessary. The class of sequential Monte Carlo (SMC) methods, also known as particle filters [27, 34, 37, 22, 21], has become a very popular tool for this purpose. Particle filters generate discrete random measures (constructed from random samples in the state space) that can be naturally used to approximate integrals with respect to (w.r.t.) the filtering measure.

The asymptotic convergence of SMC algorithms has been well studied during the past two decades. The first formal results appeared in [13, 14], while the analysis in [9] already took into account the branching (resampling) step indispensable in most practical applications. Currently, there is a broad knowledge about the convergence of particle filters in some of the forms commonly used in practical applications; see [17, 8, 15, 36, 3, 30] and the references therein. Most of these results are aimed to show that integrals of real functions w.r.t. π_t can be accurately approximated by weighted sums when the particle filter is run with a sufficiently large number of random samples (commonly referred to as particles). More recently, other types of convergence have been investigated. For instance, the convergence of particle approximations of maximum a posteriori (MAP) estimates of sequences has also been proved. Convergence in probability can be shown using random genealogical trees (see [41] and [15]) while almost sure convergence can also be guaranteed by extending the analysis in [10] (see [39]).

In most cases of interest, the filtering measure has a density, denoted p_t, w.r.t. a dominating measure (usually Lebesgue's) and practitioners have freely used various estimators of this function. Less attention has been devoted to this problem from a theoretical perspective, though. Note that the samples generated by the particle filter are not drawn directly from p_t: they can only be considered as *approximate samples*, in the sense that they can be used to estimate the value of integrals w.r.t. the measure π_t. As a consequence, the convergence of a kernel density estimate of p_t built from the output of a particle filter cannot be justified directly using the classical theory of kernel density estimation, which is concerned with samples drawn directly from the distribution of interest (see, e.g., [18, 43, 45, 44]).

The estimation of p_t is of interest by itself, since it naturally enables the computation of confidence regions, as well as MAP and maximum likelihood estimators, but also because it leads to the approximation of π_t by a continuous (instead of discrete) random measure. The convergence of continuous approximations of the filtering measure in total variation distance has been investigated in the context of regularized particle filters [36] as well as for accept/reject and auxiliary particle filters [35].

1.2. Contributions

In this paper, we analyze the approximation of p_t constructed as the sum of properly scaled kernel functions located at the particle positions. Kernel methods [43, 45] are the

most widely used techniques for the non-parametric estimation of probability density functions (pdf's) and, therefore, it seems natural to analyze their convergence when applied to the approximate samples generated by particle filters.

The pdf estimators we analyze are based on generic kernel functions which are only required to satisfy mild standard conditions (essentially the same as in classical density estimation theory [43]). We describe how to build approximations in arbitrary-dimensional spaces $\mathbb{R}^d$, $d \geq 1$, and then analyze their convergence as the number of particles is increased and the bandwidth of the kernels is decreased. In particular, we obtain point-wise convergence rates for the absolute approximation errors, both of p_t and its derivatives[1] (provided they exist). The latter results can be extended to deduce uniform (instead of point-wise) convergence rates, again both for p_t and its derivatives. Specifically, we provide explicit bounds for the supremum of the approximation error and prove that it converges almost surely (a.s.) toward 0 as the number of particles is increased. Our analysis is different from the standard methods in kernel density estimation. The latter address the bias and variance of the estimators using approximations based on Taylor series (see, e.g., [43, Chapter 4] or [45, Chapter 4]) or Edgeworth expansions [28], which enable the asymptotic approximation of the mean integrated square error (MISE) of the density estimate and yield expressions involving the number of samples and the kernel bandwidth. We directly obtain convergence rates for various estimation errors (not only the MISE), given in terms of a single index that links the number of samples and the kernel bandwidth. This link is briefly discussed in Section 3.3.

The uniform (on the support of p_t) convergence result can be exploited in a number of ways. For instance, if we let p_t^N be the approximation of p_t with N particles, then we can obtain a continuous approximation of the filtering measure $\pi_t(dx)$ as $\breve{\pi}_t^N(dx) = p_t^N(x)dx$, prove that $\breve{\pi}_t^N$ converges to π_t a.s. in total variation distance (as $N \to \infty$) and provide explicit convergence rates. A similar kind of analysis also leads to the calculation of convergence rates for the MISE of the particle-kernel density estimator p_t^N. Additionally, we prove that the (random) integrated square error (ISE) of a truncated version of p_t^N converges to 0 a.s. and provide convergence rates. A comparison of these results with the standard asymptotic approximation of the MISE for kernel estimators built from i.i.d. samples is presented at the end of Section 4.3.

The convergence in total variation distance of a continuous approximation of the filtering measure π_t was also addressed in [36] and [35]. Compared to these earlier contributions, our analysis guarantees the almost sure convergence of the (random) total variation distance toward 0, with explicit rates, rather than the convergence of its expected value (as in [36]) or its convergence in probability (as in [35]). Also, our assumptions on the Markov kernel of the state process $\{X_t\}_{t\geq 0}$ and the conditional

[1] Let us note here that the approximation of derivatives of the filter has received attention recently, related to problems of parameter estimation in state-space systems [12, 16]. In the latter context, the filtering pdf is made to depend explicitly on a parameter vector $\theta = (\theta_1, \ldots, \theta_d)$, and the interest is in the computation of the partial derivatives $\partial p_t/\partial \theta_i$ in order to implement, e.g., maximum likelihood estimation algorithms [16]. In this paper, however, we consider derivatives with respect to the state variables in $X_t = (X_{1,t}, \ldots, X_{d_x,t})$, i.e., $\partial p_t/\partial x_{i,t}$.

densities of $\{Y_t|X_t\}_{t\geq 1}$ are relatively mild and simple to check. In particular, our results also hold for light-tailed Markov kernels (e.g., Gaussian), unlike Theorems 2 and 3 in [35].

The last part of the paper is devoted to some applications of the density approximation method and the uniform convergence result. We first consider the problem of MAP estimation. We refer here to the maximization of the filtering density, a problem different from that of MAP estimation in the path space addressed, e.g., in [26, 41, 39]. We first prove that the maxima of the approximation of the filtering density actually converge, asymptotically, to the maxima of the true function p_t and then show some simulation results that illustrate the use of gradient algorithms on the estimated density function.

The second application we describe is the approximation of functionals of p_t. We provide first a generic result that guarantees the almost sure convergence of such approximations for bounded and Lipschitz continuous functionals. Then, we address the problem of approximating Shannon entropies [7], which is of practical interest in various machine learning and signal processing problems. The log function is neither bounded nor Lipschitz continuous and, therefore, the latter generic result does not apply to the computation of entropies. We specifically address this problem resoriting to a new result on the convergence of the particle approximations of integrals of the form $\int f(x)\pi_t(dx)$ when the test function f is possibly unbounded. Let us remark that a large majority of the results in the literature [17, 8, 15, 36, 3] refer exclusively to the approximation of integrals of bounded functions. Only recently, the convergence of approximate integrals of unbounded test functions has been proved [31], albeit for a modified particle filter and assuming that the product of the test and the likelihood functions *is* bounded. Here, we prove the almost sure convergence of the approximations of integrals of unbounded functions for the standard particle filter, placing only integrability assumptions on the test function. From this result, we deduce the almost sure convergence toward 0 of the errors in the approximation of Shannon entropies for densities with a compact support. A numerical illustration is given.

1.3. Organization of the paper

The rest of the paper is organized as follows. Section 2 contains background material, including a summary of notation, a description of Markov state space models and the standard particle (*bootstrap*) filter. A new lemma that establishes the convergence of the particle approximation of posterior expectations of unbounded test functions is also introduced in Section 2. The construction of particle-kernel approximations of the filtering density and its derivatives is described in Section 3, where we also review some basics of kernel density estimation and the most relevant results in [36] and [35] for density estimation with particle filters. Our formal results on the convergence of the particle-kernel density estimators and the smooth approximation of the filtering measure are introduced in Section 4. This includes the point-wise and uniform approximations of $p_t(x)$, the convergence in total variation distance of the smooth measures $\breve{\pi}_t^N$ and convergence rates for the mean integrated square error and the (random) integrated

square error of p_t^N and its truncated version, respectively. In Section 5 we discuss applications of the particle-kernel estimator of p_t and its derivatives. In particular, we consider the problem of (marginal) MAP estimation of the state variables and the approximation of functionals of the filtering density, including Shannon's entropy. Finally, brief conclusions are presented in Section 6.

2. Particle filtering

2.1. Notation

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

- Measures and integrals.
  - $\mathcal{B}(\mathbb{R}^d)$ is the σ-algebra of Borel subsets of $\mathbb{R}^d$.
  - $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures over $\mathcal{B}(\mathbb{R}^d)$.
  - $(f, \mu) \triangleq \int f(x)\mu(dx)$ is the integral of a real function $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t. a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$.
  - Take a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, a Borel set $A \in \mathcal{B}(\mathbb{R}^d)$ and a real function $f : \mathbb{R}^d \to \mathbb{R}^+$. The projective product $f \star \mu$ is a measure, absolutely continuous w.r.t. μ and proportional to f, constructed as

$$(f \star \mu)(A) = \frac{\int_A f(x)\mu(dx)}{(f, \mu)}. \tag{2.1}$$

- Functions.
  - The supremum norm of a real function $f : \mathbb{R}^d \to \mathbb{R}$ is denoted as $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$.
  - $B(\mathbb{R}^d)$ is the set of bounded real functions over $\mathbb{R}^d$, i.e., $f \in B(\mathbb{R}^d)$ if, and only if, $\|f\|_\infty < \infty$.
  - $C_b(\mathbb{R}^d)$ is the set of continuous and bounded real functions over $\mathbb{R}^d$.
- Sets.
  - Given a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, a Borel set $A \in \mathcal{B}(\mathbb{R}^d)$ and the indicator function

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

    $\mu(A) = (I_A, \mu) = \int_A \mu(dx)$ is the probability of A.
  - The Lebesgue measure of a set $A \in \mathcal{B}(\mathbb{R}^d)$ is denoted $\mathcal{L}(A)$.

  - For a set $A \in \mathbb{R}^d$, $A^{\mathsf{c}} = \mathbb{R}^d \backslash A$ denotes its complement.
- Sequences, vectors and random variables (r.v.).
  - We use a subscript notation for sequences, $x_{t_1:t_2} \triangleq \{x_{t_1}, \ldots, x_{t_2}\}$.
  - For an element $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ of an Euclidean space, its norm is denoted as $\|x\| = \sqrt{x_1^2 + \ldots + x_d^2}$.
  - The L_p norm of a real r.v. Z, with $p \geq 1$, is written as $\|Z\|_p \triangleq E[|Z|^p]^{1/p}$, where $E[\cdot]$ denotes expectation.

2.2. Filtering in discrete-time, state-space Markov models

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. The common probability measure for the pair $(\{X_t\}_{t\geq 0}, \{Y_t\}_{t\geq 1})$ is denoted P, and we assume that it is absolutely continuous w.r.t. the Lebesgue measure. We refer to the first sequence as the state 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 $\tau_t : \mathcal{B}(\mathbb{R}^{d_x}) \times \mathbb{R}^{d_x} \to [0, 1]$, defined as

$$\tau_t(A|x_{t-1}) \triangleq \mathsf{P}\left\{X_t \in A | X_{t-1} = x_{t-1}\right\}, \tag{2.2}$$

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 the state X_t, meaning that

$$\mathsf{P}\left\{Y_t \in A | X_{0:t} = x_{0:t}, \{Y_k = y_k\}_{k\neq t}\right\} = \mathsf{P}\left\{Y_t \in A | X_t = x_t\right\}, \tag{2.3}$$

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

$$\gamma_t(A|x_t) \triangleq \mathsf{P}\left\{Y_t \in A | X_t = x_t\right\}, \quad A \in \mathcal{B}(\mathbb{R}^{d_x}), \quad t = 1, 2, \ldots, \tag{2.4}$$

has a positive density w.r.t. the Lebesgue measure. We denote this density as $g_t(y|x)$, hence we write $\gamma_t(A|x_t) = \int_A g_t(y|x_t)dy$.

The filtering problem consists in the computation of the posterior probability measure of the state X_t given a sequence of observations up to time t. Specifically, for a fixed observation record $Y_{1:T} = y_{1:T}$, $T < \infty$, we seek the measures $\pi_t \in \mathcal{P}(\mathbb{R}^{d_x})$ given by

$$\pi_t(A) \triangleq \mathsf{P}\left\{X_t \in A | Y_{1:t} = y_{1:t}\right\}, \quad t = 0, 1, \ldots, T, \tag{2.5}$$

where $A \in \mathcal{B}(\mathbb{R}^{d_x})$. For many practical problems, the interest actually lies in the computation of integrals of the form (f, π_t). Note that, for $t = 0$, we recover the prior measure, i.e., $\pi_0 = \tau_0$.

2.3. Particle filters

The sequence of measures $\{\pi_t\}_{t\geq 1}$ can be numerically approximated using particle filtering. Particle filters are numerical methods based on the recursive decomposition [3]

$$\pi_t = g_t^{y_t} \star \tau_t \pi_{t-1}, \tag{2.6}$$

where $g_t^{y_t} : \mathbb{R}^{d_x} \to \mathbb{R}^+$ is the function defined as $g_t^{y_t}(x) \triangleq g_t(y_t|x)$, $\star$ denotes the projective product and $\xi_t \triangleq \tau_t \pi_{t-1}$ is the (predictive) probability measure

$$\xi_t(A) = \tau_t \pi_{t-1}(A) = \int \tau_t(A|x)\pi_{t-1}(dx), \quad A \in \mathcal{B}(\mathbb{R}^{d_x}). \tag{2.7}$$

Specifically, the simplest particle filter, often called 'standard particle filter' or 'bootstrap filter' [27] (see also [20]), can be described as follows.

1. Initialization. At time $t = 0$ draw N i.i.d. samples from the initial distribution $\tau_0 \equiv \pi_0$, denoted $x_0^{(n)}$, $n = 1, \ldots, N$.
2. Recursive step. Let $\Omega_{t-1}^N = \{x_{t-1}^{(n)}\}_{n=1,\ldots,N}$ be the particles (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 $\tau_t(\cdot|x_{t-1}^{(n)})$ and compute the normalized weight

$$w_t^{(n)} = \frac{g_t^{y_t}(\bar{x}_t^{(n)})}{\sum_{k=1}^N g_t^{y_t}(\bar{x}_t^{(k)})}. \tag{2.8}$$

   (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 [22, 19] but the convergence of the algorithm can be easily proved for various other schemes (see, e.g., the treatment of the resampling step in [8]). Using the samples in $\Omega_t^N = \{x_t^{(n)}\}_{n=1,\ldots,N}$, we construct a random approximation of π_t, namely

$$\pi_t^N(dx_t) = \frac{1}{N}\sum_{n=1}^N \delta_{x_t^{(n)}}(dx_t), \tag{2.9}$$

where $\delta_{x_t^{(n)}}$ is the delta unit-measure located at $X_t = x_t^{(n)}$. For any integrable function f in the state space, it is straightforward to approximate the integral (f, π_t) as

$$(f, \pi_t) \approx (f, \pi_t^N) = \frac{1}{N}\sum_{n=1}^N f(x_t^{(n)}). \tag{2.10}$$

The convergence of particle filters has been analyzed in a number of different ways. Most of the results to be described in this paper rely only on the convergence of the L_p

norm of the approximation errors $(f, \pi_t^N) - (f, \pi_t)$ for bounded functions. Additionally, we establish the a.s. convergence toward 0 of the approximation errors for a class of possibly unbounded functions. Specifically, let f be a real function over the state space and introduce the notation

$$\tau_t(f)(x) = \int f(z)\tau_t(dz|x)$$

for conciseness. Note that $\tau_t(f) : \mathbb{R}^{d_x} \to \mathbb{R}$ is also a real function over the state space. We define the following class of functions.

Definition 2.1. F_T^p *is a family of functions* $f : \mathbb{R}^{d_x} \to \mathbb{R}$ *that satisfy:*

(i) $(f^p, \pi_t) < \infty$ *for* $t = 0, ..., T$, *and*
(ii) if $f \in \mathsf{F}_T^p$ *then* $\tau_t(f^p) \in \mathsf{F}_T^p$ *for* $t = 1, ..., T$.

The set F_T^p includes functions that are p-integrable w.r.t. π_t, $0 \leq t \leq T$, and remain p-integrable when sequentially transformed by the kernels τ_t, $1 \leq t \leq T$. Note that if $p \leq q$ then $F_T^q \subseteq F_T^p$. It turns out that if $f \in \mathsf{F}_T^p$ for some $p \geq 4$, then the error of the particle approximations vanishes for large N at every time step. This is precisely stated by the following proposition.

Proposition 2.1. *Assume that the sequence of observations* $Y_{1:T} = y_{1:T}$ *is fixed, with* T *being some large but finite time horizon,* $g_t^{y_t} \in B(\mathbb{R}^{d_x})$ *and* $g_t^{y_t} > 0$ *(in particular,* $(g_t^{y_t}, \xi_t) > 0$*) for every* $t = 1, 2, ..., T$. *The following results hold.*

(a) For all $f \in B(\mathbb{R}^{d_x})$ *and any* $p \geq 1$,

$$\left\|(f, \pi_t^N) - (f, \pi_t)\right\|_p \leq \frac{c_t \|f\|_\infty}{\sqrt{N}} \tag{2.11}$$

for $t = 0, 1, \ldots, T$, *where* c_t *is a constant independent of* N, $\|f\|_\infty = \sup_{x \in \mathbb{R}^{d_x}} |f(x)|$ *and the expectation is taken over all possible realizations of the random measure* π_t^N. *In particular,*

$$\lim_{N \to \infty} |(f, \pi_t^N) - (f, \pi_t)| = 0 \text{ a.s. for } 0 \leq t \leq T.$$

(b) If $f \in \mathsf{F}_T^4$ *then* $\lim_{N \to \infty} \left|(f, \pi_t^N) - (f, \pi_t)\right| = 0$ *a.s. for* $0 \leq t \leq T$.

See Appendix A for a proof.

***Remark* 2.1.** Part (a) of Proposition 2.1 is fairly standard. A similar proposition was already proved in [17], albeit under additional assumptions on the state-space model. Bounds for $p = 2$ and $p = 4$ can also be found in a number of references (see, e.g., [8, 11, 15]). Part (b) establishes the almost sure convergence for the approximate integrals of unbounded functions (e.g., for the approximation of the posterior mean) as long

as they are "sufficiently integrable". A similar result can be found in [31], including convergence rates. However, the analysis in [31] is carried out for a modified particle filtering algorithm, that involves a rejection test on the generated particles, and cannot be applied to the standard particle filter presented in this section.

3. Particle-kernel approximation of the filtering density

In the sequel we will be concerned with the family of Markov state-space models for which the posterior probability measures $\{\pi_t\}_{t\geq 1}$ are absolutely continuous w.r.t. the Lebesgue measure and, therefore, there exist pdf's $p_t : \mathbb{R}^{d_x} \to [0, +\infty)$, $t = 1, 2, \ldots$, such that $\pi_t(A) = \int_A p_t(x)dx$ for any $A \in \mathcal{B}(\mathbb{R}^{d_x})$. The density p_t is referred to as the filtering pdf at time t. In this section we briefly review the basic methodology for kernel density estimation and then describe the construction of sequences of approximations of p_t using the particles generated by a particle filter and a generic kernel function. The section concludes with the discussion on the relationship between the complexity of the particle filter (i.e., the number of particles N) and the choice of kernel bandwidth for the density estimators.

