KYBERNETIKA — VOLUME *52* (2016), NUMBER 5, PAGES 735–756

ON CONVERGENCE OF KERNEL DENSITY ESTIMATES IN PARTICLE FILTERING

David Coufal

The paper deals with kernel density estimates of filtering densities in the particle filter. The convergence of the estimates is investigated by means of Fourier analysis. It is shown that the estimates converge to the theoretical filtering densities in the mean integrated squared error. An upper bound on the convergence rate is given. The result is provided under a certain assumption on the Sobolev character of the filtering densities. A sufficient condition is presented for the persistence of this Sobolev character over time.

Keywords: particle filter, kernel methods, Fourier analysis

Classification: 65C35

1. INTRODUCTION

The particle filter enables its user to efficiently compute integral characteristics (moments) of distributions of interest. In the filtering problem, these distributions are traditionally referred to as the *filtering distributions*. In the particle filter, the filtering distribution is approximated by an empirical measure. This measure is constructed in the form of a weighted sum of Dirac measures located at randomly (empirically) generated points called *particles*. Particles are generated sequentially by the algorithm which is an instance of the *sequential Monte Carlo methods* [4, 5].

The theoretical result that justifies the application of the particle filter is that the generated empirical measures converge to the theoretical filtering distribution as the number of particles goes to infinity [2, 4]. Approximating the filtering distribution by an empirical measure is useful for estimating moments of the distribution because they correspond to weighted sums of the values of moment functions over generated particles.

The filtering distribution has typically a density with respect to the corresponding Lebesgue measure. This density is called the *filtering density*. Knowing an analytical approximation of the filtering density has advantages. For example, the possibility of computing analytical approximations of densities of the related conditional distributions.

From these practical, and of course also theoretical, reasons the issue of the analytical approximation of the filtering densities is the subject of ongoing research. The problem has been addressed in [10, 11, 13] and recently in [3].

DOI: 10.14736/kyb-2016-5-0735

In [13], the authors refer to their previous works in which they introduced the particle filters that employ kernel estimates at different places in their computational schemes. The filters are called the pre-regularized and post-regularized particle filters, respectively; and differ in where exactly the kernel density estimate is applied in the classical particle filtering algorithm. They further introduce the local rejection regularized particle filter (L2RPF filter) and show that it generalizes the post-regularized particle filter and the KF filter introduced in [9]. The convergence analysis of the post-regularized filter is presented in details in [11].

In [10], the author investigates configurations when the acceptance-rejection method and importance sampling with an additional resampling step are used in the particle filter. The author shows that rejection sampling has a smaller asymptotic variance than the standard importance sampling resampling method. However, generally the computational effort for rejection sampling is greater than for the importance sampling. The author provides several convergence results for a kernel estimate to converge to the corresponding filtering density in terms of convergence in probability; and a version of the central limit theorem. However, the assumptions of Theorem 2 in [10], which applies to the importance sampling resampling method, exclude the common filtering settings that consider an additive Gaussian noise.

The summary discussion of the above papers is also presented in Section 3.1 of [3]. In fact, the paper [3] is the closest to our work as it addresses the application of kernel density estimation in particle filtering in a very similar way to what we do. However, our work, which is inspired by the book of Tsybakov [17], builds on Fourier analysis of kernel density estimates. This fact enables us to obtain a stronger version of certain results presented in [3], see Section 6 for a detailed discussion.

The paper presents two main results. The first result is the convergence of the kernel density estimates to the theoretical filtering density at any fixed time of the operation of the filter, provided that the number of generated particles goes to infinity. In fact, we present an upper bound on the MISE convergence rate which consequently implies the convergence of the estimates with an increasing number of generated particles. The result is based on the notion of the Sobolev character of the filtering density.

The second result gives the condition under which this Sobolev character is retained over time. Thus, the first result applies at any time of filter operation. As mentioned above, both results draws on the techniques of Fourier analysis.

The rest of the paper is organized as follows. In the next section, we review the basics of the particle filter's theory together with the related convergence results. Section 3 deals with a review of kernel methods with the focus on the Fourier analysis approach. Sections 4 and 5 present the announced main results. Section 6 discusses our results in the context of the results provided in [3] as these address the same problem, but using different assumptions; and the paper is concluded by Section 7.