3.1. Kernel density estimators

In order to build an approximation of the function $p_t(x)$ using a sample of size N, $\{x_t^{(n)}\}_{i=1,\ldots,N}$, we resort to the classical kernel approach commonly used in density estimation [43, 45, 44]. Specifically, given a kernel function $\phi : \mathbb{R}^{d_x} \to \mathbb{R}^+$, we build a regularized density function of the form

$$p_t^N(x) = \frac{1}{N}\sum_{n=1}^{N} \phi(x - x_t^{(n)}). \tag{3.1}$$

In the classical theory, the kernel function ϕ is often taken to be a non-negative and symmetric probability density function with zero mean and finite second order moment. Specifically, the following assumptions are commonly made [43] and we abide by them in this paper

A. 1. *The kernel ϕ is a pdf w.r.t. the Lebesgue measure. In particular, $\phi(x) \geq 0$ $\forall x \in \mathbb{R}^{d_x}$ and $\int \phi(x)dx = 1$.*

A. 2. *The probability distribution with density ϕ has a finite second order moment, i.e., $c_2 = \int \|x\|^2\phi(x)dx < \infty$.*

Given a function ϕ satisfying A.1 and A.2 it is possible to define a family of re-scaled kernels

$$\phi_{\frac{1}{h}}(x) = h^{-d_x}\phi\left(h^{-1}x\right), \tag{3.2}$$

where $h > 0$ is often referred to as the bandwidth of the kernel function. Both the kernel and the bandwidth can be optimized to minimize the mean integrated square error (MISE) between the regularized density and the target densities [45]. Specifically, the MISE is defined as

$$\text{MISE} \equiv \int E\left[\left(p_t(x) - \frac{1}{N}\sum_{n=1}^{N} \phi_{\frac{1}{h}}(x - x_t^{(n)})\right)^2\right] dx, \tag{3.3}$$

where the expectation is taken over the random sample. Although the MISE given in Eq. (3.3) is intractable in general, asymptotic approximations (as $N \to \infty$) are known [45]. Moreover, if we assume that $x_t^{(1)}, ..., x_t^{(N)}$ are i.i.d. and drawn exactly from $p_t(x)$ (beware that this is *not* the case in the particle filtering framework, though), then the MISE is minimized by the Epanechnikov kernel [43]

$$\phi_E(x) = \begin{cases} \frac{d_x+2}{2v_{d_x}}(1 - \|x\|^2), & \text{if } \|x\| < 1 \\ 0, & \text{otherwise} \end{cases}, \tag{3.4}$$

where v_{d_x} is the volume of the unit sphere in $\mathbb{R}^{d_x}$. If, additionally, $p_t(x)$ is Gaussian with unit covariance matrix, then the scaling of ϕ_E that yields the minimum MISE is given by the bandwidth [45]

$$h_{opt} = \left[8v_{d_x}^{-1}(d_x + 4)(2\sqrt{\pi})^{d_x}\right]^{\frac{1}{d_x+4}} N^{-\frac{1}{d_x+4}}.$$

In our case, $p_t(x)$ is not known (it is known not to be Gaussian in general, though) and the random sample $x_t^{(1)}, ..., x_t^{(N)}$ is not drawn from $p_t(x)$, so the standard results of [43, 45, 44] and others cannot be applied directly and a specific analysis is needed [36, 35].

In [36], two regularized particle filtering algorithms were studied, each of them yielding a different kernel estimator of p_t. Using the notation in the present paper, they can be written as

$$p_{t,pre}^N(x) \propto \frac{1}{N}\sum_{n=1}^{N} g_t(y_t|x)\phi_{\frac{1}{h}}(x - \bar{x}_t^{(n)}), \tag{3.5}$$

for the *pre-regularized particle filter*, and

$$p_{t,post}^N(x) = \sum_{n=1}^{N} w_t^{(n)}\phi_{\frac{1}{h}}(x - \bar{x}_t^{(n)}),$$

for the post-regularized particle filter. Note that $p_{t,pre}^N(x)$ is an unnormalized approximation of $p_t(x)$ (the normalization constant cannot be computed in general). For the post-regularized density estimator it can be shown that under certain regularity assumptions [36, Theorem 6.15]

$$E\left[\int \left|p_t(x) - p_{t,post}^N(x)\right| dx \quad | \quad Y_{1:t}\right] \to 0 \quad \text{a.s.}$$

(where the expectation is taken w.r.t. $p^N_{t,post}$) when $N \to \infty$ and $h \to 0$ jointly. Specifically, the mean total variation decreases as $O(N^{-\frac{1}{2}} + h^2)$. A similar result can be shown for $p^N_{t,post}$ [36, Theorem 6.9].

***Remark* 3.1.** Although we use the same notation for the particles, $\bar{x}^{(i)}_t$, $i = 1, ..., N$, as in Section 2.3, the sampling/resampling schemes in the pre-regularized and post-regularized particle filters are different from the basic 'bootstrap' filter [40, 36]. The pre-regularized filter, in particular, involves the use of a rejection sampler.

***Remark* 3.2.** The convergence results in [36] for the post-regularized density estimator $p^N_{t,post}$ hold true when the following assumptions on the state-space model are guaranteed.

- The transition kernel $R_t(x_{t-1}, A) = \int_A g_t^{Y_t}(x)\tau_t(dx|x_{t-1})$ is mixing [36, Definition 3.2].
- The likelihood satisfies $\sup_{u\in \mathsf{W}^{2,1}} \frac{g_t^{Y_t} u}{\|u\|_{2,1}} < \infty$, where $\mathsf{W}^{2,1}$ is the Sobolev space of functions defined on $\mathbb{R}^{d_x}$ which, together with their derivatives up to order 2, are integrable with respect to the Lebesgue measure, and $\|\cdot\|_{2,1}$ is the corresponding norm.
- The measure $\tau_t(dx|x_{t-1})$ is absolutely continuous w.r.t. the Lebesgue measure, with density $\tau_t^{x_{t-1}}(x) \in \mathsf{W}^{2,1}$ and $\sup_{x_{t-1}\in\mathbb{R}^{d_x}} \|\tau_t^{x_{t-1}}\|_{2,1} < \infty$.

Assuming that $\tau_t = \tau$ for every $t \geq 1$ (hence, the Markov state process is homogeneous), the analysis in [35] targets the convergence in total variation distance of the continuous measure $\rho^N_t(x)dx$, where the density estimator ρ^N_t is defined as

$$\rho^N_t(x) = c_t \sum_{n=1}^N g_t(y_t|x)\tau^{x^{(n)}_{t-1}}(x)$$

with normalization constant $c_t = \left(\sum_{n=1}^N \int g_t(y_t|x)\tau^{x^{(n)}_{t-1}}(x)dx\right)^{-1}$. This is similar to the pre-regularized approximation $p^N_{t,pre}$ but using the Markov kernel of the model, τ, for smoothing, instead of the generic kernel $\phi_{\frac{1}{h}}$. Although in most problems it is possible to draw from $\tau^{x_{t-1}}$, it is often not possible to evaluate it and, in such cases, the approximation ρ^N_t is not practical. Also note that ρ^N_t is *not* a kernel density estimator of p_t in the classical form of Eq. (3.1). The sample of size N from which the approximation is constructed corresponds to the variable X_{t-1}, rather than X_t, and smoothing is achieved by way of a prediction step (using the Markov kernel τ). It is *not* possible, in general, to write $\rho^N_t(x) \propto \sum_{n=1}^N g_t(y_t|x)\phi_{\frac{1}{h}}(x - x^{(n)}_{t-1})$ for some kernel function ϕ. Under regularity assumptions on g_t and τ, it is proved in [35, Theorem 2] that

$$\mathsf{P}\left\{\int |\rho^N_t(x) - p_t(x)|dx > \epsilon\right\} \leq c_1 \exp\{-c_2 N\}, \quad t \geq 1, \tag{3.6}$$

for any $\epsilon > 0$ and some constants $c_1, c_2 > 0$.

***Remark* 3.3.** The regularity assumptions on the state-space model in [35, Theorem 2] are the following.

(a) There are pdf's $\{b_t\}_{t\geq 1}$ and two constants $0 < c_\tau < C_\tau < \infty$ such that

$$c_\tau b_t(x) \leq \tau^{x_{t-1}}(x) \leq C_\tau b_t(x), \quad \text{for all } x, t.$$

(b) The likelihood g_t satisfies that $\sup_{t\geq 1; x,x'\in\mathbb{R}^{d_x}; y\in\mathbb{R}^{d_y}} \frac{g_t(y|x)}{g_t(y|x')} < \infty$.

The assumption in (a) excludes, e.g., models of the form $X_t = h(X_{t-1}) + V_t$ where the function $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ is not bounded or the noise process V_t is Gaussian [35, Section 4.2]. The assumption in (b) is also stronger than required for Proposition 2.1 to hold true.

3.2. Approximation of the filtering density and its derivatives

We investigate particle-kernel approximations of p_t constructed from a kernel function ϕ and the samples $x_t^{(n)}$, $n = 1, ..., N$, generated by the particle filter. Instead of restricting our attention to procedures based on a single kernel, however, we consider a sequence of functions $\phi_k : \mathbb{R}^{d_x} \to \mathbb{R}^+$, $k \in \mathbb{N}$, defined according to the notation in Eq. (3.2), i.e., $\phi_k(x) = k^{d_x}\phi(kx)$. If ϕ complies with A.1 and A.2, then we have similar properties for ϕ_k. Trivially, $\phi_k(x) \geq 0\ \forall x \in \mathbb{R}^{d_x}$, and it is also straightforward to check that $\int \phi_k(x)dx = 1$. Moreover, if we apply the change of variable $y = kx$ and note that $dy = k^{d_x}dx$, then

$$\int \|x\|^2\phi_k(x)dx = \frac{1}{k^2}\int \|y\|^2\phi(y)dy = \frac{c_2}{k^2},$$

from A.2.

The approximation of p_t generated by the particles $x_t^{(n)}$, $n = 1, ..., N$, and the k-th kernel, ϕ_k, is denoted as p_t^k and has the form

$$p_t^k(x) \triangleq \frac{1}{N}\sum_{n=1}^{N} \phi_k(x - x_t^{(n)}) = (\phi_k^x, \pi_t^N),$$

where $\phi_k^x(x') \triangleq \phi_k(x - x')$. Beware that, in our notation, we skip the dependence of p_t^k on the number of particles, N, for the sake of simplicity. In Section 4.1 we will assume a certain relationship between N and k that will be carried on through the rest of the paper and justifies the omission in the notation. Let us also remark that we do not construct p_t^k in order to approximate integrals w.r.t. the filtering measure (this is more efficiently achieved using Eq. (2.10)). Instead, we aim at applications where an explicit approximation of the density p_t is necessary. Some examples are considered in Section 5.

In order to investigate the approximation of derivatives of p_t, let us consider the multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{d_x}) \in \mathbb{N}^* \times \mathbb{N}^* \times \ldots \times \mathbb{N}^*$, where $\mathbb{N}^* = \mathbb{N} \cup \{0\}$, and introduce the partial derivative operator D^α defined as

$$D^\alpha h \triangleq \frac{\partial^{\alpha_1} \cdots \partial^{\alpha_{d_x}} h}{\partial x_1^{\alpha_1} \cdots \partial x_{d_x}^{\alpha_{d_x}}}$$

for any (sufficiently differentiable) function $h : \mathbb{R}^{d_x} \to \mathbb{R}$. The order of the derivative $D^\alpha h$ is denoted as $|\alpha| = \sum_{i=1}^{d_x} \alpha_i$. We are interested in the approximation of functions $D^\alpha p_t(x)$ which are continuous, as explicitly given below.

A. 3. *For every x in the domain of $p_t(x)$, $D^\alpha p_t(x)$ exists and is Lipschitz continuous, i.e., there exists a constant $c_{\alpha,t} > 0$ such that*

$$|D^\alpha p_t(x - z) - D^\alpha p_t(x)| \leq c_{\alpha,t}\|z\|$$

for all $x, z \in \mathbb{R}^{d_x}$.

***Remark* 3.4.** It is possible to check whether A.3 holds by inspecting the transition kernel τ_t and the likelihood function $g_t^{y_t}$. For example, assume that $\tau_t(dx|x')$ has an associated density w.r.t. the Lebesgue measure, denoted $\tau_t^{x'}$. A sufficient condition for $D^\alpha p_t$ to be Lipschitz is that both $g_t^{y_t}$ and $\tau_t^{x'}$ be bounded with bounded derivatives up to order $1 + |\alpha|$. Specifically, it is sufficient that $g_t^{y_t} \in B(\mathbb{R}^{d_x})$ and, for any $\beta = (\beta_1, ..., \beta_{d_x})$ such that $0 \leq |\beta| \leq 1 + |\alpha|$, $D^\beta g_t^{y_t} \in B(\mathbb{R}^{d_x})$ and there exist constants c_β, independent of x and x', such that $D^\beta \tau_t^{x'} \leq c_\beta$.

For the same α, we also impose the following condition on the kernel ϕ.

A. 4. *$D^\alpha \phi \in C_b(\mathbb{R}^{d_x})$, i.e., $D^\alpha \phi$ is a continuous and bounded function. In particular, $\|D^\alpha \phi\|_\infty = \sup_{x \in \mathbb{R}^{d_x}} |D^\alpha \phi(x)| < \infty$.*

***Remark* 3.5.** Trivially, if $D^\alpha \phi \in C_b(\mathbb{R}^{d_x})$ then $D^\alpha \phi_k \in C_b(\mathbb{R}^{d_x})$ for any finite k. In particular, $\|D^\alpha \phi_k\|_\infty = k^{d_x + |\alpha|} \|D^\alpha \phi\|_\infty$.

The approximation of $D^\alpha p_t$ computed from the samples $x_t^{(n)}$, $n = 1, ..., N$, and the k-th kernel, ϕ_k, has the form

$$D^\alpha p_t^k(x) = \frac{1}{N} \sum_{n=1}^{N} D^\alpha \phi_k^x(x_t^{(n)}) = (D^\alpha \phi_k^x, \pi_t^N). \tag{3.7}$$

3.3. Complexity of the particle filter and choice of kernel bandwidth

In the sequel, we will be concerned with the convergence of the sequence of approximations $\{D^\alpha p_t^k\}_{k \geq 1}$ under the generic assumptions A.1–A.4. The convergence results introduced in Sections 4 and 5 are given either as limits, for $k \to \infty$, or as error bounds that decrease with k.

Recall, however, that $p_t^k(x) = (\phi_k^x, \pi_t^N)$, i.e., the density estimator p_t^k depends both on the kernel bandwidth $h = \frac{1}{k}$ and the number of particles N. A distinctive feature of the analysis in Sections 4 and 5 is that it links both indices by way of the

inequality $N \geq k^{2(d_x+|\alpha|+1)}$, where $|\alpha| = \sum_{i=1}^{d_x} \alpha_i$ is the order of the derivative D^α. For $\alpha = (0, \ldots, 0)$, $D^\alpha p_t^k = p_t^k$ and

$$N \geq k^{2(d_x+1)}. \tag{3.8}$$

Obviously, $k \to \infty$ implies that $N \to \infty$ and $h \to 0$.

This connection is useful to provide simple bounds for the approximation errors, but also because it yields guidance for the numerical implementation of the density estimators. In particular, for $|\alpha| = 0$ and a fixed kernel bandwidth $h = \frac{1}{k}$, the inequality in (3.8) determines the minimum number of particles N that are needed in the particle filter in order to guarantee that convergence, at the rates given by the Theorems of Sections 4 and 5, holds. A lesser number of samples (i.e., some $N < k^{2(d_x+1)}$) would result in an under-smoothed density $p_t^k(x)$ with a bigger approximation error.

If the computational complexity of the particle filter is limited by practical considerations, then N is given and the error bounds to be introduced only hold when $k \leq N^{\frac{1}{2(d_x+1)}}$ or, equivalently, when the kernel bandwidth is lower-bounded as $h = \frac{1}{k} \geq N^{-\frac{1}{2(d_x+1)}}$. A smaller bandwidth would, again, result in an under-smoothed approximation $p_t^k(x)$. On the other hand, since over-smoothing also increases the approximation error of kernel density estimators [43], it is convenient to choose the smallest possible bandwidth h. For given N, we should therefore select[2] $h = h(N) = N^{-\frac{1}{2(d_x+1)}}$.

4. Convergence of the approximations

Starting from Proposition 2.1, we prove that the kernel approximations of the filtering pdf, $p_t^k(x)$, and its derivates converge a.s. for every x in the domain of p_t, both point-wise and uniformly on $\mathbb{R}^{d_x}$. We also prove that the smoothed approximating measure $\breve{\pi}_t^{N(k)}(dx) = p_t^k(x)dx$ converges to π_t in total variation distance and that the integrated square error of a sequence of truncated density estimators converges quadratically (in k) toward 0 a.s. Explicit convergence rates for the approximations are given.

4.1. Almost sure convergence

In this section we obtain convergence rates for the particle-kernel approximation $D^\alpha p_t^k(x)$ of Eq. (3.7). Depending on the support of the density $p_t(x)$, these rates may be point-wise or uniform (for all x). In both cases, convergence is attained a.s. based on the following auxiliary result.

Lemma 4.1. *Let $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ be a sequence of non-negative random variables such that, for $p \geq 2$,*

$$E\left[(\theta^k)^p\right] \leq \frac{c}{k^{p-\nu}}, \tag{4.1}$$

[2] In practice, an adaptive choice of the kernel bandwidth (see, e.g., [5, 47]) is generally more efficient. In this paper, however, we restrict our attention to fixed-bandwidth kernels.

where $c > 0$ and $0 \leq \nu < 1$ are constant w.r.t. k. Then, there exists a non-negative and a.s. finite random variable U^{ε}, independent of k, such that

$$\theta^k \leq \frac{U^{\varepsilon}}{k^{1-\varepsilon}}, \tag{4.2}$$

where $\frac{1+\nu}{p} < \varepsilon < 1$ is also a constant w.r.t. k.

Proof: See Appendix B.

***Remark* 4.1.** In Lemma 4.1, if the inequality (4.1) holds for all $p \geq 2$ then the constant ε in (4.2) can be made arbitrarily small, i.e., we can choose $0 < \varepsilon < 1$.

Using Lemma 4.1, it is possible to prove that $D^{\alpha}p_t^k(x) \to D^{\alpha}p_t(x)$ a.s. and obtain explicit convergence rates. In order to establish a connection between the sequence of kernels $\phi_k(x)$, $k \in \mathbb{N}$, and the sequence of measure approximations π_t^N, $N \in \mathbb{N}$, we define the number of particles to be a function of the kernel index and denote it as $N(k)$. To be specific, for a given multi-index α, we assume that $N(k) \geq k^{2(d_x+|\alpha|+1)}$. In this way, all the convergence rates to be presented in this paper are primarily given in terms of the kernel index k. We first show that $D^{\alpha}p_t^k \to D^{\alpha}p_t$ point-wise for $x \in \mathbb{R}^{d_x}$.

Theorem 4.1. *Under assumptions A.1, A.2, A.3, A.4 and $N(k) \geq k^{2(d_x+|\alpha|+1)}$, the inequality*

$$\left|D^{\alpha}p_t^k(x) - D^{\alpha}p_t(x)\right| \leq \frac{V^{x,\alpha,\varepsilon}}{k^{1-\varepsilon}} \tag{4.3}$$

holds true, with $V^{x,\alpha,\varepsilon}$ an a.s. finite, non-negative random variable and a constant $0 < \varepsilon < 1$. In particular,

$$\lim_{k\to\infty} \left|D^{\alpha}p_t^k(x) - D^{\alpha}p_t(x)\right| = 0 \quad a.s. \tag{4.4}$$

Proof: Let us construct an approximation of $p_t(x)$ using the kernel ϕ_k and the true filtering measure π_t, namely, $\tilde{p}_t^k(x) = (\phi_k^x, \pi_t)$. Since $\pi_t(dx) = p_t(x)dx$, the approximation $\tilde{p}_t^k$ is actually a convolution integral and can be written in two alternative ways using the commutative property, namely

$$\tilde{p}_t^k(x) = \int \phi_k(x-z)p_t(z)dz = \int \phi_k(z)p_t(x-z)dz. \tag{4.5}$$

Let us now consider the derivative $D^{\alpha}p_t$. If we apply the operator D^{α} to $\tilde{p}_t^k$ in (4.5) we readily obtain

$$D^{\alpha}\tilde{p}_t^k(x) = \int \phi_k(z)D^{\alpha}p_t(x-z)dz$$

and, using the latter expression, we find an upper bound for the error $|D^\alpha \tilde{p}_t{}^k(x) - D^\alpha p_t(x)|$. In particular,

$$\begin{aligned}
\left|D^\alpha \tilde{p}_t{}^k(x) - D^\alpha p_t(x)\right| &= \left|\int \phi_k(z) D^\alpha p_t(x-z)dz - D^\alpha p_t(x)\right| \\
&\leq \int \phi_k(z) \left|D^\alpha p_t(x-z) - D^\alpha p_t(x)\right| dz \qquad (4.6)\\
&\leq c_{\alpha,t} \int \phi_k(z)\|z\| dz \qquad (4.7)\\
&\leq c_{\alpha,t}\sqrt{\int \phi_k(z)\|z\|^2 dz} \qquad (4.8)\\
&= c_{\alpha,t}\frac{\sqrt{c_2}}{k}, \qquad (4.9)
\end{aligned}$$

where Eq. (4.6) follows from A.1 (namely, $\phi \geq 0$), (4.7) is obtained from the Lipschitz assumption A.3, (4.8) follows from Jensen's inequality and, finally, the bound in (4.9) is obtained from assumption A.2. Note that $c_{\alpha,t}$ and c_2 are constants with respect to both x and k. As a consequence of (4.9),

$$\lim_{k\to\infty} D^\alpha \tilde{p}_t^k(x) = D^\alpha p_t(x).$$

Consider now the approximation, with $N(k)$ particles, $D^\alpha p_t^k = (D^\alpha \phi_k^x, \pi_t^{N(k)})$ of the integral $(D^\alpha \phi_k^x, \pi_t)$. From Proposition 2.1 and assumption A.4 we obtain

$$\begin{aligned}
\left\|D^\alpha p_t^k(x) - D^\alpha \tilde{p}_t^k(x)\right\|_p &= \left\|(D^\alpha \phi_k^x, \pi_t^{N(k)}) - (D^\alpha \phi_k^x, \pi_t)\right\|_p \\
&\leq \frac{\bar{c}_t k^{d_x+|\alpha|}\|D^\alpha \phi\|_\infty}{\sqrt{N(k)}}, \qquad (4.10)
\end{aligned}$$

where we have used Remark 3.5 and the constant $\bar{c}_t$ is independent of $N(k)$ and x.

A straightforward application of the triangle inequality now yields

$$\left\|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right\|_p \leq \left\|D^\alpha p_t^k(x) - D^\alpha \tilde{p}_t^k(x)\right\|_p + \left\|D^\alpha \tilde{p}_t^k(x) - D^\alpha p_t(x)\right\|_p. \qquad (4.11)$$

The first term on the right-hand side of (4.11) can be bounded using (4.10), while the second term also has an upper bound given by[3] (4.9). Taking both bounds together, we arrive at

$$\left\|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right\|_p \leq \frac{\bar{c}_t k^{d_x+|\alpha|}\|D^\alpha \phi\|_\infty}{\sqrt{N(k)}} + \frac{c_{\alpha,t}\sqrt{c_2}}{k} \leq \frac{\bar{c}_{\alpha,t}}{k}, \qquad (4.12)$$

[3]Note that $\left\|D^\alpha \tilde{p}_t^k(x) - D^\alpha p_t(x)\right\|_p = \left|D^\alpha \tilde{p}_t^k(x) - D^\alpha p_t(x)\right|$ because $D^\alpha \tilde{p}_t{}^k(x)$ does not depend on $\pi_t^{N(k)}$.

where the second inequality follows from the assumption $N(k) \geq k^{2(d_x+|\alpha|+1)}$ and $\bar{c}_{\alpha,t} = \bar{c}_t \|D^\alpha \phi\|_\infty + c_{\alpha,t}\sqrt{c_{2,\alpha}}$ is a constant.

The inequality (4.12) immediately yields

$$E\left[\left|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right|^p\right] \leq \frac{\bar{c}_{\alpha,t}^p}{k^p} \tag{4.13}$$

and we can apply Lemma 4.1, with $\theta^k = \left|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right|$, $\nu = 0$ and arbitrarily large $p \geq 2$, to obtain

$$\left|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right| \leq \frac{V^{\alpha,x,\varepsilon}}{k^{1-\varepsilon}}, \tag{4.14}$$

where $V^{\alpha,x,\varepsilon}$ is a non-negative and a.s. finite random variable and $0 < \varepsilon < 1$ is a constant, both of them independent of k. The limit in Eq. (4.4) follows immediately from the inequality (4.14).

□

***Remark* 4.2.** The convergence rate for the approximation error $\|D^\alpha p_t^k(x) - D^\alpha p_t(x)\|_p$ given by inequality (4.12) can be improved if we place additional assumptions on the filter density and the kernel, and increase the number of particles $N(k)$. In particular, if in addition to A.1–A.4 we assume that

- $p_t(x)$ has continuous and bounded derivatives up to order $|\alpha| + 2$,
- the kernel satisfies $\int z_i \phi(z) dz = 0$, for $i = 1, ..., d_x$, and
- $N(k) \geq k^{2(d_x+|\alpha|+2)}$,

then it can be shown, using the multivariate version of Taylor's theorem, that

$$\|D^\alpha p_t^k(x) - D^\alpha p_t(x)\|_p \leq \frac{\bar{C}_{\alpha,t}}{k^2}$$

for some constant $\bar{C}_{\alpha,t}$ independent of k. A specific result that relies on these extended assumptions is given in Theorem 4.6 (see Section 4.3).

***Remark* 4.3.** The constant $\bar{c}_{\alpha,t}$ of Eq. (4.12) is independent of the index k and the point $x \in \mathbb{R}^{d_x}$. The random variable $V^{\alpha,x,\varepsilon}$ is also independent of the kernel index k, as explicitly given by Lemma 4.1. However, it may depend on the multi-index α, the dimension of the state space d_x and the point x where the derivative of the density is approximated, hence the notation.

***Remark* 4.4.** For $\alpha = (0, \ldots, 0) = \mathbf{0}$, the inequality (4.3) implies that that we can construct a particle approximation of $p_t(x)$ that converges point-wise. In particular, $D^{\mathbf{0}} p_t(x) = p_t(x)$ and $D^{\mathbf{0}} p_t^k(x) = p_t^k(x) = (\phi_k^x, \pi_t^{N(k)})$, hence Eq. (4.4) becomes

$$\lim_{k\to\infty} |p_t^k(x) - p_t(x)| = 0 \quad \text{a.s.} \tag{4.15}$$

for every $x \in \mathbb{R}^{d_x}$.

***Remark* 4.5.** The proof of Theorem 4.1 does not demand that the assumptions A.3, A.4 and $N(k) \geq k^{2(d_x+|\alpha|+1)}$ hold for every possible α, but only for the particular derivative we need to approximate. For instance, if we only aim to approximate $p_t(x)$ (i.e., $\alpha = \mathbf{0}$), assumption A.2 implies that the distribution with density ϕ must have a finite second order moment, assumption A.3 means that p_t must be Lipschitz, assumption A.4 implies that the basic kernel function ϕ must be continuous and bounded, and it suffices that the number of particles satisfies the inequality $N(k) \geq k^{2(d_x+1)}$.

Most of the results to be given in the remaining of this paper are conditional on the assumptions A.1, A.2, A.3, A.4 and $N(k) \geq k^{2(d_x+|\alpha|+1)}$, the same as Theorem 4.1. However, they refer only to properties of p_t and its first order derivatives and, as a consequence, it is enough to assume that A.3 and A.4 hold true for $\alpha = \mathbf{0}$ and $\alpha = \mathbf{1} = (1, \ldots, 1)$ alone. For the same reason, it suffices to assume $N(k) \geq k^{2(2d_x+1)}$.

Through the rest of the paper, we say that the "standard conditions" are satisfied when

- A.1 and A.2 hold true;
- A.3 and A.4 hold true for, at least, $\alpha = \mathbf{0}$ and $\alpha = \mathbf{1}$; and
- $N(k) \geq k^{2(2d_x+1)}$.

If we restrict x to take values on a sequence of compact subsets of $\mathbb{R}^{d_x}$, then we can obtain a convergence rate for the error $|p_t^k(x) - p_t(x)|$ that is uniform on x, instead of point-wise like in Theorem 4.1. For the following result we fix $p \geq 2$ and consider the sequence of hypercubes

$$\mathcal{K}_k = [-M_k, +M_k] \times \ldots \times [-M_k, +M_k] \subset \mathbb{R}^{d_x},$$

where $M_k = \frac{1}{2} k^{\frac{\beta}{d_x p}}$, and $0 \leq \beta < 1$ is a positive constant independent of k. Note that, for any fixed p and $\beta > 0$, $\lim_{k\to\infty} \mathcal{K}_k = \mathbb{R}^{d_x}$.

Theorem 4.2. *If the standard conditions are satisfied, then*

$$\sup_{x\in\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| \leq \frac{U^\varepsilon}{k^{1-\varepsilon}},$$

where $U^\varepsilon \geq 0$ is an a.s. finite random variable and $0 < \varepsilon < 1$ is a constant, both of them independent of k and x. In particular,

$$\lim_{k\to\infty} \sup_{x\in\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| = 0 \quad a.s.$$

Proof: For any $x = (x_1, \ldots, x_{d_x}) \in \mathcal{K}_k$ and a function $f : \mathbb{R}^{d_x} \to \mathbb{R}$ continuous, bounded and differentiable,

$$f(x) - f(0) = \int_{-M_k}^{x_1} \cdots \int_{-M_k}^{x_{d_x}} D^{\mathbf{1}} f(z) dz - \int_{-M_k}^{0} \cdots \int_{-M_k}^{0} D^{\mathbf{1}} f(z) dz.$$

In particular, for $x_i \in [-M_k, M_k]$, $i = 1, ..., d_x$, and the assumption A.4 with $\alpha = \mathbf{1}$,

$$\left|p_t^k(x) - p_t(x)\right| \leq 2 \int_{-M_k}^{M_k} \cdots \int_{-M_k}^{M_k} \left|D^{\mathbf{1}} p_t^k(z) - D^{\mathbf{1}} p_t(z)\right| dz + \left|p_t^k(0) - p_t(0)\right| \tag{4.16}$$

and, as a consequence,

$$\sup_{x \in \mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| \leq 2A^k + \left|p_t^k(0) - p_t(0)\right|, \tag{4.17}$$

where

$$A^k = \int_{-M_k}^{M_k} \cdots \int_{-M_k}^{M_k} \left|D^{\mathbf{1}} p_t^k(z) - D^{\mathbf{1}} p_t(z)\right| dz.$$

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

$$\left(\frac{1}{2^{d_x} M_k^{d_x}} A^k\right)^p \leq \frac{1}{2^{d_x} M_k^{d_x}} \int_{-M_k}^{M_k} \cdots \int_{-M_k}^{M_k} \left|D^{\mathbf{1}} p_t^k(z) - D^{\mathbf{1}} p_t(z)\right|^p dz,$$

hence

$$\left(A^k\right)^p \leq 2^{d_x(p-1)} M_k^{d_x(p-1)} \sum_{\ell=0}^{2^{d_x}-1} \int_{-M_k}^{M_k} \cdots \int_{-M_k}^{M_k} \left|D^{\mathbf{1}} p_t^k(z) - D^{\mathbf{1}} p_t(z)\right|^p dz. \tag{4.18}$$

Since, from inequality (4.12) in the proof of Theorem 4.1,

$$E\left[\left|D^{\mathbf{1}} p_t^k(s_\ell(z)) - D^{\mathbf{1}} p_t(s_\ell(z))\right|^p\right] \leq \frac{\bar{c}_{\mathbf{1},t}^p}{k^p}, \tag{4.19}$$

we can combine (4.19) and (4.18) to arrive at

$$E\left[(A^k)^p\right] \leq \frac{2^{d_x p} M_k^{d_x p} \bar{c}_{\mathbf{1},t}^p}{k^p} = \frac{\bar{c}_{\mathbf{1},t}^p}{k^{p-\beta}},$$

where the equality follows from the relationship $M_k = \frac{1}{2} k^{\frac{\beta}{d_x p}}$. Using Lemma 4.1 with $\theta_k = A^k$, $p \geq 2$, $\nu = \beta$ and $c = \bar{c}_{\mathbf{1},t}^p$, we obtain a constant $\varepsilon_1 \in \left(\frac{1+\beta}{p}, 1\right)$ and a non-negative and a.s. finite random variable V^{A,ε_1}, both of them independent of k, such that

$$A^k \leq \frac{V^{A,\varepsilon_1}}{k^{1-\varepsilon_1}}. \tag{4.20}$$

Since, from Proposition 2.1,

$$E\left[\left|p_t^k(x) - p_t(x)\right|^p\right] \leq \frac{\bar{c}_{\mathbf{0},t}^p}{k^p},$$

we can apply Lemma 4.1 again, with $\theta^k = |p_t^k(0) - p_t(0)|$, $p \geq 2$, $\nu = 0$ and $c = \bar{c}_{\mathbf{0},t}^p$ to obtain that

$$\left|p_t^k(0) - p_t(0)\right| \leq \frac{V^{p_t(0),\varepsilon_2}}{k^{1-\varepsilon_2}}, \tag{4.21}$$

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

If we choose $\varepsilon = \varepsilon_1 = \varepsilon_2 \in \left(\frac{1+\beta}{p}, 1\right)$ and define $U^\varepsilon = V^{A,\varepsilon_1} + V^{p_t(0),\varepsilon_2}$, then the combination of Eqs. (4.17), (4.20) and (4.21) yields

$$\sup_{x \in \mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| \leq \frac{U^\varepsilon}{k^{1-\varepsilon}},$$

where U^ε is a.s. finite. Note that U^ε and ε are independent of k. Moreover, we can choose p as large as we wish and $\beta > 0$ as small as needed, hence we can select $\varepsilon \in (0, 1)$.

□

***Remark* 4.6.** Assuming that A.3 and A.4 hold for the multi-index $\alpha' = \alpha + \mathbf{1}$, the argument of the proof of Theorem 4.2 can also be adapted to show that

$$\sup_{x \in \mathcal{K}_k} \left|D^\alpha p_t^k(x) - D^\alpha p_t(x)\right| \leq \frac{\tilde{U}^\varepsilon}{k^{1-\varepsilon}},$$

where the constant $0 < \varepsilon < 1$ and the a.s. finite random variable $\tilde{U}^\varepsilon \geq 0$ are independent of k.

***Remark* 4.7.** Theorem 4.2 also holds for a fixed compact subset $\mathcal{K} \subset \mathbb{R}^{d_x}$ instead of the sequence $\mathcal{K}_1, \mathcal{K}_2, \ldots$ In particular, the presented proof is easily adapted to a fixed hypercube $\mathcal{K} = [-M, +M] \times \ldots [-M, +M]$. Therefore,

$$\sup_{x \in \mathcal{K}} \left|p_t^k(x) - p_t(x)\right| \leq \frac{\tilde{U}^\varepsilon}{k^{1-\varepsilon}}, \tag{4.22}$$

where the constant $0 < \varepsilon < 1$ and the a.s. finite random variable $\tilde{U}^\varepsilon \geq 0$ are independent of k.

4.2. Convergence in total variation distance

The total variation distance (TVD) between two measures $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ on the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ is defined as

$$d_{TV}(\mu_1, \mu_2) \triangleq \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mu_1(A) - \mu_2(A)|.$$

Correspondingly, a sequence of measures $\mu^n \in \mathcal{P}(\mathbb{R}^d)$ converges toward $\mu \in \mathcal{P}(\mathbb{R}^d)$ in TVD when $\lim_{n\to\infty} d_{TV}(\mu^n, \mu) = 0$. It can be shown that if μ^n and μ have densities w.r.t. the Lebesgue measure, denoted q^n and q, respectively, then

$$d_{TV}(\mu^n, \mu) = \frac{1}{2} \int |q^n(x) - q(x)|\, dx$$

and, therefore, the sequence μ^n converges to μ in TVD if, and only if,

$$\lim_{n\to\infty} \int |q^n(x) - q(x)|\, dx = 0. \tag{4.23}$$

Consider the smooth approximating measures

$$\breve{\pi}_t^{N(k)}(dx) = p_t^k(x)dx, \quad k = 1, 2, \ldots$$

In this section we show that the sequence $\breve{\pi}_t^{N(k)}$ converges toward π_t in TVD, as $k \to \infty$, by proving first that $\int \left|p_t^k - p_t\right| dx \to 0$ under the same assumptions of Theorem 4.2. This result is established by Theorem 4.3 below. The same as in the proof of Theorem 4.2, we consider an increasing sequence of hypercubes $\mathcal{K}_1 \subset \ldots \subset \mathcal{K}_k \subset \ldots \subset \mathbb{R}^{d_x}$, where $\mathcal{K}_k = [-M_k, +M_k] \times \ldots \times [-M_k, +M_k]$ and $M_k = \frac{1}{2}k^{\frac{\beta}{d_x p}}$, with constants $0 < \beta < 1$ and $p > 3$. Also, recall that, for a set $A \in \mathbb{R}^d$, $A^{\mathsf{c}} = \mathbb{R}^d \backslash A$ denotes its complement and, given a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\mu(A) = \int_A \mu(dx)$ is the probability of A.

Theorem 4.3. *If the standard conditions are satisfied and $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b}{2}k^{-\gamma}$, where $b > 0$ and $\gamma > 0$ are arbitrary but constant w.r.t. k, then*

$$\int \left|p_t^k(x) - p_t(x)\right| dx < \frac{Q^\varepsilon}{k^{\min\{1-\varepsilon,\gamma\}}},$$

where $Q^\varepsilon > 0$ is an a.s. finite random variable and $0 < \varepsilon < 1$ is a constant, both of them independent of k. In particular,

$$\lim_{k\to\infty} \int \left|p_t^k(x) - p_t(x)\right| dx = 0 \quad a.s.$$

and, as a consequence,

$$\lim_{k\to\infty} d_{TV}(\breve{\pi}_t^{N(k)}, \pi_t) = 0 \quad a.s.$$

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

$$\begin{aligned}
\int \left|p_t^k(x) - p_t(x)\right| dx &= \int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx + \int_{\mathcal{K}_k^{\mathsf{c}}} \left|p_t^k(x) - p_t(x)\right| dx \\
&\leq \int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx + 2\int_{\mathcal{K}_k^{\mathsf{c}}} p_t(x)dx \\
&\quad + \int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t^k(x) - p_t(x)\right) dx,
\end{aligned}$$

where the equality follows from $\mathcal{K}_k \cup \mathcal{K}_k^{\mathsf{c}} = \mathbb{R}^{d_x}$ and the inequality is obtained from the fact that p_t and p_t^k are non-negative. Moreover

$$\int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t^k(x) - p_t(x)\right) dx \leq \int_{\mathcal{K}_k^{\mathsf{c}}} \left|p_t^k(x) - p_t(x)\right| dx \leq \int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx$$

hence

$$\int \left|p_t^k(x) - p_t(x)\right| dx \leq 2 \int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx + 2 \int_{\mathcal{K}_k^{\mathsf{c}}} p_t(x) dx \tag{4.24}$$

The first term on the right-hand side (4.24) can be bounded by

$$\int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx \leq \mathcal{L}(\mathcal{K}_k) \sup_{x \in \mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right|, \tag{4.25}$$

where $\mathcal{L}(\mathcal{K}_k) = (2M_k)^{d_x} = k^{\frac{\beta}{p}}$ is the Lebesgue measure of $\mathcal{K}_k$. From Theorem 4.2, the supremum in (4.25) can be bounded as $\sup_{x \in \mathcal{K}_k} |p_t^k(x) - p_t(x)| \leq V^{\varepsilon_1}/k^{1-\varepsilon_1}$, where $V^{\varepsilon_1} \geq 0$ is an a.s. finite random variable and $\frac{1+\beta}{p} < \varepsilon_1 < 1$ is a constant, both independent of k. Therefore, the inequality (4.25) can be extended to yield

$$\int_{\mathcal{K}_k} \left|p_t^k(x) - p_t(x)\right| dx \leq \frac{V^{\varepsilon_1}}{k^{1-\varepsilon_1-\frac{\beta}{p}}} = \frac{V^{\varepsilon}}{k^{1-\varepsilon}}, \tag{4.26}$$

where $\varepsilon = \varepsilon_1 + \frac{\beta}{p}$ and $V^{\varepsilon} = V^{\varepsilon_1}$. If we choose $\varepsilon_1 < 1 - \frac{\beta}{p}$, then $\varepsilon \in \left(\frac{1+2\beta}{p}, 1\right)$. Note that, for $\beta < 1$ and $p > 3$, $1 - \frac{\beta}{p} - \frac{1+\beta}{p} > 1 - \frac{3}{p} > 0$, hence both ε_1 and ε are well defined.