2. PARTICLE FILTER

The basics of the particle filter and general filtering theory can be found, for example, in [2, 4, 5, 6] and [15]. Nevertheless, we present here the essential framework of the related methodology in order that the paper be self-contained.

2.1. Filtering problem

The filtering problem lies in determining the optimal estimate of the inaccessible state of a stochastic process on the basis of accessible observations. The observations constitute a stochastic process called the *observation process*. The observation process is interconnected with a principal stochastic process which is called the *signal process*. The states of the signal process are then subject to estimation. Let us be more specific.

Let $\{\boldsymbol{X}_t\}_{t=0}^{\infty}$, $\{\boldsymbol{Y}_t\}_{t=1}^{\infty}$ be two stochastic processes specified on a common probabilistic space $(\Omega, \mathcal{A}, P)$. The first process $\{\boldsymbol{X}_t\}_{t=0}^{\infty}$, $\boldsymbol{X}_t : (\Omega, \mathcal{A}) \to (\mathbb{R}^{d_x}, \mathcal{B}(\mathbb{R}^{d_x}))$, $t \in \mathbb{N}_0$, $d_x \in \mathbb{N}$ is the signal process. It represents generally an inhomogeneous Markov chain with a continuous state space $\mathbb{R}^{d_x}$, endowed with its standard Borel σ-algebra $\mathcal{B}(\mathbb{R}^{d_x})$. The probabilistic behavior of the chain is determined by the initial distribution $\pi_0(\mathrm{d}\boldsymbol{x}_0)$ of $\boldsymbol{X}_0$ and by the set of transition kernels $K_{t-1} : \mathcal{B}(\mathbb{R}^{d_x}) \times \mathbb{R}^{d_x} \to [0,1]$, $t \in \mathbb{N}$. We denote by $K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ the measure induced by the transition kernel K_{t-1} for $\boldsymbol{x}_{t-1} \in \mathbb{R}^{d_x}$ being fixed.

The second process $\{\boldsymbol{Y}_t\}_{t=1}^{\infty}$, $\boldsymbol{Y}_t : (\Omega, \mathcal{A}) \to (\mathbb{R}^{d_y}, \mathcal{B}(\mathbb{R}^{d_y}))$, $t \in \mathbb{N}$, $d_y \in \mathbb{N}$ is the observation process. As above, $\mathbb{R}^{d_y}$ is the state space of the process and $\mathcal{B}(\mathbb{R}^{d_y})$ the corresponding Borel σ-algebra. The process is specified on the basis of the signal process by the formula

$$\boldsymbol{Y}_t = h_t(\boldsymbol{X}_t) + \boldsymbol{V}_t, \quad t \in \mathbb{N} \tag{1}$$

where $h_t : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$, $t \in \mathbb{N}$ are Borel functions and $\boldsymbol{V}_t : (\Omega, \mathcal{A}) \to (\mathbb{R}^{d_y}, \mathcal{B}(\mathbb{R}^{d_y}))$ are i.i.d. random variables that are independent from $\boldsymbol{X}_{0:t} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_t)$ for all $t \in \mathbb{N}$. The independence of $\boldsymbol{V}_t$ transfers on observations, due to (1) and the Markov character of $\{\boldsymbol{X}_t\}_{t=0}^{\infty}$, in the following way:

$$P(\boldsymbol{Y}_t \in \mathrm{d}\boldsymbol{y}_t|\boldsymbol{X}_{0:t}, \boldsymbol{Y}_{1:t-1}) = P(\boldsymbol{Y}_t \in \mathrm{d}\boldsymbol{y}_t|\boldsymbol{X}_t). \tag{2}$$

For $t=1$, the left-hand side reads as $P(\boldsymbol{Y}_1 \in \mathrm{d}\boldsymbol{y}_1|\boldsymbol{X}_{0:1})$.

In the paper, the colon is used to denote finite sequences, e. g., $\boldsymbol{Y}_{1:t-1} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{t-1})$ or $\boldsymbol{y}_{1:t} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$, etc.

2.2. Filtering distribution and filtering density

As stated, the purpose of filtering is to provide the optimal estimate of the current state $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ of the signal process using the current and past observations $\boldsymbol{y}_{1:t} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$. This is done at each time instant $t \in \mathbb{N}$. It is a classical result that under the assumption of L_2 integrability of $\boldsymbol{X}_t$, the L_2-optimal estimate corresponds to the conditional expectation $\mathbb{E}[\boldsymbol{X}_t|\boldsymbol{Y}_{1:t}]$. In what follows we will assume that $\boldsymbol{X}_t \in L_2(\Omega, \mathcal{A}, P)$ for each $t \in \mathbb{N}_0$.

For fixed observations $\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}$, the conditional expectation $\mathbb{E}[\boldsymbol{X}_t|\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}]$ is determined by the related conditional distribution $P(\boldsymbol{X}_t \in \mathrm{d}\boldsymbol{x}_t|\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t})$. This distribution then represents the filtering distribution at time $t \in \mathbb{N}$ and is approximated by an empirical measure generated by the particle filter.

For the filtering problem discussed in this paper, it is assumed that all the involved finite-dimensional distributions have *bounded densities* with respect to the corresponding Lebesgue measures. Namely, we assume that $\pi_0(\mathrm{d}\boldsymbol{x}_0) = p_0(\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_0$, $K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) =$

$K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_t$ and $P(\boldsymbol{V}_t \in \mathrm{d}\boldsymbol{v}_t) = g_t^v(\boldsymbol{v}_t)\,\mathrm{d}\boldsymbol{v}_t$. For the purposes of the convergence results in Section 2.5, it is assumed the g_t^v is *strictly positive*, i. e., $g_t^v(\boldsymbol{v}_t) > 0$ for all t. This enables us to express the respective filtering density, i. e., the density of $P(\boldsymbol{X}_t \in \mathrm{d}\boldsymbol{x}_t|\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t})$.

The conditional density of $P(\boldsymbol{Y}_t \in \mathrm{d}\boldsymbol{y}_t|\boldsymbol{X}_t = \boldsymbol{x}_t)$ is determined by formula (1). The density is denoted by $g_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$ and writes

$$g_t(\boldsymbol{y}_t|\boldsymbol{x}_t) = g_t^v(\boldsymbol{y}_t - h_t(\boldsymbol{x}_t)). \tag{3}$$

The joint density of $(\boldsymbol{X}_{0:t}, \boldsymbol{Y}_{1:t})$ has the form

$$p(\boldsymbol{x}_{0:t}, \boldsymbol{y}_{1:t}) = p_0(\boldsymbol{x}_0) \prod_{k=1}^{t} g_k(\boldsymbol{y}_k|\boldsymbol{x}_k)\, K_{k-1}(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}). \tag{4}$$

These specifications are induced by the conditional independence of observations (2) and by the standard theory of Markov chains with a continuous state space.

The filtering density at time $t \in \mathbb{N}$ is denoted by $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. Employing the joint density (4), it is expressed as

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_{1:t})}{p(\boldsymbol{y}_{1:t})} = \frac{\int p(\boldsymbol{x}_{0:t}, \boldsymbol{y}_{1:t})\,\mathrm{d}\boldsymbol{x}_{0:t-1}}{\int p(\boldsymbol{x}_{0:t}, \boldsymbol{y}_{1:t})\,\mathrm{d}\boldsymbol{x}_{0:t}}. \tag{5}$$

The above integrals are generally inexpressible in a closed form. However, certain recursive analytical relations can be stated. These relations are called the *filtering equations* and are addressed in the next section.

2.3. Filtering equations

The filtering equations describe recursively development of the filtering density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ over time. They consist of the *prediction formula* (6) and the *update formula* (7).

The prediction formula gives the expression for the so-called prediction density which is the density of $P(\boldsymbol{X}_t \in \mathrm{d}\boldsymbol{x}_t|\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t-1})$. The update formula then gives the specification of the filtering density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$.

Lemma 2.1. Let the joint density of $(\boldsymbol{X}_{0:t}, \boldsymbol{Y}_{1:t})$ be given by formula (4), then

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) = \int K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}, \tag{6}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{g_t(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{\int g_t(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})\,\mathrm{d}\boldsymbol{x}_t}, \quad t \in \mathbb{N} \tag{7}$$

with $p(\boldsymbol{x}_1|\boldsymbol{y}_{1:0})$ understood as $p(\boldsymbol{x}_1)$ and $p(\boldsymbol{x}_0|\boldsymbol{y}_{1:0})$ as $p(\boldsymbol{x}_0)$.