For the second integral in Eq. (4.24), note that $\int_{\mathcal{K}_k^{\mathsf{c}}} p_t(x) dx = \pi_t(\mathcal{K}_k^{\mathsf{c}})$ and, therefore, it can be bounded directly from the assumptions in the Theorem, i.e.,

$$2 \int_{\mathcal{K}_k^{\mathsf{c}}} p_t(x) dx \leq b k^{-\gamma}, \tag{4.27}$$

where $b > 0$ and $\gamma > 0$ are constant w.r.t. k. Putting together Eqs. (4.24), (4.26) and (4.27) yields the desired result.

□

***Remark* 4.8.** The condition $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b}{2} k^{-\gamma}$ in the statement of Theorem 4.3 is satisfied for any t when

- it is satisfied at time $t = 0$, i.e., there exists some constant b_0 such that $\pi_0(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b_0}{2} k^{-\gamma}$,
- the likelihood is bounded, i.e., $g_t^{y_t} \in B(\mathbb{R}^{d_x})$,
- and the kernels $\tau_t(dx|x')$ have sufficiently light tails for every t and every $x' \in \mathbb{R}^{d_x}$.

The latter can be made more precise using a standard induction argument. For example, let $\mathcal{K}_k = [-\frac{1}{2} k^{\frac{\beta}{d_x p}}, +\frac{1}{2} k^{\frac{\beta}{d_x p}}]$ with $p \geq 2$ and $0 \leq \beta < 1$, and assume that for any $x' \in \mathbb{R}^{d_x}$

the kernel τ_t satisfies that $\tau_t(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b(x')}{2} k^{-\gamma}$ for some function $b : \mathbb{R}^{d_x} \to (0, \infty)$. If $b(x')$ can be upper bounded by a polynomial function, say $b(x') \leq c\left(1 + \left(\sum_{i=1}^{d_x} |x_i'|\right)^a\right)$, for some constant $c > 0$ and degree $a < \frac{d_x p(\gamma-1)}{\beta}$, then there exists a constant $b_t < \infty$ such that $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b_t}{2} k^{-\gamma}$.

4.3. Integrated square error

A standard figure of merit for the assessment of kernel density estimators is the mean integrated square error (MISE) [43, 44]. If we assume that both $p_t(x)$ and the kernel $\phi(x)$ take values on a compact set $\mathcal{K}$, then it is relatively simple to prove that the MISE of the sequence of approximations $D^\alpha p_t^k$ converges toward 0 quadratically with the index k. In particular, we have the following result.

Theorem 4.4. *Assume that A.1, A.2, A.3, A.4 and $N(k) \geq k^{2(d_x+|\alpha|+1)}$ hold true. If both $p_t(x)$ and the kernel $\phi(x)$ have a compact support set $\mathcal{K} \subset \mathbb{R}^{d_x}$, then*

$$MISE \equiv \int_{\mathcal{K}} E\left[\left(D^\alpha p_t^k(x) - D^\alpha p_t(x)\right)^2\right] dx \leq \frac{c_{\alpha,\mathcal{K},t}}{k^2},$$

where $c_{\alpha,\mathcal{K},t} > 0$ is constant w.r.t. k.

Proof: Since any compact set is contained in a larger hypercube, we can choose $\mathcal{K} = [-M, +M] \times \cdots \times [-M, +M]$ without loss of generality. Furthermore, since the assumptions of Theorem 4.1 are satisfied, we can recall the inequality in (4.13), which, selecting $p = 2$, yields

$$E\left[\left(D^\alpha p_t^k(x) - D^\alpha p_t(x)\right)^2\right] \leq \frac{\bar{c}_{\alpha,t}^2}{k^2},$$

where the constant $\bar{c}_{\alpha,t}^2$ is independent of k and x. Therefore,

$$\int_{\mathcal{K}} E\left[\left(D^\alpha p_t^k(x) - D^\alpha p_t(x)\right)^2\right] dx \leq \frac{\bar{c}_{\alpha,t}^2}{k^2} \mathcal{L}(\mathcal{K}) \leq \frac{c_{\alpha,\mathcal{K},t}}{k^2},$$

where $\mathcal{L}(\mathcal{K}) = (2M)^{d_x}$ is the Lebesgue measure of $\mathcal{K}$ and $c_{\alpha,\mathcal{K},t} = (2M)^{d_x} \bar{c}_{\alpha,t}^2$.

□

It is also possible to establish a quadratic convergence rate (w.r.t. k) for the integrated square error (ISE) of a sequence of truncated density approximations. In particular, consider the usual hypercubes $\mathcal{K}_k = [-M_k, +M_k] \times \cdots \times [-M_k, +M_k]$ with $M_k = \frac{1}{2} k^{\frac{\beta}{d_x p}}$, for some $p > \frac{5}{2}$ and a constant $0 < \beta < 1$, and define the truncated density estimators

$$p_t^{\top,k}(x) = I_{\mathcal{K}_k}(x) p_t^k(x) = \begin{cases} p_t^k(x), & \text{if } x \in \mathcal{K}_k, \\ 0, & \text{otherwise} \end{cases}.$$

Since $\lim_{k\to\infty} \mathcal{K}_k = \mathbb{R}^{d_x}$, it follows that $\lim_{k\to\infty} |p_t^{\top,k}(x) - p_t^k(x)| = 0$ and we can make $p_t^{\top,k}$ arbitrarily close to the original approximation. The theorem below states that $p_t^{\top,k}$ converges a.s. toward p_t, with a quadratic rate.

Theorem 4.5. *If the standard conditions are satisfied, $p_t \in B(\mathbb{R}^{d_x})$ and $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq bk^{-\gamma}$, where $b > 0$ and $\gamma > 0$ are arbitrary but constant w.r.t. k, then*

$$ISE \equiv \int \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx \leq \frac{U^\varepsilon}{k^{\min\{2-\varepsilon,\gamma\}}},$$

where $U^\varepsilon \geq 0$ is an a.s. finite random variable, independent of k, and $0 < \varepsilon < 2$ is an arbitrarily small constant. In particular,

$$\lim_{k\to\infty} \int \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx = 0 \quad a.s.$$

Proof: We start with the trivial decomposition

$$\begin{aligned}\int \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx &= \int_{\mathcal{K}_k} \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx \\ &\quad + \int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx, \end{aligned} \tag{4.28}$$

where $\mathcal{K}_k^{\mathsf{c}} = \mathbb{R}^{d_x}\backslash\mathcal{K}_k$ is the complement of $\mathcal{K}_k$, and, expanding the square in the last integral of Eq. (4.28), we obtain

$$\begin{aligned}\int \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx &= \int_{\mathcal{K}_k} \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx \\ &\quad + \int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t(x) - p_t^{\top,k}(x)\right) p_t(x) dx \\ &\quad + \int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t^{\top,k}(x) - p_t(x)\right) p_t^{\top,k}(x) dx. \end{aligned} \tag{4.29}$$

In the rest of the proof, we compute upper bounds for each of the integrals on the right-hand side of Eq. (4.29).

For the first term in (4.29) we note that $p_t^{\top,k}(x) = p_t^k(x)$ for all $x \in \mathcal{K}_k$, hence

$$\begin{aligned}\int_{\mathcal{K}_k} \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx &= \int_{\mathcal{K}_k} \left(p_t^k(x) - p_t(x)\right)^2 dx \\ &\leq \mathcal{L}(\mathcal{K}_k) \left(\sup_{x\in\mathcal{K}_k} \left|p_t^{\top,k}(x) - p_t(x)\right|\right)^2, \end{aligned} \tag{4.30}$$

where $\mathcal{L}(\mathcal{K}_k) = (2M_k)^{d_x} = k^{\frac{\beta}{p}}$. Using Theorem 4.2, we obtain an upper bound for the supremum in Eq. (4.30), namely $\sup_{x\in\mathcal{K}_k} |p_t^k(x) - p_t(x)| \leq V^{\varepsilon_1}/k^{1-\varepsilon_1}$, where $V^{\varepsilon_1} \geq 0$ is an a.s. finite random variable and $\frac{1+\beta}{p} < \varepsilon_1 < 1$ is a constant. Both V^{ε_1} and ε_1 are independent of k. We then extend the inequality in (4.30) as

$$\int_{\mathcal{K}_k} \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx \leq k^{\frac{\beta}{p}} \frac{(V^{\varepsilon_1})^2}{k^{2-2\varepsilon_1}} = \frac{\tilde{U}^\varepsilon}{k^{2-\varepsilon}}, \tag{4.31}$$

where $\varepsilon = 2\varepsilon_1 + \frac{\beta}{p}$ and $\tilde{U}^\varepsilon = (V^{\varepsilon_1})^2$. If we choose $\varepsilon_1 < 1 - \frac{\beta}{2p}$, then $\varepsilon \in \left(\frac{2+3\beta}{p}, 2\right)$. Note that, for $\beta < 1$ and $p > \frac{5}{2}$, $2 - \frac{2+3\beta}{p} > 0$, hence ε is well defined.

For the second term on the right-hand side of Eq. (4.29) we simply note that $p_t^{\top,k}(x) = 0$ for all $x \in \mathcal{K}_k^{\mathsf{c}}$ and $p_t(x) < \|p_t\|_\infty < \infty$, since $p_t \in B(\mathbb{R}^{d_x})$. Therefore,

$$\int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t(x) - p_t^{\top,k}(x)\right) p_t(x) dx \leq \|p_t\|_\infty \int_{\mathcal{K}_k^{\mathsf{c}}} p_t(x) dx = \|p_t\|_\infty \pi_t(\mathcal{K}_k^{\mathsf{c}}),$$

and using the assumption $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq bk^{-\gamma}$ we obtain

$$\int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t(x) - p_t^{\top,k}(x)\right) p_t(x) dx \leq \frac{b\|p_t\|_\infty}{k^\gamma}. \tag{4.32}$$

The third term is trivial. Since $p_t^{\top,k}(x) = 0$ for all $x \in \mathcal{K}_k^{\mathsf{c}}$, it follows that

$$\int_{\mathcal{K}_k^{\mathsf{c}}} \left(p_t^{\top,k}(x) - p_t(x)\right) p_t^{\top,k}(x) = 0. \tag{4.33}$$

Substituting Eqs. (4.31), (4.32) and (4.33) into Eq. (4.29) yields

$$\int \left(p_t^{\top,k}(x) - p_t(x)\right)^2 dx \leq \frac{\tilde{U}^\varepsilon}{k^{2-\varepsilon}} + \frac{b\|p_t\|_\infty}{k^\gamma} \leq \frac{U^\varepsilon}{k^{\min\{2-\varepsilon,\gamma\}}},$$

where $U^\varepsilon = \tilde{U}^\varepsilon + b\|p_t\|_\infty$ and $0 < \varepsilon < 2$.

□

The classical asymptotic approximation of the MISE (AMISE) for kernel density estimators built from i.i.d. samples is (see, e.g., [23] and note that we restrict ourselves to diagonal bandwidth matrices)

$$\text{AMISE} \equiv h^4 c(\phi, p_o) + \frac{c(\phi)}{Nh^{d_x}}, \tag{4.34}$$

where $h > 0$ is the bandwidth parameter, $c(\phi, p_o) > 0$ is a constant that depends on the kernel ϕ and the target density (denoted p_o here and assumed twice differentiable), $c(\phi) > 0$ is another constant depending on ϕ alone and N is the number of samples. If we substitute $h = 1/k$ and $N = k^{2d_x+2}$, as given by our analysis, into the expression above, then we find that the MISE converges asymptotically as $\frac{\tilde{c}(\phi,p_o)}{k^4}$, for some constant $\tilde{c}(\phi, p_o) > 0$. We note, however, that

- Eq. (4.34) is only an asymptotic approximation of the MISE, whereas Theorems 4.4 and 4.5 give actual upper bounds for the MISE and the ISE that are valid for every k;
- the AMISE of Eq. (4.34) is derived under the assumption that a size N sample drawn from the density p_t is available [45], whereas Theorems 4.4 and 4.5 hold true for the smoothing of any random measure π_t^N that satisfies $\|(f, \pi_t^N) - (f, \pi_t)\|_p \leq \frac{c\|f\|_\infty}{\sqrt{N}}$ for some constant c and $f \in B(\mathbb{R}^{d_x})$.

Nevertheless, the convergence rate for the MISE in Theorem 4.4 can be improved if we place some additional assumptions on the kernel $\phi(x)$, assume that the filter density is sufficiently smooth and increase the number of particles $N(k)$ in the filter. To be specific, we consider the approximation of $p_t(x)$ alone for clarity and make the following assumptions.

- a.1 The kernel $\phi(x)$ satisfies A.1 ($\phi > 0$, $\int \phi(x)dx = 1$), A.2 ($\int \|x\|^2\phi(x)dx \leq C_2 < \infty$ for some constant C_2) and it is a bounded function. Additionally, $\int x_i\phi(x)dx = 0$ for every $i \in \{1, ..., d_x\}$.
- a.2 The filter density p_t has continuous and bounded derivatives up to order 2, i.e., $D^\alpha p_t \in C_b(\mathbb{R}^{d_x})$ for every α such that $|\alpha| \leq 2$.
- a.3 The number of particles is selected to guarantee that $N = N(k) \geq k^{2(d_x+2)}$.

Then we have the following refinement of Theorem 4.4 for $\alpha = 0$.

Theorem 4.6. *If both $p_t(x)$ and the kernel $\phi(x)$ have a compact support set $\mathcal{K} \subset \mathbb{R}^{d_x}$ and assumptions a.1, a.2 and a.3 hold, then*

$$MISE \equiv \int_{\mathcal{K}} E\left[\left(p_t^k(x) - p_t(x)\right)^2\right] dx \leq \frac{C_{\mathcal{K},t}}{k^4},$$

where $C_{\mathcal{K},t} < \infty$ is constant w.r.t. k.

Proof: Recall the deterministic approximation $\tilde{p}_t^k(x) = (\phi_k^x, \pi_t)$ of $p_t(x)$. Using the multivariate version of Taylor's theorem, the difference $\tilde{p}_t^k(x) - p_t(x)$ can be written as

$$\begin{aligned}\tilde{p}_t^k(x) - p_t(x) &= \int \phi_k(z)\left(p_t(x-z) - p_t(x)\right)dz \\ &= \int \phi_k(z)\left(\sum_{\alpha:|\alpha|=1} D^\alpha p_t(x)(-z)^\alpha + \sum_{\alpha:|\alpha|=2} R_\alpha(x-z)(-z)^\alpha\right)dz, \end{aligned} \tag{4.35}$$

where $z^\alpha = z_1^{\alpha_1} \cdots z_{d_x}^{\alpha_{d_x}}$ and the remainder terms, R_α, satisfy

$$|R_\alpha(x-z)| \leq \max_{\alpha:|\alpha|=2} \|D^\alpha p_t\|_\infty. \tag{4.36}$$

From assumption a.1, $\int \phi_k(z)z_i dz = 0$ for any $1 \leq i \leq d_x$, hence

$$\sum_{\alpha:|\alpha|=1} D^\alpha p_t(x) \int \phi_k(z)(-z)^\alpha dz = -\sum_{i=1}^{d_x} \frac{\partial p_t}{\partial x_i}(x) \int \phi_k(z)z_i dz = 0. \tag{4.37}$$

Substituting (4.37) and (4.36) into (4.35) and taking the absolute value of the difference yields

$$|\tilde{p}_t^k(x) - p_t(x)| \leq \left(\max_{\alpha:|\alpha|=2} \|D^\alpha p_t\|_\infty\right) \sum_{i,j\in\{1,\ldots,d_x\}} \int \phi_k(z)|z_i z_j|dz.$$

However, $\max_{\alpha:|\alpha|=2} \|D^\alpha p_t\|_\infty < \infty$ from assumption a.2 and $\int \phi_k(z)|z_i z_j|dz \leq \frac{C_2}{k^2}$ from assumption a.1. Therefore, we obtain

$$|\tilde{p}_t^k(x) - p_t(x)| \leq \frac{C_{2,t}}{k^2}, \tag{4.38}$$

where the constant $C_{2,t} = \max_{\alpha:|\alpha|=2} \|D^\alpha p_t\|_\infty d_x^2 C_2 < \infty$ is independent of k. Combining (4.38) with the inequalities (4.10) (for $\alpha = 0$) and (4.11) yields

$$\|p_t^k(x) - p_t(x)\|_p \leq \frac{\bar{c}_t k^{d_x} \|\phi\|_\infty}{\sqrt{N(k)}} + \frac{C_{2,t}}{k^2},$$

where $\bar{c}_t$ is constant w.r.t. to k (and $N(k)$). From assumption a.3, $N(k) \geq k^{2(d_x+2)}$, we arrive at

$$\|p_t^k(x) - p_t(x)\|_p \leq \frac{\bar{C}_{2,t}}{k^2}, \tag{4.39}$$

where $\bar{C}_{2,t} = \bar{c}_t \|\phi\|_\infty + C_{2,t} < \infty$ is a constant.

Similarly to the proof of Theorem 4.4, we choose $\mathcal{K} = [-M, +M] \times \cdots \times [-M, +M]$ without loss of generality. Using the inequality (4.39) with $p = 2$, we readily obtain

$$\int_{\mathcal{K}} E\left[\left(p_t^k(x) - p_t(x)\right)^2\right] dx \leq \frac{\bar{C}_{2,t}^2}{k^4} \mathcal{L}(\mathcal{K}) \leq \frac{C_{\mathcal{K},t}}{k^4},$$

where $\mathcal{L}(\mathcal{K}) = (2M)^{d_x}$ is the Lebesgue measure of $\mathcal{K}$ and $C_{\mathcal{K},t} = (2M)^{d_x} \bar{C}_{2,t}^2$ is constant w.r.t. k.

□

Note that the improvement of the convergence rate in Thorem 4.6 (k^{-4} versus k^{-2} in Theorem 4.4) is obtained at the expense of slightly increasing the computational cost of the particle filter ($N(k) \geq k^{2(d_x+2)}$ are needed, versus $N(k) \geq k^{2(d_x+1)}$ in Theorem 4.4 for $\alpha = \mathbf{0}$).

4.4. Convergence with the number of particles N

The results stated in this section are given in terms of the index k because this leads to concise expressions for the upper bounds of the approximation errors and it also yields a straightforward connection with classical kernel density estimation results in terms of the kernel bandwidth (recall that $h = 1/k$), as explicitly exploited in Section 4.3.

However, for the use of numerical schemes it may be useful to re-state, or at least interpret, some of these results in terms of the number particles, N, in the particle filter, since it is this parameter that determines the computational complexity of the algorithm. Fortunately, there is a straightforward (and deterministic) connection between the values of N and k, as already discussed in Section 3.3. Here, we elaborate on this issue and provide versions of Theorems 4.2 (uniform convergence over the state space), 4.3 (convergence in total variation distance of the continuous particle approximation of

π_t) and 4.5 (convergence of the ISE) with rates given in terms of N. They are given as corollaries, as their proofs are straightforward from the original theorems.