Proof. The basic proof can be found, for example, in [15], see Theorem 4.1 on page 54; or in Section 2.6.2 of [4] where it is presented in a more general form. □

Development of the filtering density over time is split into two sub-steps by the filtering equations. The prediction density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$ is obtained in the first sub-step

and, in the second one, is updated to the filtering density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ on the basis of the current observation $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$.

Speaking in the language of distributions, the filtering distribution is usually denoted by π_t, i. e., $\pi_t(\mathrm{d}\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})\,\mathrm{d}\boldsymbol{x}_t$. The filtering distribution is also alternatively referred to as the *update distribution (measure)*. The prediction density then corresponds to the density of the so-called *prediction distribution (measure)* denoted by $\overline{\pi}_t$, i. e., $\overline{\pi}_t(\mathrm{d}\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})\,\mathrm{d}\boldsymbol{x}_t$.

2.4. Particle filter

Time development of the filtering distribution can be seen as a recursive alternation between the prediction and update distributions $\overline{\pi}_t$ and π_t. This characterization fits to particle filter operation because the filter alternately generates empirical prediction and update measures.

In the particle filter, empirical measures are constructed as weighted sums of Dirac measures localized at particles generated by the filter. The justification of this representation stems from the Strong Law of Large Numbers (SLLN). Assuming that $\{\boldsymbol{X}_i = \boldsymbol{x}_i\}_{i=1}^n$, $n \in \mathbb{N}$ is an i.i.d. sample from a given distribution μ, i. e., $\boldsymbol{X}_i \sim \mu$, and constructing the empirical measure $\delta_n(\mathrm{d}\boldsymbol{x})$ as

$$\delta_n(d\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}(\mathrm{d}\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}(\mathrm{d}\boldsymbol{x}) \tag{8}$$

the SLLN states that for any integrable function f, the integral over this empirical measure converges a.s. to the integral over the distribution μ. Note that in (8), the second expression points out the random character of $\delta_n(\mathrm{d}\boldsymbol{x})$, in fact, $\delta_n(\mathrm{d}\boldsymbol{x})$ is a random measure.

Dealing with the filtering problem practically, we are not able to directly generate i.i.d. samples from π_t because we do not have any closed-form representation of the filtering density at our disposal. However, due to the product character of the joint density $p(\boldsymbol{x}_{0:t}, \boldsymbol{y}_{1:t})$, one can state an algorithm which recursively generates samples (particles) that are used for constructing empirical counterparts of $\overline{\pi}_t$ and π_t.

The construction of empirical measures proceeds sequentially. The particles generated in the previous cycle of operation are employed in the current cycle. A stochastic update of particles and their weights is taken in each cycle. The weights are updated on the basis of the current observation. The procedure is in fact an instance of the sequential Monte Carlo methods applied in the context of the filtering problem [4]; and the algorithm follows the recursion described by the filtering equations. However, there is one extension.

In the raw mode of operation, the update measure is constructed as a non-uniformly weighted sum of Dirac measures. As explained in [4], as $t \in \mathbb{N}$ increases, the distribution of weights becomes more and more skewed and after a few time steps only a single particle has a non-zero weight. To avoid this degeneracy, the *resampling step* is introduced.

During the resampling step, a non-uniformly weighted empirical measure is resampled into its uniformly weighted counterpart. The basic type of resampling draws on the idea of discarding particles with low weights (with respect to $1/n$) and promote

Algorithm 1 : Operation of the particle filter.

0. declarations
$n \in \mathbb{N}$ - the number of particles,
$T \in \mathbb{N}$ - the computational horizon,
p_0 - the initial density of $\boldsymbol{X}_0$,
$K_{t-1}(\cdot\,|\boldsymbol{x}_{t-1})$, $t = 1, \dots T$ - the transition densities.

1. initialization
$t = 0$,
sample $\{\overline{\boldsymbol{x}}_0^i \sim p_0\}_{i=1}^n$,
constitute $\widehat{\pi}_0^n(\mathrm{d}\boldsymbol{x}_0) = \frac{1}{n}\sum_{i=1}^n \delta_{\overline{\boldsymbol{x}}_0^i}(\mathrm{d}\boldsymbol{x}_0)$,
set $\pi_0^n(\mathrm{d}\boldsymbol{x}_0) = \widehat{\pi}_0^n(\mathrm{d}\boldsymbol{x}_0)$, i. e., $\{\boldsymbol{x}_0^i = \overline{\boldsymbol{x}}_0^i\}_{i=1}^n$.