Under the standard conditions in Remark 4.5, the number of particles N and the inverse bandwidth k satisfy the inequality $N \geq k^{2(d_x+1)}$, and they are both integer quantities. Therefore, given N, the largest inverse bandwidth that we can choose is

$$k(N) = \lfloor N^{\frac{1}{2(d_x+1)}} \rfloor, \tag{4.40}$$

where $\lfloor z \rfloor = \sup\{m \in \mathbb{Z} : m \leq z\}$. It is apparent that $\lim_{N\to\infty} k(N) = \infty$. For conciseness in the notation, let us write

$$\hat{p}_t^N(x) = p_t^{k(N)}(x) = (\phi^x_{k(N)}, \pi_t^N)$$

for the kernel approximation of p_t with N particles determined by the map (4.40). Similarly, consider the sequence of hypercubes

$$\hat{\mathcal{K}}_N = [-\hat{M}_N, +\hat{M}_N] \times \ldots \times [-\hat{M}_N, +\hat{M}_N],$$

where $\hat{M}_N = \frac{1}{2}k(N)^{\frac{\beta}{d_x p}}$, with positive constants $p \geq 2$ and $0 \leq \beta < 1$. This is the counterpart of the sequence $\mathcal{K}_k$ in Section 4.1. Then, the next result follows readily from Theorem 4.2.

Corollary 4.1. *If the standard conditions are satisfied, then*

$$\sup_{x\in\hat{\mathcal{K}}_N} \left|\hat{p}_t^N(x) - p_t(x)\right| \leq \frac{U^\varepsilon}{k(N)^{1-\varepsilon}},$$

where $k(N) = \lfloor N^{\frac{1}{2(d_x+1)}} \rfloor$, $U^\varepsilon \geq 0$ *is an a.s. finite random variable and* $0 < \varepsilon < 1$ *is a constant, both of them independent of* N *and* x. *In particular,*

$$\lim_{N\to\infty} \sup_{x\in\hat{\mathcal{K}}_N} \left|\hat{p}_t^N(x) - p_t(x)\right| = 0 \quad a.s.$$

If we write $\breve{\pi}_t^N(dx) = \hat{p}_t^N(x)dx$ for the continuous approximation of $\pi_t(dx)$ constructed from the approximate density for a given number of particles N, then we have the corollary below, that follows immediately from Theorem 4.3.

Corollary 4.2. *If the standard conditions are satisfied and* $\pi_t(\hat{\mathcal{K}}_N^{\mathsf{c}}) \leq \frac{b}{2}k(N)^{-\gamma}$, *where* $k(N) = \lfloor N^{\frac{1}{2(d_x+1)}} \rfloor$ *and* $b > 0$ *and* $\gamma > 0$ *are constants independent of* N, *then*

$$\int \left|\hat{p}_t^N(x) - p_t(x)\right| dx < \frac{Q^\varepsilon}{k(N)^{\min\{1-\varepsilon,\gamma\}}},$$

where Q^ε *is an a.s. finite random variable and* $0 < \varepsilon < 1$ *is a constant, both of them independent of* N. *In particular,*

$$\lim_{N\to\infty} \int \left|\hat{p}_t^N(x) - p_t(x)\right| dx = 0 \quad a.s.$$

and, as a consequence,

$$\lim_{N\to\infty} d_{TV}(\breve{\pi}_t^N, \pi_t) = 0 \quad a.s.$$

We can also give a version of Theorem 4.5 with the error bound explicitly given in terms of the number of particles, N. To write it, let $\hat{p}_t^{\top,N}(x) = p_t^{\top,k(N)}(x) = I_{\hat{\mathcal{K}}_N}(x)\hat{p}_t^N(x)$ be the truncation of the approximate density within the compact hypercube $\hat{\mathcal{K}}_N$. Then we have the corollary below, which is proved in a trivial way from Theorem 4.5.

Corollary 4.3. *If the standard conditions are satisfied, $p_t \in B(\mathbb{R}^{d_x})$ and $\pi_t(\hat{\mathcal{K}}_N^{\mathsf{c}}) \leq bk(N)^{-\gamma}$, where $k(N) = \lfloor N^{\frac{1}{2(d_x+1)}} \rfloor$ and $b > 0$ and $\gamma > 0$ are constants independent of N, then*

$$ISE \equiv \int \left(\hat{p}_t^{\top,N}(x) - p_t(x)\right)^2 dx \leq \frac{U^\varepsilon}{k(N)^{\min\{2-\varepsilon,\gamma\}}},$$

where $U^\varepsilon \geq 0$ is an a.s. finite random variable, independent of N, and $0 < \varepsilon < 2$ is an arbitrarily small constant. In particular,

$$\lim_{N\to\infty} \int \left(\hat{p}_t^{\top,N}(x) - p_t(x)\right)^2 dx = 0 \quad a.s.$$

4.5. A simple example

There are several possible choices for the kernel function $\phi(x)$ that comply with assumptions A.1 and A.2. In particular, the standard multivariate Gaussian density with unit covariance,

$$\phi_G(x) = \frac{1}{(2\pi)^{\frac{d_x}{2}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{d_x} x_j^2\right\},$$

the d_x-dimensional Laplacian pdf,

$$\phi_L(x) = \left(\frac{1}{2b}\right)^{d_x} \exp\left\{-\frac{1}{b}\sum_{j=1}^{d_x} |x_j|\right\},$$

where $b = \sqrt{\frac{1}{2d_x}}$, and the Epanechnikov kernel $\phi_E(x)$ of Eq. (3.4) are densities with bounded second order moment.

It is also straightforward to check assumption A.4 for $\alpha = \mathbf{0}$ and $\alpha = \mathbf{1}$. In particular, for $\alpha = \mathbf{0}$, it is apparent that $\phi_G, \phi_L, \phi_E \in C_b(\mathbb{R}^{d_x})$. For $\alpha = \mathbf{1}$, the partial derivatives of the Gaussian and Laplacian kernels yield

$$\begin{aligned} D^{\mathbf{1}}\phi_G(x) &= \frac{(-1)^{d_x}}{(2\pi)^{\frac{d_x}{2}}} \prod_{l=1}^{d_x} x_l \exp\left\{-\frac{1}{2}\sum_{j=1}^{d_x} x_j^2\right\} \quad \text{and} \\ D^{\mathbf{1}}\phi_L(x) &= \frac{(-1)^{n^+}}{2^{d_x} b^{2d_x}} \exp\left\{-\frac{1}{b}\sum_{j=1}^{d_x} |x_j|\right\}, \quad x \neq 0, \end{aligned}$$

respectively, where $n^+ = |\{l \in \{1, \ldots, d_x\} : x_l > 0\}|$ is the number of positive elements of $x \in \mathbb{R}^{d_x}$. It is not hard to verify that $D^{\mathbf{1}}\phi_G \in C_b(\mathbb{R}^{d_x})$, while $D^{\mathbf{1}}\phi_L \in B(\mathbb{R}^{d_x})$. As for the Epanechnikov kernel, it is easy to show that $D^{\mathbf{1}}\phi_E(x) = 0 \; \forall x \in \mathbb{R}^{d_x}$.

In the sequel, we consider a simple example consisting in the approximation of a Gaussian filtering density using the Epanechnikov kernel.

***Example* 4.1.** Consider the state-space system

$$p_0(x_0) = N(x_0; 0, \mathcal{I}_2), \quad X_t = AX_{t-1} + U_t, \quad Y_t = BX_t + V_t, \quad t = 1, 2, ... \tag{4.41}$$

where $N(x_0; 0, \mathcal{I}_2)$ is the bivariate Gaussian pdf with mean 0 and 2×2 identity covariance matrix, $\mathcal{I}_2$; the matrices $A, B \in \mathbb{R}^{2\times 2}$ are

$$A = \begin{bmatrix} 0.50 & -0.35 \\ 0.39 & -0.45 \end{bmatrix}, \quad B = \begin{bmatrix} 0.50 & 0.30 \\ -0.80 & 0.20 \end{bmatrix},$$

and U_t, V_t, $t = 1, 2, ...$, are sequences of independent and identically distributed 2×1 Gaussian vectors with zero mean and covariance $\mathcal{I}_2$. For this class of (linear and Gaussian) models the filtering pdf p_t, $t \geq 1$, can be computed exactly using the Kalman filter [33] and, therefore, we have a reference for comparison with the approximations p_t^k produced by the particle filter with $N(k) = k^{2(d_x+1)} = k^6$ samples.

For the simulation, we generated two sequences, $x_0, x_1, \ldots, x_T$ and $y_1, \ldots, y_T$ for $T = 50$, according to the model (4.41). Then, using the fixed data $y_{1:T}$, we run a Kalman filter to compute the Gaussian pdf $p_T(x) = N(x; \bar{x}_T, \Sigma_T)$ exactly, where $\bar{x}_T$ and Σ_T are the posterior mean and covariance at time T, respectively. For the same sequence $y_{1:T}$, we run independent particle filters with various values of k and $N(k) = k^6$ particles each.

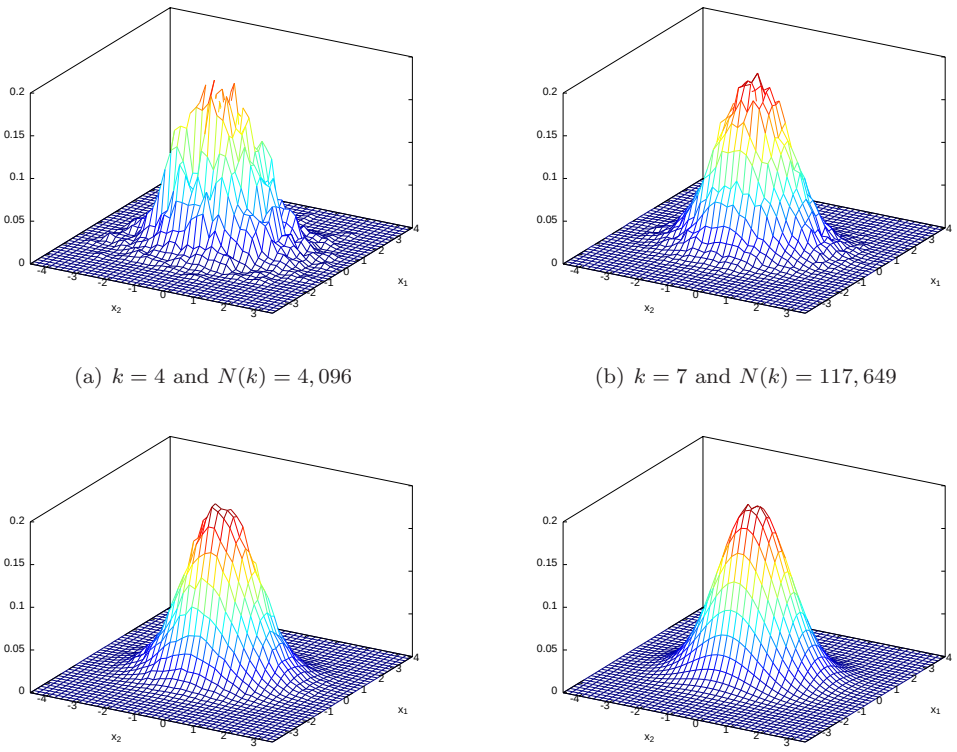
Figure 1 shows plots of the approximations $p_T^k(x)$ for $k = 4, 7, 10$ (constructed using the Epanechnikov kernel, ϕ_E) and the true pdf $p_T(x)$. The plots are drawn from a regular grid of points in $\mathbb{R}^2$, namely

$$x \in G_T = \{(x_1, x_2) : x_1 = -2.92 + 0.2n, \quad x_2 = -3.54 + 0.2n, \quad 1 \leq n \leq 42\} \tag{4.42}$$

(the offsets -2.92 and -3.54 correspond, approximately, to the true posterior mean of X_t). We can see that there is an obvious error for small k, while for $k = 10$ the difference between $p_T(x)$ and its approximation is negligible.

5. Applications

We illustrate the use of the convergence results in Section 3 by addressing two application problems: the computation of maximum a posteriori (MAP) estimators and the approximation of functionals of the filtering density, p_t. All through this section, we implicitly assume that the standard conditions of Remark 4.5 are satisfied.



(a) $k = 4$ and $N(k) = 4,096$

(b) $k = 7$ and $N(k) = 117,649$

(c) $k = 10$ and $N(k) = 1,000,000$

(d) $p_T(x)$ (exact)

Figure 1: Plots (a)-(c) display the approximations of the filtering density produced by the particle filter, $p_T^k(x)$, with an increasing number of particles $N(k) = k^6$, and an Epanechnikov kernel, ϕ_E. The true pdf, $p_T(x)$, is shown in Plot (d) for comparison. The plots correspond to the discrete grid G_T in Eq. (4.42).

5.1. MAP estimation

We tackle the problem of approximating the maximum a posteriori (MAP) estimator of the r.v. X_t. In particular, we address the numerical search of elements of the set

$$S_t = \arg\max_{x\in\mathbb{R}^{d_x}} p_t(x), \tag{5.1}$$

where $s \in S_t$ if, and only if, $p_t(s) = \max_{x\in\mathbb{R}^{d_x}} p_t(x)$. Note that this is a relevant problem since MAP estimates are often used, e.g., in signal processing and engineering applications (see, e.g., [25, 38, 24]), and the density $p_t(x)$ cannot be analytically found in general.

Let

$$S_t^k = \arg\max_{x\in\mathbb{R}^{d_x}} p_t^k(x) \tag{5.2}$$

be the set of MAP estimates for the approximation density $p_t^k(x)$ and note that $\hat{x}_k \in S_t^k$ if, and only if, $p_t^k(\hat{x}_k) = \max_{x\in\mathbb{R}^{d_x}} p_t^k(x)$. We can build a sequence of approximate estimates, denoted $\{\hat{x}_k\}_{k\geq 1}$, by taking one element from each set S_t^k, $k = 1, 2, ...$, at time t. If S_t is nonempty, then any convergent subsequence of $\{\hat{x}_k\}_{k\geq 1}$ yields an arbitrarily accurate approximation of a true MAP estimator $s \in S_t$, as stated below.

Theorem 5.1. *Assume that $S_t \neq \emptyset$ and take any convergent subsequence of $\{\hat{x}_k\}_{k\geq 1}$, denoted $\{\hat{x}_{k_i}\}_{i\geq 1}$. Let $\hat{x} = \lim_{i\to\infty} \hat{x}_{k_i}$ be the limit of such subsequence. If $p_t \in C_b(\mathbb{R}^{d_x})$, then $p_t(\hat{x}) = \max_{x\in\mathbb{R}^{d_x}} p_t(x)$. In particular, if $p_t(x)$ has a unique maximum, then S_t is a singleton and $\lim_{i\to\infty} \hat{x}_{k_i} = \arg\max_{x\in\mathbb{R}^{d_x}} p_t(x)$.*

Proof: We prove the theorem by contradiction. Specifically, assume that $p_t(\hat{x}) < \max_{x\in\mathbb{R}^{d_x}} p_t(x)$. Then, choose some $s \in S_t$, so that $p_t(s) = \max_{x\in\mathbb{R}^{d_x}} p_t(x)$ and $p_t(\hat{x}) < p_t(s)$, and let

$$\epsilon \triangleq \frac{p_t(s) - p_t(\hat{x})}{3} > 0. \tag{5.3}$$

Now, choose a compact subset $\mathcal{K} \subset \mathbb{R}^{d_x}$ that contains s, $\{\hat{x}_{k_i}\}_{i\geq 1}$ and $\hat{x}$. From Remark 4.7, $\lim_{k\to\infty} \sup_{x\in\mathcal{K}} |p_t^k(x) - p_t(x)| = 0$ a.s., hence there exists m such that for all $k \geq m$

$$\sup_{x\in\mathcal{K}} \left|p_t^k(x) - p_t(x)\right| < \epsilon. \tag{5.4}$$

Moreover, since $p_t(x)$ is continuous at every point $x \in \mathcal{K}$, we can choose an integer i_0 such that for all $i \geq i_0$ we obtain

$$|p_t(\hat{x}_{k_i}) - p_t(\hat{x})| < \epsilon. \tag{5.5}$$

Now, choose an index ℓ such that $\ell \geq i_0$ and $\ell \geq m$. Then, for every $i, k_i > \ell$, we have

$$\begin{aligned} p_t^{k_i}(\hat{x}_{k_i}) - p_t^{k_i}(s) &= \overbrace{p_t^{k_i}(\hat{x}_{k_i}) - p_t(\hat{x}_{k_i})}^{<\epsilon} + \overbrace{p_t(\hat{x}_{k_i}) - p_t(\hat{x})}^{<\epsilon} \\ &\quad + \overbrace{p_t(\hat{x}) - p_t(s)}^{=-3\epsilon} + \overbrace{p_t(s) - p_t^{k_i}(s)}^{<\epsilon} < 0, \end{aligned} \tag{5.6}$$

where the first term on the right-hand side, $p_t^{k_i}(\hat{x}_{k_i}) - p_t(\hat{x}_{k_i}) < \epsilon$, follows from inequality (5.4), the second term, $p_t(\hat{x}_{k_i}) - p_t(\hat{x}) < \epsilon$, follows from inequality (5.5), the third term, $p_t(\hat{x}) - p_t(s) = -3\epsilon$, is due to the definition in (5.3) and for the fourth term, $p_t(s) - p_t^{k_i}(s) < \epsilon$, is obtained from the inequality (5.4). Therefore, $\hat{x}_{k_i} \notin \arg\max_{x\in\mathbb{R}^{d_x}} p_t^{k_i}(x)$ and we arrive at a contradiction. Hence, $p_t(\hat{x}) = \max_{x\in\mathbb{R}^{d_x}} p_t(x)$.
□

***Remark* 5.1.** Note that the whole sequence $\{\hat{x}_k\}$ may not converge to a MAP estimate since it may, e.g., alternate between different elements of S_t.

Many global optimization algorithms, such as simulated annealing [6, 29] or accelerated random search [2], rely only on the evaluation of the objective function and Theorem 5.1 justifies their use with the approximation $p_t^k(x)$. Many other optimization procedures are based on the evaluation of derivates of the objective function. For example, we may want to use a gradient search method to find a local maximum of $p_t(x)$, i.e., to find a solution of the equation

$$\nabla_x p_t(x) = 0, \tag{5.7}$$

where $x = (x_1, \ldots, x_{d_x})$ and

$$\nabla_x p_t(x) = \begin{bmatrix} \frac{\partial p_t}{\partial x_1} \\ \vdots \\ \frac{\partial p_t}{\partial x_{d_x}} \end{bmatrix}(x) = \begin{bmatrix} D^{\alpha_1} p_t \\ \vdots \\ D^{\alpha_{d_x}} p_t \end{bmatrix}(x),$$

with $\alpha_i = (0, \ldots, \overbrace{1}^{i-\text{th}}, \ldots, 0)$. Let x^* be a solution of (5.7), i.e., $\nabla_x p_t(x^*) = 0$. Under the assumptions of Theorem 4.1, for every $\epsilon > 0$ there exists k_ϵ such that, $\forall k > k_\epsilon$,

$$-\epsilon < D^{\alpha_i} p_t^k(x^*) < \epsilon \quad \text{a.s..}$$

Therefore,

$$\|\nabla_x p_t^k(x^*)\| = \sqrt{\sum_{i=1}^{d_x} \left(D^{\alpha_i} {p_t}^k(x^*)\right)^2} < \epsilon\sqrt{d_x}, \quad \forall k < k_\epsilon,$$

and, since ϵ can be chosen as small as we wish,

$$\lim_{k\to\infty} \|\nabla_x p_t^k(x^*)\| = 0 \quad \text{a.s.},$$

which justifies the application of a gradient search procedure using the approximation of the filtering pdf.