2. sampling
$t = t + 1$,
sample $\{\overline{\boldsymbol{x}}_t^i \sim K_{t-1}(\cdot\,|\boldsymbol{x}_{t-1}^i)\}_{i=1}^n$,
for $i = 1\!:\!n$ compute

$$\widetilde{w}(\overline{\boldsymbol{x}}_t^i) = \frac{g_t(\boldsymbol{y}_t - h_t(\overline{\boldsymbol{x}}_t^i))}{\sum_{j=1}^n g_t(\boldsymbol{y}_t - h_t(\overline{\boldsymbol{x}}_t^j))},$$

constitute $\widehat{\pi}_t^n(\mathrm{d}\boldsymbol{x}_t) = \sum_{i=1}^n \widetilde{w}(\overline{\boldsymbol{x}}_t^i)\,\delta_{\overline{\boldsymbol{x}}_t^i}(\mathrm{d}\boldsymbol{x}_t)$.

3. resampling
using $\mathcal{M}(n, \widetilde{w}(\overline{\boldsymbol{x}}_t^1), \dots, \widetilde{w}(\overline{\boldsymbol{x}}_t^n))$, resample $\{\boldsymbol{x}_t^i\}_{i=1}^n$ from $\{\overline{\boldsymbol{x}}_t^i\}_{i=1}^n$ and constitute $\pi_t^n(\mathrm{d}\boldsymbol{x}_t) = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_t^i}(\mathrm{d}\boldsymbol{x}_t)$.

4. if $t = T$ end, else go to step 2.

those with high weights. Practically it is done by sampling with replacement from the set of original particles with the probabilities of selection given by the original particles' weights. It means that the resampled particles might be duplicated. In fact, the numbers of duplicates corresponds to a sample from the multinomial distribution $\mathcal{M}(n, \widetilde{w}(\overline{\boldsymbol{x}}_t^1), \dots, \widetilde{w}(\overline{\boldsymbol{x}}_t^n))$. Let us stress here that the resampled particles *does not constitute an i.i.d. sample.*

We are now ready to present the operation of the particle filter in the algorithmic way, see Algorithm 1. Note that in the presented pseudocode the particles, i. e., points from $\mathbb{R}^{d_x}$, are denoted by bold lowercase letters with the ^{i} superscript.

The particle filter sequentially generates three empirical measures in each single cycle of its operation. These are the empirical prediction measure $\overline{\pi}_t^n$, the empirical update measure before resampling $\widehat{\pi}_t^n$ and the empirical update measure after resampling π_t^n. The third measure then forms the empirical counterpart of the filtering distribution π_t.

A comparison of developments of the empirical measures and the theoretical distributions is presented in Figure 1.

$$\begin{array}{ccccccccccc}
\pi_0 & \to & \overline{\pi}_1^n & \to & \widehat{\pi}_1^n \to \pi_1^n & \to \ldots \to & \overline{\pi}_t^n & \to & \widehat{\pi}_t^n \to \pi_t^n \\
\pi_0 & \to & \overline{\pi}_1 & \to & \pi_1 & \to \ldots \to & \overline{\pi}_t & \to & \pi_t
\end{array}$$

Fig. 1. Development of the empirical and theoretical distributions in the particle filter.

2.5. Convergence results

In the particle filter, it is known that the empirical measures $\overline{\pi}_t^n$ and π_t^n converge weakly a.s. (they are the random measures) to their theoretical counterparts as the number of generated particles goes to infinity. We will not go into details of the proof of the assertion, we only mention the result and its L_2 variant related to our research. To present the convergence theorems, we denote the class of real bounded functions on $\mathbb{R}^{d_x}$ by $B(\mathbb{R}^{d_x})$, the class of real bounded and continuous functions on $\mathbb{R}^{d_x}$ by $\mathcal{C}_b(\mathbb{R}^{d_x})$, the supremum norm of a function $f : \mathbb{R}^{d_x} \to \mathbb{R}$ by $||f||_\infty$ and the integral of f over the measure μ by μf. Further, it is assumed that the transition kernels of the signal process possess the Feller property. That is, $K_{t-1}f \in \mathcal{C}_b(\mathbb{R}^{d_x})$ for any $f \in \mathcal{C}_b(\mathbb{R}^{d_x})$ and $t \in \mathbb{N}$ where $(K_{t-1}f)(\boldsymbol{x}_{t-1}) = \int f(\boldsymbol{x}_t)K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. The other assumption is that the densities $g_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$ of (3), $t \in \mathbb{N}$ are bounded, continuous and strictly positive functions.