***Example* 5.1.** We illustrate the application of a gradient search procedure using the same example as in Section 4.5. In particular, we consider the approximation of the maximum of the Gaussian filtering pdf $p_T(x)$, $T = 50$, using a steepest descent method.

Given an approximation $p_T^k(x)$ of the filtering density constructed with the Gaussian kernel ϕ_G, we run the iterative algorithm

$$\hat{x}_T(i+1)^k = \hat{x}_T(i)^k + a\, \nabla_x p_T^k(x)\big|_{x=\hat{x}_T(i)^k}, \quad i = 0, 1, 2, \ldots \tag{5.8}$$

with initial condition $\hat{x}_T(0)^k = (-2, -2)^\top$ and step-size parameter $a = 0.1$. This procedure yields a sequence of approximations $\hat{x}_T(1)^k, \ldots, \hat{x}_T(i)^k, \ldots$ of the MAP estimator $\hat{x}_T$. Since for the model of Eq. (4.41) it is possible to obtain $p_T(x)$ exactly, we have also run a steepest descent search over the true filtering pdf, namely,

$$\hat{x}_T(i+1) = \hat{x}_T(i) + a\, \nabla_x p_T(x)|_{x=\hat{x}_T(i)}, \quad i = 0, 1, 2, \ldots, \tag{5.9}$$

that generates the estimates $\hat{x}_T(1), \ldots, \hat{x}_T(i), \ldots$ for the same initial condition and step size.

The results, using the same sequence of observations as in Section 4.5, are shown in Figure 2. Specifically, Figures 2(a) and 2(b) show the trajectories described by the estimates $\hat{x}_T(1)^k, \ldots, \hat{x}_T(i)^k, \ldots$ superimposed over the contour plots of the approximate pdf $p_T^k(x)$ for $k = 5$ and $k = 9$, respectively (and $N(k) = k^6$). For comparison, Figure 2(c) depicts the sequence $\hat{x}_T(1), \ldots, \hat{x}_T(i), \ldots$ obtained from the search over the true density $p_T(x)$, together with the corresponding contour plot. We observe that both the pdf's and the trajectories described by the search algorithms are very similar.

In practice, problem (5.2) may turn out difficult to solve because the approximation $p_t^k(x)$ can be rough, plagued with local maxima, when the number of particles $N(k)$ is not sufficiently large (see, e.g., Figure 1(a)). In such cases, one may have to resort to computationally expensive global optimization methods instead of (simpler) gradient-based techniques. A computationally less demanding approach consists in performing the search of the maximum of $p_t^k(x)$ over the discrete set of particles $\Omega_t^{N(k)} = \{x_t^{(n)}\}_{n=1,\ldots,N(k)}$ (where $N(k) \geq k^{2(2d_x+1)}$) instead of over the complete (continuous) state space[4] [1]. To be specific, it is straightforward (e.g., by a linear search) to obtain the set of particle values for which the approximate density is maximum, namely

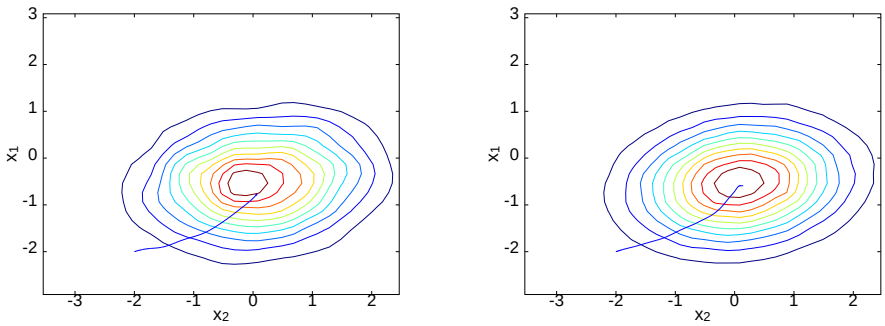
$$\tilde{S}_t^k = \arg \max_{x \in \Omega_t^{N(k)}} p_t^k(x). \tag{5.10}$$

In the classical setup, when the target density is approximated using i.i.d. samples drawn directly from the desired distribution, it can be shown that the elements of $\tilde{S}_t^k$ become arbitrarily close to the elements of S_t^k as $k \to \infty$ (and, hence, as $N(k) \to \infty$) [1]. The following theorem yields a similar asymptotic result when $\Omega_t^{N(k)}$ is generated by the standard particle filter.

Theorem 5.2. *Assume that $S_t \neq \emptyset$ and $p_t \in C_b(\mathbb{R}^{d_x})$. If $s_t \in S_t$, $s_t^k \in S_t^k$ and $\tilde{s}_t^k \in \tilde{S}_t^k$, then,*

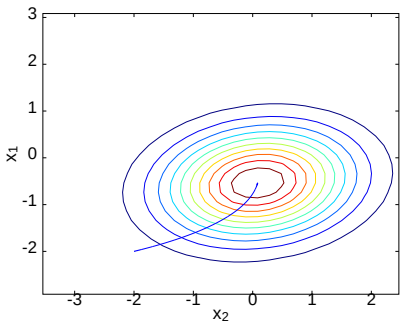
$$\lim_{k\to\infty} p_t(\tilde{s}_t^k) = \lim_{k\to\infty} p_t(s_t^k) = p_t(s_t) \quad a.s. \tag{5.11}$$

[4]This alternative approximation of the MAP estimator of X_t was pointed out to us by one of the anonymous reviewers of the original manuscript.



(a) $k = 5$, $N(k) = 15,625$.

(b) $k = 9$, $N(k) = 531,441$.



(c) True pdf p_t.

Figure 2: Trajectories of the gradient search algorithms. Plot (a) shows the estimates produced by the gradient search algorithm of Eq. (5.8) superimposed over a contour representation of $p_T^k(x)$ for $k = 5$. Plot (b) displays the estimates and contour graph for $p_T^k(x)$ for $k = 9$. Plot (c) shows the estimates produced by the gradient search algorithm of Eq. (5.9) superimposed over a contour representation of $p_T(x)$, for comparison.

Proof: Let us introduce the additional approximation of the MAP estimator

$$\check{S}_t^k = \arg \max_{x \in \Omega_t^{N(k)}} p_t(x).$$

The set $\check{S}_t^k$ cannot be computed in practice because $p_t(x)$ cannot be evaluated, but it will be auxiliary in proving that Eq. (5.11) holds. Specifically, we first show (using an argument taken from [39]) that the sequence $\{p_t(\check{s}_t^k) : \check{s}_t^k \in \check{S}_t^k, k \geq 1\}$ converges to $p_t(s_t)$ a.s. when $k \to \infty$. Then, we use the latter result to show that (5.11) holds.

We proceed to prove that $\lim_{k\to\infty} p_t(\check{s}_t^k) = p_t(s_t)$ a.s. Choose any MAP estimate $s_t \in S_t$ and define the open ball

$$B_m(s_t) = \left\{ x \in \mathbb{R}^{d_x} : \|x - s_t\| < \frac{1}{m} \right\},$$

where m is a positive integer. From Proposition 2.1, the integral $(I_{B_m(s_t)}, \pi_t)$ (where $I_A(x) = 1$ if $x \in A$ and 0 otherwise) can be approximated with asymptotically vanishing error. Specifically, since $k \to \infty$ implies that $N(k) \to \infty$, we have

$$\lim_{k\to\infty} (I_{B_m(s_t)}, \pi_t^{N(k)}) = \lim_{k\to\infty} \frac{|B_m(s_t) \cap \Omega_t^{N(k)}|}{N(k)} = (I_{B_m(s_t)}, \pi_t) \quad \text{a.s.},$$

where $|B_m(s_t) \cap \Omega_t^{N(k)}|$ yields the number of particles inside the ball $B_m(s_t)$. Since p_t is continuous and positive at s_t, then $(I_{B_m(s_t)}, \pi_t) > 0$, hence

$$\lim_{k\to\infty} \frac{|B_m(s_t) \cap \Omega_t^{N(k)}|}{N(k)} > 0 \quad \text{a.s.} \tag{5.12}$$

for any integer m.

The inequality (5.12) means that the set $B_m(s_t) \cap \Omega_t^{N(k)}$, consisting of particles which are "close" to s_t (namely, at a distance lesser than $1/m$), is asymptotically non-empty, with probability 1, no matter how large we choose m. Therefore, let us choose a point $s_t^{k,m} \in B_m(s_t) \cap \Omega_t^{N(k)}$. Obviously, $p_t(s_t^{k,m}) \leq p_t(s_t)$, but also $p_t(s_t^{k,m}) \leq p_t(\check{s}_t^k)$ by construction, hence

$$p_t(s_t^{k,m}) \leq p_t(\check{s}_t^k) \leq p_t(s_t). \tag{5.13}$$

Since p_t is continuous at s_t, for any arbitrarily small $\epsilon > 0$ we can choose $m > 0$ such that if $x \in B_m(s_t)$ then $p_t(s_t) - p_t(x) < \epsilon$. However, for every m there exists k_m such that when $k > k_m$ the intersection $B_m(s_t) \cap \Omega_t^{N(k)}$ is a.s. non-empty, hence there exists a particle $s_t^{k,m} \in B_m(s_t) \cap \Omega_t^{N(k)}$ and the inequality (5.13) yields $0 \leq p_t(s_t) - p_t(\check{s}_t^k) \leq p_t(s_t) - p_t(s_t^{k,m}) < \epsilon$. Therefore,

$$\lim_{k\to\infty} p_t(\check{s}_t^k) = p_t(s_t) \quad \text{a.s.} \tag{5.14}$$

Now we prove the convergence of $p_t(\tilde{s}_t^k)$ and $p_t(s_t^k)$ toward $p_t(s_t) = \max_{x\in\mathbb{R}^{d_x}} p_t(x)$. Consider first the non-negative difference

$$0 \leq p_t(s_t) - p_t(\tilde{s}_t^k) = \left(p_t(s_t) - p_t(\check{s}_t^k)\right) + \left(p_t(\check{s}_t^k) - p_t^k(\check{s}_t^k)\right) + \left(p_t^k(\check{s}_t^k) - p_t^k(\tilde{s}_t^k)\right) + \left(p_t^k(\tilde{s}_t^k) - p_t(\tilde{s}_t^k)\right) \tag{5.15}$$

where the inequality follows from the definition of S_t, and let us look into each term on the right hand side of (5.15) separately.

Choose any arbitrarily small $\epsilon > 0$. From (5.14), there exists k_1 such that for every $k > k_1$,

$$0 \leq p_t(s_t) - p_t(\check{s}_t^k) < \frac{\epsilon}{6}. \tag{5.16}$$

Let us now select, without loss of generality, a compact set $\mathcal{K} \supset S_t \cup S_t^k \cup \tilde{S}_t^k \cup \check{S}_t^k$. From Remark 4.7,

$$|p_t(\check{s}_t^k) - p_t^k(\check{s}_t^k)| \leq \sup_{x \in \mathcal{K}} |p_t(x) - p_t^k(x)| \leq \frac{\tilde{U}^\varepsilon}{k^{1-\varepsilon}},$$

where $\tilde{U}^\varepsilon$ is an a.s. finite random variable and $0 < \varepsilon < 1$ is arbitrary but constant. Hence, there exists k_2 such that, for every $k > k_2$,

$$-\frac{\epsilon}{6} < p_t(\check{s}_t^k) - p_t^k(\check{s}_t^k) < \frac{\epsilon}{6}. \tag{5.17}$$

By the same argument, there is some k_3 such that, for every $k > k_3$,

$$-\frac{\epsilon}{6} < p_t(\tilde{s}_t^k) - p_t^k(\tilde{s}_t^k) < \frac{\epsilon}{6}. \tag{5.18}$$

Since, by construction,

$$p_t^k(\check{s}_t^k) - p_t^k(\tilde{s}_t^k) \leq 0, \tag{5.19}$$

substituting (5.16)–(5.19) into the inequality (5.15) and solving for $p_t^k(\check{s}_t^k) - p_t^k(\tilde{s}_t^k)$ yields

$$0 \geq p_t^k(\check{s}_t^k) - p_t^k(\tilde{s}_t^k) > -\frac{\epsilon}{2}, \tag{5.20}$$

for every $k > \max\{k_1, k_2, k_3\}$. However,

$$\left|p_t(s_t) - p_t(\tilde{s}_t^k)\right| \leq \left|p_t(s_t) - p_t(\check{s}_t^k)\right| + \left|p_t(\check{s}_t^k) - p_t^k(\check{s}_t^k)\right| + \left|p_t^k(\check{s}_t^k) - p_t^k(\tilde{s}_t^k)\right| + \left|p_t^k(\tilde{s}_t^k) - p_t(\tilde{s}_t^k)\right| \tag{5.21}$$

and substituting (5.16)–(5.18) and (5.20) into (5.21) yields $|p_t(s_t) - p_t(\tilde{s}_t^k)| < \epsilon$ a.s. for every $k > \max\{k_1, k_2, k_3\}$, hence

$$\lim_{k \to \infty} p_t(\tilde{s}_t^k) = p_t(s_t) \quad \text{a.s.}$$

A similar argument proves the convergence of $p_t(s_t^k) \to p_t(s_t)$. In particular, if we choose a compact set $\mathcal{K} \supset S_t \cup S_t^k$ we can again apply Remark 4.7 to show that, for any $\epsilon > 0$ there exists k_4 such that, for every $k > k_4$,

$$-\frac{\epsilon}{4} < p_t(s_t) - p_t^k(s_t) < \frac{\epsilon}{4} \tag{5.22}$$

and there exists k_5 such that, for every $k > k_5$,

$$-\frac{\epsilon}{4} < p_t^k(s_t^k) - p_t(s_t^k) < \frac{\epsilon}{4}. \tag{5.23}$$

	$p_T(s_T)$	$p_T(s_T) - p_T(s_T^k)$	$p_T(s_T) - p_T(\tilde{s}_T^k)$
$k = 5$	0.201937	0.005090	0.004500
$k = 9$	0.201937	0.001030	0.002679

Table 1. Approximation of the maximum posterior density $p_T(s_T) = \max_{x \in \mathbb{R}^{d_x}} p_T(x)$ by way of Eqs. (5.2) and (5.10) ($p_T(s_T^k)$ and $p_T(\tilde{s}_T^k)$, respectively). The approximate MAP estimate s_T^k has been computed via the gradient search method of Eq. (5.8).

However,

$$0 \leq p_t(s_t) - p_t(s_t^k) = \left(p_t(s_t) - p_t^k(s_t)\right) + \left(p_t^k(s_t) - p_t^k(s_t^k)\right) + \left(p_t^k(s_t^k) - p_t(s_t^k)\right) \quad (5.24)$$

and, since $p_t^k(s_t) - p_t^k(s_t^k) \leq 0$ by definition of S_t^k, substituting (5.22) and (5.23) into (5.24) and solving for $p_t^k(s_t) - p_t^k(s_t^k)$ yields

$$-\frac{\epsilon}{2} < p_t^k(s_t) - p_t^k(s_t^k) \leq 0 \quad (5.25)$$

for every $k > \max\{k_4, k_5\}$. Finally, since

$$\left|p_t(s_t) - p_t^k(s_t^k)\right| \leq \left|p_t(s_t) - p_t^k(s_t^k)\right| + \left|p_t^k(s_t) - p_t^k(s_t^k)\right| + \left|p_t^k(s_t^k) - p_t(s_t^k)\right|,$$

we obtain that $|p_t(s_t) - p_t^k(s_t^k)| \leq \epsilon$ for every $k > \max\{k_4, k_5\}$, hence

$$\lim_{k \to \infty} p_t^k(s_t^k) = p_t(s_t) \quad \text{a.s.}$$

□

***Example* 5.2.** We consider, again, the Gaussian density $p_T(x)$, with $T = 50$, of Examples 4.1 and 5.1 in order to compare numerically the approximations $p_T(s_t^k)$ and $p_T(\tilde{s}_T^k)$ with the true maximum $p_T(s_T)$. The results are displayed in Table 1, which shows the maximum $p_T(s_T) = \max_{x \in \mathbb{R}^{d_x}} p_T(x)$ and the differences $p_T(s_T) - p_T(\tilde{s}_t^k)$ and $p_T(s_T) - p_T(s_T^k)$ for $k = 5$ ($N = 15,625$) and $k = 9$ ($N = 531,441$).

5.2. Functionals of p_t

The result of Theorem 4.3 allows us to construct (rigorous) approximations of functionals of the form $(f \circ p_t, \pi_t)$, where $\circ$ denotes composition and f is a Lipschitz-continuous and bounded real function. In order to provide rates for the convergence of the particle-kernel approximations $(f \circ p_t^k, \pi_t^{N(k)})$, we again work with the sequence of hypercubes $\mathcal{K}_k = [-M_k, M_k] \times \cdots \times [-M_k, M_k] \subset \mathbb{R}^{d_x}$ where $M_k = \frac{1}{2} k^{\frac{\beta}{d_x p}}$ and $0 < \beta < 1$, $p > 3$ are constants with respect to k. Specifically, we have the following result.

Theorem 5.3. *Choose any bounded, Lipschitz continuous function f, i.e., $f \in B(\mathbb{R})$ and $\forall x, y \in \mathbb{R}$*

$$|f(x) - f(y)| \leq c_f |x - y|,$$

for some finite constant $c_f > 0$. If $p_t \in B(\mathbb{R}^{d_x})$ and $\pi_t(\mathcal{K}_k^{\mathsf{c}}) \leq \frac{b}{2}k^{-\gamma}$ for some constants $\gamma, b > 0$ then

$$\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t, \pi_t)\right| \leq \frac{Q_f^\varepsilon}{k^{\min\{1-\varepsilon,\gamma\}}} \tag{5.26}$$

where $0 < \varepsilon < 1$ is an arbitrarily small constant and Q_f^ε is an a.s. finite random variable independent of k. In particular,

$$\lim_{k\to\infty} \left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t, \pi_t)\right| = 0 \quad a.s.$$

Proof: Consider first the absolute difference

$$\begin{aligned} \left|(f \circ p_t^k, \pi_t) - (f \circ p_t, \pi_t)\right| &= \left|\int \left[(f \circ p_t^k)(x) - (f \circ p_t)(x)\right] p_t(x)dx\right| \\ &\leq \int \left|(f \circ p_t^k)(x) - (f \circ p_t)(x)\right| p_t(x)dx, \end{aligned} \tag{5.27}$$

where the inequality holds because $p_t(x) \geq 0$. Using the Lipschitz continuity of f in the integral of Eq. (5.27) yields

$$\begin{aligned} \left|(f \circ p_t^k, \pi_t) - (f \circ p_t, \pi_t)\right| &\leq c_f \int \left|p_t^k(x) - p_t(x)\right| p_t(x)dx \\ &\leq c_f \|p_t\|_\infty \int \left|p_t^k(x) - p_t(x)\right| dx, \end{aligned} \tag{5.28}$$

where the second inequality follows from the assumption $p_t \in B(\mathbb{R}^{d_x})$ (hence $\|p_t\|_\infty < \infty$). Eq. (5.28) together with Theorem 4.3 readily yields

$$\left|(f \circ p_t^k, \pi_t) - (f \circ p_t, \pi_t)\right| \leq \frac{c_f \|p_t\|_\infty Q^\varepsilon}{k^{\min\{1-\varepsilon,\gamma\}}} \tag{5.29}$$

where $0 < \varepsilon < 1$ is a constant and Q^ε is an a.s. finite random variable.

As a second step, consider the difference $\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t^k, \pi_t)\right|$. Since $f \in B(\mathbb{R})$, it follows that $\|f \circ p_t^k\|_\infty \leq \|f\|_\infty$ independently of k and an application of Proposition 2.1 yields

$$E\left[\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t^k, \pi_t)\right|^q\right] \leq \frac{c_t^q \|f\|_\infty^q}{N(k)^{\frac{q}{2}}} \leq \frac{c_t^q \|f\|_\infty^q}{k^{q(2d_x+1)}},$$

where $q \geq 1$ and the second inequality holds because $N(k) \geq k^{2(d_x+|\mathbf{1}|+1)}$. Using Lemma 4.1 with $c = c_t^q \|f\|_\infty^q$ and $\nu = 0$ (note that $q(2d_x + 1) \geq 2$ for any $q, d_x \geq 1$), we readily obtain the convergence rate for the absolute error, i.e.,

$$\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t^k, \pi_t)\right| \leq \frac{U^\varepsilon}{k^{1-\varepsilon}} \tag{5.30}$$

where $0 < \varepsilon < 1$ is an arbitrarily small constant and $U^\varepsilon \geq 0$ is an a.s. finite random variable.

To conclude, consider the triangle inequality

$$\begin{aligned}\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t, \pi_t)\right| &\leq \left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t^k, \pi_t)\right| \\ &\quad + \left|(f \circ p_t^k, \pi_t) - (f \circ p_t, \pi_t)\right|. \end{aligned} \tag{5.31}$$

Substituting (5.29) and (5.30) into (5.31) yields

$$\left|(f \circ p_t^k, \pi_t^{N(k)}) - (f \circ p_t, \pi_t)\right| \leq \frac{U^\varepsilon}{k^{1-\varepsilon}} + \frac{c_f \|p_t\|_\infty Q^\varepsilon}{k^{\min\{1-\varepsilon,\gamma\}}} \leq \frac{Q_f^\varepsilon}{k^{\min\{1-\varepsilon,\gamma\}}}, \tag{5.32}$$

where the random variable $Q_f^\varepsilon = U^\varepsilon + c_f \|p_t\|_\infty Q^\varepsilon \geq 0$ is a.s. finite and independent of k.

□

In statistical signal processing, machine learning and information theory it is often of interest to evaluate the Shannon entropy of a probability measure π [4, 32, 42]. Assuming that π has a density p w.r.t. the Lebesgue measure, the entropy of the probability distribution is

$$\mathcal{H}(\pi) = -(\log p, \pi) = -\int_{\mathcal{S}} p(x) \log [p(x)]\, dx,$$

where $\mathcal{S}$ is the support of p. In the case of the filtering measure π_t, it is natural to think of a particle approximation of the entropy $\mathcal{H}(\pi_t)$ constructed as

$$\mathcal{H}(\pi_t)^k = -(\log p_t^k, \pi_t^{N(k)}) = -\frac{1}{N(k)} \sum_{n=1}^{N(k)} \log p_t^k(x_t^{(n)}).$$

Unfortunately, the log function is neither bounded nor Lipschitz continuous and, therefore, Theorem 5.3 does not guarantee the convergence $\mathcal{H}(\pi_t)^k \to \mathcal{H}(\pi_t)$. Such a result, however, can be obtained, with a more specific argument, if we assume the support of the density p_t to be compact.

Theorem 5.4. *Let the sequence of observations $Y_{1:T} = y_{1:T}$ (for some large but finite T) be fixed and assume that $g_t^{y_t}$ is positive and bounded and $\log p_t \in \mathsf{F}_T^4$ for $1 \leq t \leq T$. If there exists a compact set $\mathcal{S} \subset \mathbb{R}^{d_x}$ such that $\int_{\mathcal{S}} p_t(x)dx = 1$ and $\inf_{x\in\mathcal{S}} p_t(x) > 0$, then*

$$\lim_{k\to\infty} \left|\mathcal{H}(\pi_t)^k - \mathcal{H}(\pi_t)\right| = 0 \quad a.s.$$

Proof: We apply the triangle inequality to obtain

$$\begin{aligned}\left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t)\right| &\leq \left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t^{N(k)})\right| \\ &\quad + \left|(-\log p_t, \pi_t^{N(k)}) - (-\log p_t, \pi_t)\right| \end{aligned} \tag{5.33}$$

and then analyze the two terms on the right-hand side of (5.33).

The first one can be expanded to yield

$$\begin{aligned}\left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t^{N(k)})\right| &= \left|\frac{1}{N(k)}\sum_{i=1}^{N(k)} \log \frac{p_t(x_t^{(i)})}{p_t^k(x_t^{(i)})}\right| \\ &\leq \frac{1}{N(k)}\sum_{i=1}^{N(k)} \left|\log \frac{p_t(x_t^{(i)})}{p_t^k(x_t^{(i)})}\right|. \end{aligned} \tag{5.34}$$

The logarithm of a ratio x/y can be upper bounded as

$$\log \frac{x}{y} \leq \frac{\max\{x,y\}}{\min\{x,y\}} - 1, \tag{5.35}$$

hence applying (5.35) into (5.34) we arrive at

$$\left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t^{N(k)})\right| \leq \frac{1}{N(k)}\sum_{i=1}^{N(k)} \left|\frac{\max\left\{p_t^k(x_t^{(i)}), p_t(x_t^{(i)})\right\}}{\min\left\{p_t^k(x_t^{(i)}), p_t(x_t^{(i)})\right\}} - 1\right|. \tag{5.36}$$

However, from Theorem 4.2 and Remark 4.7,

$$\lim_{k\to\infty} p_t^k(x)/p_t(x) = \lim_{N\to\infty} p_t(x)/p_t^k(x) = 1 \quad \text{a.s.}$$

for every $x \in \mathcal{S}$. Moreover, since we have assumed $\inf_{x\in\mathcal{S}} p_t(x) > 0$, it follows that for any $\epsilon > 0$ there exists k_ϵ independent of x such that, for all $k > k_\epsilon$,

$$\frac{\max\left\{p_t^k(x_t^{(i)}), p_t(x_t^{(i)})\right\}}{\min\left\{p_t^k(x_t^{(i)}), p_t(x_t^{(i)})\right\}} \leq 1 + \epsilon. \tag{5.37}$$

Substituting (5.37) into (5.36) yields, for all $k > k_\epsilon$,

$$\left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t^{N(k)})\right| \leq \epsilon \quad \text{a.s.}$$

Since ϵ can be as small as we wish,

$$\lim_{k\to\infty} \left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t^{N(k)})\right| = 0 \quad \text{a.s.} \tag{5.38}$$

The second term in (5.33) converges to 0 because of Proposition 2.1, part (b), i.e.,

$$\lim_{k\to\infty} \left|(-\log p_t, \pi_t^{N(k)}) - (-\log p_t, \pi_t)\right| = 0 \quad \text{a.s.} \tag{5.39}$$

and taking together Eqs. (5.38), (5.39) and (5.33) we arrive at

$$\lim_{k\to\infty} \left|(-\log p_t^k, \pi_t^{N(k)}) - (-\log p_t, \pi_t)\right| = 0 \quad \text{a.s.}$$

□

	$k=3$	$k=4$	$k=5$
mean	0.0616	0.0370	0.0128
std	0.0453	0.0249	0.0091

Table 2. Empirical mean and standard deviation of the entropy-approximation error, $\left|\mathcal{H}(\pi_T) - \mathcal{H}(\pi_T)^k\right|$, averaged over 30 independent simulations. The entropies are evaluated in nats.

***Example* 5.3.** We continue to use the model of Section 4.5 to numerically illustrate the particle approximation of $\mathcal{H}(\pi_t)$. Since the densities p_t for this example are Gaussian with known covariance matrices Σ_t, $t = 1, 2, ...$, we can compute their associated Shannon entropies exactly, namely

$$\mathcal{H}(\pi_t) = \frac{1}{2}\log\left((2\pi \mathrm{e})^{d_x}|\Sigma_t|\right),$$

where $|\Sigma_t|$ is the determinant of matrix Σ_t. Taking $t = T = 50$ and using the same sequence of observations $y_{1:50}$ as in Section 4.5, the resulting entropy is $\mathcal{H}(\pi_T) = 2.5998$ nats.

Let us point out that, obviously, the Gaussian distribution has an infinite support and, therefore, the convergence result of Theorem 5.4 cannot be rigorously applied. However, the Gaussian pdf is light-tailed and, as can be observed from Figure 1(d), it can be truncated within a compact (rectangular) support and still yield a faithful representation of the original distribution.

Table 2 displays the empirical mean and standard deviation of the absolute error $\left|\mathcal{H}(\pi_T) - \mathcal{H}(\pi_T)^k\right|$ obtained through computer simulations for $N(k) = k^6$ and $k = 3, 4, 5$. To be specific, we carried out 30 independent simulation runs for each value of k. We observe how both the mean error and its standard deviation reduce quickly as k is increased.

6. Summary

We have addressed the approximation of the sequence of filtering pdf's of a Markov state-space model using a particle filter. The numerical technique is conceptually simple. We collect the N particles generated by the sequential Monte Carlo algorithm and approximate the desired density as the sum of N scaled kernel functions located at the particle positions. The main contribution of the paper is the analysis of the convergence of such particle-kernel approximations. In particular, we have first proved the point-wise convergence of the approximation of the filtering density and its derivatives as the number of particles is increased and the kernel bandwidth is correspondingly decreased. Explicit convergence rates are provided and they are sufficient to prove that the approximation errors vanish a.s. Under mild additional assumptions on the chosen kernel, it is possible to extend the latter result to prove that the approximation error converges uniformly on the support of the filtering density (rather than point-wise) and a.s. to 0. We have also found an explicit convergence rate for the supremum of the approximation error. The analysis establishes a connection between the complexity of the particle filter and

the bandwidth of the kernel function used for estimating the filtering pdf. For a given number of particles N, this relationship yields an optimal value of the bandwidth.

The uniform approximation result has a number of applications. We have first exploited it to prove the convergence, in total variation distance, of the continuous measure generated by the estimated density toward the true filtering measure. In a similar vein, we have also shown that the MISE of the sequence of approximate densities converges (quadratically with the kernel bandwidth) toward 0 when the state space is compact. For a truncated version of the density approximation, the (random) ISE is also shown to converge a.s. toward 0 without assuming compactness of the support. Although the convergence rate found for the ISE is only quadratic (versus fourth order for the asymptotic approximation of the MISE in classical kernel density estimation theory), one should be aware that all the results obtained in this paper remain valid whenever the density estimator is obtained by smoothing a discrete random measure π_t^N that is "good enough" to estimate integrals of bounded functions in such a way that the L_p norms of the approximation error convergence as $\frac{c}{\sqrt{N}}$ (in particular, we do not require to have samples from the target density p_t). As a consequence, the results obtained here can be applied to, e.g., kernel density estimators built from importance samples as in [46], or to the analysis of bootstrapped estimators as considered in [28]. Convergence of the MISE with the fourth power of the bandwidth (i.e., the same as for the AMISE in the classical theory) can also be obtained at the expense of a slight increase in the computational load of the particle filter and some additional assumptions on the kernel function and the smoothness of the filter density.

We have also proved that the maxima of the approximate filtering density converge a.s. toward the true ones. Therefore, MAP estimation of the state at time t can be carried out using, e.g., gradient search methods on the approximate filtering pdf. We remark that it is sound to apply such methods on the approximate function, since we have proved convergence also for its derivatives. The last application we consider is the approximation of functionals of the filtering pdf. We provide a general result that guarantees the convergence of the particle-kernel approximations for general bounded and Lipschitz continuous functionals of the filtering density. Finally, we prove that it is also possible to use the proposed constructs to approximate the Shannon entropy of densities with a compact support. In order to arrive at this result, we have also proved the convergence of the particle filter approximations of integrals of unbounded test functions under very mild assumptions (essentially, the integrability of the function up to fourth order). This is a departure from most existing approaches, which assume bounded test functions.

Appendix A: Proof of Proposition 2.1

Part (a) of Proposition 2.1 is a straightforward consequence of [39, Lemma 1], hence we focus here on part (b). We start with the following Lemma, which is used as an auxiliary result in the proof of the Proposition.

Lemma A.1. *Let $\{\theta_n; n = 1, ..., N\}$ be a set of random variables, assumed centered and i.i.d. conditionally on some σ-algebra $\mathcal{G}$. If $E[\theta_n^4] < \infty$, $n = 1, ..., N$, then*

$$E\left[\left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^4\right] \leq \frac{cE\left[\theta_1^4\right]}{N^2}, \tag{A.1}$$

where c is a constant independent of N. In particular,

$$\lim_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^{N}\theta_n\right| = 0 \quad a.s.$$

Proof: Conditional on $\mathcal{G}$, the variables are zero mean and independent, hence it is straightforward to show that

$$E\left[\left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^4|\mathcal{G}\right] = \frac{1}{N^4}\sum_{n=1}^{N}E[\theta_n^4|\mathcal{G}] + \frac{6}{N^4}\sum_{1\leq j<k\leq N}E[\theta_j^2|\mathcal{G}]E[\theta_k^2|\mathcal{G}]. \tag{A.2}$$

Since the conditional (on $\mathcal{G}$) distributions of the θ_n's are identical, we can rewrite (A.2) in terms of $E[\theta_1^4|\mathcal{G}]$ and $E[\theta_1^2|\mathcal{G}]$ alone, namely

$$E\left[\left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^4|\mathcal{G}\right] = \frac{1}{N^3}E[\theta_1^4|\mathcal{G}] + \frac{6(N-1)}{N^3}E^2[\theta_1^2|\mathcal{G}]. \tag{A.3}$$

However, $E[\theta_n^4|\mathcal{G}] \geq E^2[\theta_n^2|\mathcal{G}]$ (from Jensen's inequality), which readily yields the bound

$$E\left[\left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^4|\mathcal{G}\right] \leq E[\theta_1^4|\mathcal{G}]\frac{1+6(N-1)}{N^3} \leq \frac{cE[\theta_1^4|\mathcal{G}]}{N^2}, \tag{A.4}$$

for any constant $c \geq 6$. Taking unconditional expectations on the right-hand and left-hand sides of (A.4) leads to the desired inequality (A.1).

Finally, a standard Borel-Cantelli argument yields $\lim_{N\to\infty} E\left[\left|\frac{1}{N}\sum_{n=1}^{N}\theta_n\right|\right] = 0$ a.s. □

Lemma A.1 enables us to prove the convergence of particle approximations, $\lim_{N\to\infty}(f, \pi_t^N) \to (f, \pi_t)$ a.s., when $f \in \mathsf{F}_T^4$. We follow an induction argument to prove the latter result. In particular, let

$$\bar{\pi}_t^N(dx) = \sum_{n=1}^{N} w_t^{(n)}\delta_{\bar{x}_t^{(n)}}(dx) \tag{A.5}$$

be the random measure resulting from assigning importance weights $w_t^{(n)} = \frac{g_t^{y_t}(\bar{x}_t^{(n)})}{\sum_{k=1}^{N} g_t^{y_t}(\bar{x}_t^{(k)})}$ to the particles $\bar{x}_t^{(n)}$. We prove that:

1. At time $t = 0$, $\lim_{N\to\infty} |(f, \pi_0^N) - (f, \pi_0)| = 0$ a.s. and, at time $t = 1$,
$$\lim_{N\to\infty} \left|(f, \bar{\pi}_1^N) - (f, \pi_1)\right| = 0 \quad \text{a.s.}$$
2. If $\lim_{N\to\infty} |(f, \bar{\pi}_t^N) - (f, \pi_t)| = 0$ a.s. at some time $1 \le t < T$, then
$$\lim_{N\to\infty} \left|(f, \bar{\pi}_{t+1}^N) - (f, \pi_{t+1})\right| = 0 \quad \text{a.s.}$$

In the induction step it is explicitly shown that $\lim_{N\to\infty} |(f, \bar{\pi}_t^N) - (f, \pi_t)| = 0$ a.s. implies $\lim_{N\to\infty} |(f, \pi_t^N) - (f, \pi_t)| = 0$. The latter is the result in the statement of Proposition 2.1, hence the argument above yields a complete proof.

A.1. Base case ($t \le 1$)

For $t = 0$, the samples $x_0^{(n)}$, $n = 1, ..., N$, are i.i.d. with common distribution π_0, hence the approximation π_0^N is constructed as $\pi_0^N(dx) = \frac{1}{N}\sum_{n=1}^N \delta_{x_0^{(n)}}(dx)$. Consider the random variables $\theta_{n,0} = f(x_0^{(n)}) - (f, \pi_0)$. It is apparent that they are i.i.d. and $E[\theta_{n,0}] = 0$. Also

$$\begin{aligned} E[\theta_{n,0}^4] &= E\left[\left(f(x_0^{(n)}) - (f, \pi_0)\right)^4\right] \\ &\le 2^4\left(E[f(x_0^{(n)})^4] + (f, \pi_0)^4\right) < \infty, \end{aligned}$$

where the last inequality follows from $E[f(x_0^{(n)})^4] = (f^4, \pi_0)$ and the assumption $f \in \mathsf{F}_T^4$. Since the variables $\{\theta_{n,0}; n = 1, ..., N\}$ satisfy the assumptions of Lemma A.1, we readily obtain

$$\lim_{N\to\infty} \left|(f, \pi_0^N) - (f, \pi_0)\right| = \lim_{N\to\infty} \left|\frac{1}{N}\sum_{n=1}^N f(x_0^{(n)}) - (f, \pi_0)\right| = 0 \quad \text{a.s.} \tag{A.6}$$

Consider the (predictive) measure $\xi_{t+1} = \tau_{t+1}\pi_t$ as defined in Eq. (2.7). After the sampling step, the particle filter produces a random approximation

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

i.e., $\xi_{t+1}^N = \tau_{t+1}\pi_t^N$. We look now into the approximation error $|(f, \xi_1^N) - (f, \xi_1)|$.

Let $\mathcal{F}_t = \sigma\left(x_{0:t}^{(n)}, \bar{x}_{1:t}^{(n)} : \quad 1 \le n \le N\right)$ denote the σ-algebra generated by the random variables $x_s^{(n)}$ and $\bar{x}_r^{(n)}$ for $n = 1, ..., N$, $s = 0, ..., t$ and $r = 1, ..., t$. It is apparent that $E[f(\bar{x}_{t+1}^{(n)})|\mathcal{F}_t] = \tau_{t+1}(f)(x_t^{(n)})$, hence the conditional mean an of (f, ξ_{t+1}^N) is

$$\begin{aligned} E[(f, \xi_{t+1}^N)|\mathcal{F}_t] &= \frac{1}{N}\sum E[f(\bar{x}_{t+1}^{(n)})|\mathcal{F}_t] \\ &= \frac{1}{N}\sum_{n=1}^N \tau_{t+1}(f)(x_t^{(n)}) = (f, \tau_{t+1}\pi_t^N), \end{aligned}$$

and it is natural to use the triangular inequality

$$\left|(f,\xi_{t+1}^N)-(f,\xi_{t+1})\right| \leq \left|(f,\xi_{t+1}^N)-(f,\tau_{t+1}\pi_t^N)\right| + \left|(f,\tau_{t+1}\pi_t^N)-(f,\tau_{t+1}\pi_t)\right| \quad \text{(A.7)}$$

to analyze the approximation error $|(f,\xi_{t+1}^N)-(f,\xi_{t+1})|$.