Theorem 2.2. Let $\{\overline{\pi}_t^n\}_{t=1}^T$ and $\{\pi_t^n\}_{t=1}^T$ be the sequences of empirical measures generated by the particle filter for some fixed observation history $\{\boldsymbol{Y}_t = \boldsymbol{y}_t\}_{t=1}^T$, $T \in \mathbb{N}$. Then for all $t \in \{1, \ldots, T\}$ and $f \in B(\mathbb{R}^{d_x})$,

$$\lim_{n\to\infty} |\overline{\pi}_t^n f - \overline{\pi}_t f| = 0 \ \text{ a.s.}, \quad \lim_{n\to\infty} |\pi_t^n f - \pi_t f| = 0 \ \text{ a.s.}$$

P r o o f . See [4], Chapter 2 for a broader discussion of the convergence theorems. Other source is [2], Section IV. Paper [3] has even the proof of the a.s. convergence for certain unbounded functions, see Proposition 1(b). □

Theorem 2.3. Let $\{\pi_t^n\}_{t=1}^T$ be the sequence of empirical measures generated by the particle filter for some fixed observation history $\{\boldsymbol{Y}_t = \boldsymbol{y}_t\}_{t=1}^T$, $T \in \mathbb{N}$. Then for all $t \in \{1, \ldots, T\}$ and $f \in B(\mathbb{R}^{d_x})$,

$$\mathbb{E}[|\pi_t^n f - \pi_t f|^2] \le \frac{c_t^2 ||f||_\infty^2}{n} \tag{9}$$

with $c_t > 0$ being a constant for fixed $t \in \{1, \ldots, T\}$.

P r o o f . In this formulation, the theorem is presented in [2], Section V (the authors use c_t instead ours c_t^2). □

Corollary. Theorem 2.3 holds also if $f \in B^{\mathbb{C}}(\mathbb{R}^{d_x})$, i. e., if f is a bounded complex function of real variables on $\mathbb{R}^{d_x}$.

Proof. If $f \in B^{\mathbb{C}}(\mathbb{R}^{d_x})$, then $f(\boldsymbol{x}) = h(\boldsymbol{x}) + \mathrm{i}g(\boldsymbol{x})$, where i denotes the imaginary unit; and $f, g \in B(\mathbb{R}^{d_x})$. Inequality (9) then holds because for the squared modulus of $\pi_t^n f - \pi_t f$ one has $|\pi_t^n f - \pi_t f|^2 = (\pi_t^n h - \pi_t h)^2 + (\pi_t^n g - \pi_t g)^2$. □

Remark that the L_1 version of Theorem 2.3, i. e., $\mathbb{E}[|\pi_t^n f - \pi_t f|]$, is treated in [4], Theorem 2.4.1. The theorem is further mentioned for general L_p norm, $p \geq 1$ in [3], Proposition 1(a).

3. KERNEL METHODS

Kernel methods are widely used for nonparametric estimation of densities of probability distributions with the vast literature available on the topic. Here we review the very basics of the related methodology. We focus in more details on the application of Fourier analysis in this field. Our review is mainly based on the standard works of [16] and [18], and the recent book by Tsybakov [17].

3.1. Basics of kernel methods

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n \in \mathbb{N}$ be a set of independent random variables identically distributed as the real random variable $\boldsymbol{X} : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \in \mathbb{N}$. Let the distribution of $\boldsymbol{X}$ have the density $f : \mathbb{R}^d \to [0, \infty)$ with respect to the d-dimensional Lebesgue measure. A nonparametric kernel density estimate of f is constructed on the basis of an i.i.d. sample $\{\boldsymbol{X}_i = \boldsymbol{x}_i\}_{i=1}^n$ from the distribution of $\boldsymbol{X}$. The estimate is constructed as a generalization of the classical histogram by replacing the indicator function, which specifies individual bins of the histogram, by a more general function $K : \mathbb{R}^d \to \mathbb{R}$ that is commonly referred to as the *kernel function* or simply as the *kernel*.