We proceed with the case $t=0$. If we look into the the second term on the right hand side of (A.7) we observe that

$$(f,\tau_1\pi_0) = \int \tau_1(f)(x_0)\pi_0(dx_0) = (\tau_1(f),\pi_0)$$

and, similarly, $(f,\tau_1\pi_0^N) = (\tau_{t+1}(f),\pi_t^N)$. Since $f\in \mathsf{F}_T^4$, it follows that $\tau_{t+1}(f)\in\mathsf{F}_T^4$ as well and, from Eq. (A.6), we readily see that

$$\lim_{N\to\infty}\left|(f,\tau_1\pi_0^N)-(f,\tau_1\pi_0)\right| = 0 \quad \text{a.s.} \quad \text{(A.8)}$$

In order to analyze the first term on the right hand side of (A.7) (for $t=0$), let us choose the random variables

$$\bar\theta_{1,n} = f(\bar x_1^{(n)}) - \tau_1(f)(x_0^{(n)}), \quad n=1,...,N.$$

It is straightforward to check that they are unconditionally i.i.d. To see that they are centered, simply observe that, for every $n=1,...,N$,

$$E[\bar\theta_{1,n}] = E\left[E[\bar\theta_{1,n}|\mathcal{F}_0]\right] = 0,$$

since $E[\bar\theta_{1,n}|\mathcal{F}_0] = E[f(\bar x_1^{(n)})|\mathcal{F}_0] - \tau_1(f)(x_0^{(n)})$ and $E[f(\bar x_1^{(n)})|\mathcal{F}_0] = \tau_1(f)(x_0^{(n)})$. Moreover,

$$E\left[\bar\theta_{1,n}^4|\mathcal{F}_0\right] \leq 2^4\left(\tau_1(f^4)(x_0^{(n)}) + (\tau_1(f)^4,\pi_0)\right),$$

hence

$$E\left[\bar\theta_{1,n}^4\right] = E\left[E\left[\bar\theta_{1,n}^4|\mathcal{F}_0\right]\right] \leq 2^4\left((\tau_1(f^4),\pi_0) + (\tau_1(f)^4,\pi_0)\right) < \infty,$$

where the last inequality holds because $f\in\mathsf{F}_T^4$ and, as a consequence $(\tau_1(f)^4,\pi_0)\leq(\tau_1(f^4),\pi_0)<\infty$. As the variables $\bar\theta_{1,n} = f(\bar x_1^{(n)}) - \tau_1(f)(x_0^{(n)})$ satisfy the assumptions of Lemma A.1, we readily obtain that

$$\lim_{N\to\infty}\left|(f,\xi_1^N)-(f,\tau_1\pi_0^N)\right| = 0 \quad \text{a.s.} \quad \text{(A.9)}$$

Taking together (A.7), (A.9) and (A.8) we obtain

$$\lim_{N\to\infty}\left|(f,\xi_1^N)-(f,\xi_1)\right| = 0 \quad \text{a.s.} \quad \text{(A.10)}$$

After computing the importance weights we obtain the random measure $\bar\pi_1^N(dx)$ defined in Eq. (A.5) (with $t=1$). Integrals w.r.t. $\bar\pi_1^N$ can be written in terms of $g_1^{y_1}$ and ξ_1^N, namely

$$(f,\bar\pi_1^N) = \sum_{n=1}^N \frac{g_1^{y_1}(\bar x_1^{(n)})}{\sum_{k=1}^N g_1^{y_1}(\bar x_1^{(k)})} f(\bar x_1^{(n)}) = \frac{(fg_1^{y_1},\xi_1^N)}{(g_1^{y_1},\xi_1^N)}.$$

Similarly, for π_1 and ξ_1 Bayes' theorem yields

$$(f, \pi_1) = \frac{(fg_1^{y_1}, \xi_1)}{(g_1^{y_1}, \xi_1)},$$

and the difference $(f, \bar{\pi}_1^N) - (f, \pi_1)$ can be written as

$$\begin{aligned}(f, \bar{\pi}_1^N) - (f, \pi_1) &= \frac{(fg_1^{y_1}, \xi_1^N)}{(g_1^{y_1}, \xi_1^N)} - \frac{(fg_1^{y_1}, \xi_1)}{(g_1^{y_1}, \xi_1)} \pm \frac{(fg_1^{y_1}, \xi_1^N)}{(g_1^{y_1}, \xi_1)} \\ &= \frac{(fg_1^{y_1}, \xi_1^N) - (fg_1^{y_1}, \xi_1)}{(g_1^{y_1}, \xi_1)} + \frac{(fg_1^{y_1}, \xi_1^N)}{(g_1^{y_1}, \xi_1^N)} \times \frac{(g_1^{y_1}, \xi_1) - (g_1, \xi_1^N)}{(g_1^{y_1}, \xi_1)} \quad \text{(A.11)}\end{aligned}$$

Note that, since $g_1^{y_1} \in B(\mathbb{R}^{d_x})$ (hence $g_1^{y_1} \in \mathsf{F}_T^4$), Eq. (A.10) yields

$$\lim_{N\to\infty} \left|(g_1^{y_1}, \xi_1) - (g_1, \xi_1^N)\right| = 0 \quad \text{a.s.} \tag{A.12}$$

and, since we have assumed

$$(g_1^{y_1}, \xi_1) > 0, \tag{A.13}$$

it follows that

$$\lim_{N\to\infty} (g_1, \xi_1^N) > 0 \quad \text{a.s.} \tag{A.14}$$

Also as a consequence of the likelihood $g_1^{y_1}$ being bounded, we have $fg_1^{y_1} \in \mathsf{F}_T^4$ and (A.10) guarantees that

$$\lim_{N\to\infty} \left|(fg_1^{y_1}, \xi_1) - (fg_1, \xi_1^N)\right| = 0 \quad \text{a.s.} \tag{A.15}$$

Taking Eqs. (A.13) and (A.14) together, we deduce that $\lim_{N\to\infty} \frac{(fg_1^{y_1}, \xi_1^N)}{(g_1^{y_1}, \xi_1^N)} < \infty$ a.s. This result, combined with (A.12) and (A.15) yields

$$\lim_{N\to\infty} \left|(f, \bar{\pi}_1^N) - (f, \pi_1)\right| = 0 \quad \text{a.s.} \tag{A.16}$$

A.2. Induction step ($t > 1$)

Let us assume that $\lim_{N\to\infty} \left|(f, \bar{\pi}_t^N) - (f, \pi_t)\right| = 0$ a.s. for some $1 \le t < T$.

We first show that the difference $|(f, \pi_t^N) - (f, \bar{\pi}_t)|$ converges to 0 a.s. Recall that π_t^N is obtained from the equally-weighted particles after the resampling step. Let us introduce the generated σ-algebra $\bar{\mathcal{F}}_t = \sigma\left(x_{0:t-1}^{(n)}, \bar{x}_{1:t}^{(n)}; \quad 1 \le n \le N\right)$ and the random variables

$$\theta_{t,n} = f(x_t^{(n)}) - (f, \bar{\pi}_t^N), \quad n = 1, ..., N.$$

It is simple to check that

$$E\left[f(x_t^{(n)})|\bar{\mathcal{F}}_t\right] = (f, \bar{\pi}_t^N), \quad n = 1, \ldots, N,$$

hence $\theta_{t,n}$, $n = 1, ..., N$, are centered (and obviously i.i.d.) given $\bar{\mathcal{F}}_t$. Also, $E[\theta_{t,n}^4|\bar{\mathcal{F}}_t] < \infty$. Specifically, $(f^4, \bar{\pi}_t^N)$ is $\bar{\mathcal{F}}_t$-measurable hence

$$E\left[f(x_t^{(n)})^4|\bar{\mathcal{F}}_t\right] = (f^4, \bar{\pi}_t^N),$$

and, from the induction hypothesis and $f \in \mathsf{F}_T^4$,

$$\lim_{N\to\infty} (f^4, \bar{\pi}_t^N) = (f^4, \pi_t) < \infty \quad \text{a.s.}$$

Therefore, for sufficiently large N, $(f^4, \bar{\pi}_t^N) < \infty$. However,

$$E[\theta_{t,n}^4|\bar{\mathcal{F}}_t] \leq 2^4 \left((f^4, \bar{\pi}_t^N) + (f, \bar{\pi}_t^N)^4\right),$$

hence $E[\theta_{t,n}^4] = E\left[E[\theta_{t,n}^4|\bar{\mathcal{F}}_t]\right] < \infty$ for sufficiently large N. As the conditions of Lemma A.1 are satisfied for $\theta_{t,n}$, $n = 1, ..., N$, we obtain

$$\lim_{N\to\infty} \left|(f, \pi_t^N) - (f, \bar{\pi}_t^N)\right| = \lim_{N\to\infty} \left|\frac{1}{N}\sum_{n=1}^{N} \theta_{t,n}\right| = 0 \quad \text{a.s.} \tag{A.17}$$

Finally, taking together the induction hypothesis, (A.17) and the triangle inequality

$$\left|(f, \pi_t^N) - (f, \pi_t)\right| \leq \left|(f, \pi_t^N) - (f, \bar{\pi}_t^N)\right| + \left|(f, \bar{\pi}_t^N) - (f, \pi_t)\right|,$$

readily yields

$$\lim_{N\to\infty} \left|(f, \pi_t^N) - (f, \pi_t)\right| = 0 \quad \text{a.s.} \tag{A.18}$$

Next, we prove that $\lim_{N\to\infty} |(f, \xi_{t+1}^N) - (f, \xi_{t+1})| = 0$ a.s. We resort again to the triangular inequality (A.7). Since $(f, \tau_{t+1}\pi_t) = (\tau_{t+1}(f), \pi_t)$, $(f, \tau_{t+1}\pi_t^N) = (\tau_{t+1}(f), \pi_t^N)$ and $\tau_{t+1}(f) \in \mathsf{F}_T^4$, it is a straightforward consequence of (A.18) that

$$\lim_{N\to\infty} \left|(f, \tau_{t+1}\pi_t^N) - (f, \tau_{t+1}\pi_t)\right| = 0 \quad \text{a.s.} \tag{A.19}$$

To show that the error $(f, \xi_{t+1}^N) - (f, \tau_{t+1}\pi_t^N)$ also vanishes, let us choose the random variables $\bar{\theta}_{t+1,n} = f(\bar{x}_{t+1}^{(n)}) - \tau_{t+1}(f)(x_t^{(n)})$. These are i.i.d.[5] conditional on $\bar{\mathcal{F}}_t$. They are also centered, since $E[\bar{\theta}_{t+1,n}|\mathcal{F}_t] = E[f(\bar{x}_{t+1}^{(n)})|\mathcal{F}_t] - \tau_{t+1}(f)(x_t^{(n)}) = 0$ and $\bar{\mathcal{F}}_t \subset \mathcal{F}_t$. Therefore, we just need to check that $E[\bar{\theta}_{t+1,n}^4] < \infty$ in order to apply Lemma A.1. We note that

$$E[\bar{\theta}_{t+1,n}^4|\mathcal{F}_t] \leq 2^4 \left(\tau_{t+1}(f^4)(x_t^{(n)}) + \tau_{t+1}(f)^4(x_t^{(n)})\right)$$

[5]In particular, note that

- $\{\bar{x}_{t+1}^{(n)}\}_{n=1,...,N}$ can be viewed as i.i.d. samples from the probability measure $m_{t+1}(dx) = \sum_{n=1}^N w_t^{(n)} \tau_{t+1}(dx|\bar{x}_t^{(n)})$, where both $w_t^{(n)}$ and $\bar{x}_t^{(n)}$, $1 \leq n \leq N$, are $\bar{\mathcal{F}}_t$-measurable, and
- $\{x_t^{(n)}\}_{n=1,...,N}$ are also i.i.d. given $\bar{\mathcal{F}}_t$.

and then readily obtain

$$\begin{aligned} E[\bar{\theta}^4_{t+1,n}|\bar{\mathcal{F}}_t] &= E\left[E[\bar{\theta}_{t+1,n}|\mathcal{F}_t]|\bar{\mathcal{F}}_t\right] \\ &\leq 2^4\left((\tau_{t+1}(f^4),\bar{\pi}^N_t)+(\tau_{t+1}(f)^4,\bar{\pi}^N_t)\right). \end{aligned} \tag{A.20}$$

However, $\tau_{t+1}(f)^4 \leq \tau_{t+1}(f^4)$ and $f \in \mathsf{F}^4_T$ implies that $(\tau_{t+1}(f^4),\pi_t) < \infty$. Moreover, the induction hypothesis yields

$$\lim_{N\to\infty}\left|(\tau_{t+1}(f^4),\bar{\pi}^N_t)-(\tau_{t+1}(f^4),\bar{\pi}_t)\right| = 0 \quad \text{a.s.}$$

hence $(\tau_{t+1}(f)^4,\bar{\pi}^N_t) \leq (\tau_{t+1}(f^4),\bar{\pi}^N_t) < \infty$ for sufficiently large N. As a consequence, $E[\bar{\theta}^4_{t+1,n}] < \infty$ and the conditions of Lemma A.1 are satisfied for the random variables $\bar{\theta}_{t+1,n}$ and the σ-algebra $\bar{\mathcal{F}}_t$. In particular, we have

$$\lim_{N\to\infty}\left|(f,\xi^N_{t+1})-(f,\tau_{t+1}\pi^N_t)\right| = \lim_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^N \bar{\theta}_{t+1,n}\right| = 0 \quad \text{a.s.} \tag{A.21}$$

Taking together (A.21), (A.19) and (A.7) yields

$$\lim_{N\to\infty}\left|(f,\xi^N_{t+1})-(f,\xi_{t+1})\right| = 0 \quad \text{a.s.} \tag{A.22}$$

Finally, given (A.22), it is straightforward to prove that $\lim_{N\to\infty}|(f,\bar{\pi}^N_{t+1}) - (f,\pi_{t+1})| = 0$ a.s. using the same argument as in the base case for $\bar{\pi}^N_1$.

Appendix B: Proof of Lemma 4.1

Choose a constant β such that $\nu < \beta < p-1$ and define

$$U^{\beta,p} = \sum_{m=1}^{\infty} m^{p-1-\beta}(\theta^m)^p.$$

The random variable $U^{\beta,p}$ is obviously non-negative and, additionally, it has a finite mean, $E[U^{\beta,p}] < \infty$. Indeed, from Fatou's lemma

$$E\left[U^{\beta,p}\right] \leq \sum_{m=1}^{\infty} m^{p-1-\beta}E[(\theta^m)^p] \leq c\sum_{m=1}^{\infty} m^{-1-\beta+\nu},$$

where the second inequality follows from Eq. (4.1). Since $\beta-\nu > 0$, it follows that $\sum_{m=1}^{\infty} m^{-1-(\beta-\nu)} < \infty$, hence $E[U^{\beta,p}] < \infty$.

We use the so-defined random variable $U^{\beta,p}$ in order to determine the convergence rate of θ^k. Obviously, $k^{p-1-\beta}(\theta^k)^p \leq U^{\beta,p}$ and solving for θ^k yields

$$\theta^k \leq \frac{\left(U^{\beta,p}\right)^{\frac{1}{p}}}{k^{1-\frac{1+\beta}{p}}}.$$

If we define $\varepsilon = \frac{1+\beta}{p}$ and $U^{\varepsilon} = \left(U^{\beta,p}\right)^{\frac{1}{p}}$, the we obtain the inequality

$$\theta^k \leq \frac{U^{\varepsilon}}{k^{1-\varepsilon}}.$$

Since $E[U^{\beta,p}] < \infty$, it follows that $E\left[(U^{\varepsilon})^p)\right] < \infty$, hence U^{ε} is a.s. finite. Also, we recall that $\nu < \beta < p-1$, therefore $\frac{1+\nu}{p} < \varepsilon < 1$.

Acknowledgements

The work of D. Crisan was partially supported by the EPSRC Grant No: EP/H0005500/1. The work of J. Míguez was partially supported by *Ministerio de Economía y Competitividad* of Spain (program Consolider-Ingenio 2010 CSD2008-00010 COMONSENS and project TEC2012-38883-C02-01 COMPREHENSION) and *Ministerio de Educación, Cultura y Deporte* of Spain (*Programa Nacional de Movilidad de Recursos Humanos* PRX12/00690).

References

[1] C. Abraham, G. Biau, and B. Cadre. On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics*, 8:1–11, 2004.

[2] M. J. Appel, R. Labarre, and D. Radulovic. On accelerated random search. *SIAM Journal on Optimization*, 14(3):708–730, 2003.

[3] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2008.

[4] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–40, 1997.

[5] M. J. Brewer. A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing*, 10(4):299–309, 2000.

[6] A. Corana, M. Marchesi, C. Martini, and S. Ridella. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software*, 13(3):262–280, September 1987.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York (USA), 1991.

[8] D. Crisan. Particle filters - a theoretical perspective. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 2, pages 17–42. Springer, 2001.

[9] D. Crisan, P. Del Moral, and T.J. Lyons. Non-linear filtering using branching and interacting particle systems. *Markov Processes Related Fields*, 5(3):293–319, 1999.

[10] D. Crisan and A. Doucet. Convergence of sequential Monte Carlo methods. Technical Report Cambridge University (CUED/FINFENG/TR381), 2000.

[11] D. Crisan and A. Doucet. A survey of convergence results on particle filtering. *IEEE Transactions Signal Processing*, 50(3):736–746, March 2002.

[12] T. A. Dean, S. S. Singh, A. Jasra, and G. W. Peters. Parameter estimation for hidden Markov models with intractable likelihoods. *arXiv*, 1103.5399v1[math.ST], 2011.

[13] P. Del Moral. Non-linear filtering: interacting particle solution. *Markov Processes and Related Fields*, 2:555–580, 1996.

[14] P. Del Moral. Non-linear filtering using random particles. *Theory of Probability and its Applications*, 40(4):690–701, 1996.

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

[16] P. Del Moral, A. Doucet, and S. Singh. Uniform stability of a particle approximation of the optimal filter derivative. *arXiv*, 1106.2525v1[math.ST], 2011.

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

[18] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, 1985.

[19] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4^{th} International Symposium on Image and Signal Processing and Analysis*, pages 64–69, September 2005.

[20] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 1, pages 4–14. Springer, 2001.

[21] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York (USA), 2001.

[22] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[23] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

[24] L. Frenkel and M. Feder. Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking. *IEEE Transactions Signal Processing*, 47(2):306–320, February 1999.

[25] J. L. Gauvain and C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11(2-3):205–213, 1992.

[26] S. Godsill, A. Doucet, and M. West. Maximum a posteriori sequence estimation using Monte Carlo particle filters. *Annals of the Institute of Statistical Mathematics*, 53(1):82–96, March 2001.

[27] N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

[28] P. Hall and K. H. Kang. Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *The Annals of Statistics*, 29(5):1443–1468, 2001.

[29] A.-R. Hedar and M. Fukushima. Derivative-free filter simulated annealing method for constrained continuous global optimization. *Journal of Global Optimization*, 35:521–549, 2006.

[30] K. Heine and D. Crisan. Uniform approximations of discrete-time filters. *Advances in Applied Probability*, 40(4):979–1001, 2008.

[31] X.L. Hu, T.B. Schon, and L. Ljung. A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing*, 56(4):1337–1348, 2008.

[32] M. M. Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.

[33] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[34] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models. *J. Comput. Graph. Statist.*, 1:1–25, 1996.

[35] H. R. Künsch. Recursive Monte Carlo filters: Algorithms and theoretical bounds. *The Annals of Statistics*, 33(5):1983–2021, 2005.

[36] F. LeGland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Annals of Applied Probability*, pages 144–187, 2004.

[37] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, September 1998.

[38] A. Logothetis and V. Krishnamurthy. Expectation maximization algorithms for MAP estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, 47(8):2139–2156, 1999.

[39] J. Míguez, D. Crisan, and P. M. Djurić. On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.

[40] C. Musso, N. Oudjane, and F. LeGland. Improving regularised particle filters. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 12, pages 247–272. Springer, 2001.

[41] K. Najim, E. Ikonen, and P. Del Moral. Open-loop regulation and tracking control based on a genealogical decision tree. *Neural Computing and Applications*, 15:339–349, 2006.

[42] M. Nilsson and W. B. Kleijn. On the estimation of differential entropy from data located on embedded manifolds. *IEEE Transactions on Information Theory*, 53(7):2330–2341, 2007.

[43] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, Boca Raton (FL, USA), 1986.

[44] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, New York, 1996.

[45] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman-Hall/CRC, 1995.

[46] M. West. Approximating posterior distributions by mixture. *Journal of the Royal Statistical Society B*, pages 409–422, 1993.

[47] X. Zhang, M. L. King, and R. J. Hyndman. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 50(11):3009 – 3031, 2006.