The definition formula of the standard d-variate nonparametric kernel density estimate writes as

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{10}$$

In the formula, the second expression points out the random character of the estimate. That is, for each $\boldsymbol{x} \in \mathbb{R}^d$, the estimate $\hat{f}_n(\boldsymbol{x})$ constitutes a random variable whose distribution is determined by the distribution of $\boldsymbol{X}$ and by the value of the parameter $h > 0$ which is called the *bandwidth*.

Due to the random character of $\hat{f}_n(\boldsymbol{x})$, there is relevant the question on consistency and unbiasedness of the estimate. In the univariate case, the classical result of Parzen [14] (see also [16]) states the conditions under which the estimate is consistent. The result extends to the multivariate case, see e. g. [7]. Certain conditions are imposed on the properties of the kernel function and on development of the bandwidth h as a function of the sample size $n \in \mathbb{N}$. We mention only that h is required to develop in such a way that 1) $\lim_{n\to\infty} h(n) = 0$ and 2) $\lim_{n\to\infty} nh^d(n) = \infty$.

The investigation on the bias of $\hat{f}_n(\boldsymbol{x})$ is closely related to the investigation on the quality of the estimate in terms of the *mean squared error* (MSE). For a fixed point

$\boldsymbol{x} \in \mathbb{R}^d$, the error is specified as $\mathrm{MSE}_{\boldsymbol{x}}(\hat{f}_n) = \mathbb{E}[(\hat{f}_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2]$. Employing properties of mean and variance, it writes as

$$\mathrm{MSE}_{\boldsymbol{x}}(\hat{f}_n) = (\mathbb{E}[\hat{f}_n(\boldsymbol{x})] - f(\boldsymbol{x}))^2 + var[\hat{f}_n(\boldsymbol{x})] = (b\hat{f}_n(\boldsymbol{x}))^2 + \sigma^2\hat{f}_n(\boldsymbol{x}) \tag{11}$$

where the term $b\hat{f}_n(\boldsymbol{x}) = \mathbb{E}[\hat{f}_n(\boldsymbol{x})] - f(\boldsymbol{x})$ is the *bias* and $\sigma^2\hat{f}_n(\boldsymbol{x}) = var[\hat{f}_n(\boldsymbol{x})]$ the *variance* of the kernel density estimate $\hat{f}_n(\boldsymbol{x})$ at the point $\boldsymbol{x} \in \mathbb{R}^d$.

The $\mathrm{MSE}_{\boldsymbol{x}}(\hat{f}_n)$ is the local measure of the quality of the estimate. It is desirable to introduce also a corresponding global measure. Expectedly, such measure deals with local errors accumulated over the whole domain of the estimated density. Mathematically, the accumulation is performed by integration. This leads to the notion of the *mean integrated squared error* (MISE) of a kernel density estimate.

The MISE of the kernel density estimate $\hat{f}_n$ is expressed on the basis of (11) using the Fubini's theorem as

$$\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\int (\hat{f}_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x} = \int (b\hat{f}_n(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x} + \int \sigma^2\hat{f}_n(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{12}$$

The formula consists of two summands which are the integrated versions of the squared bias and variance terms of the $\mathrm{MSE}_{\boldsymbol{x}}(\hat{f}_n)$. The value of the $\mathrm{MISE}(\hat{f}_n)$ depends on the value of the bandwidth h.

Because $\mathrm{MISE}(\hat{f}_n)$ represents the global error of the estimate, one tries to minimize it by localizing the minimizer h^*_{MISE} of (12). Analytical solution to this task is known only in some specific cases, e. g., when the estimated density corresponds to a convex sum of normal densities, see [16] or [18] for exact formulas. To deal with the minimization problem generally, the widely used approach is to investigate the asymptotic behavior of the $\mathrm{MISE}(\hat{f}_n)$ with respect to the sample size $n \in \mathbb{N}$ going to infinity. This is called AMISE analysis and leads to the specification of the asymptotic minimizer h^*_{AMISE}.

However, in Section 1.2.4 of his book [17], Tsybakov provides a deeper criticism of the asymptotic approach. It stems from the fact that the optimality of h^*_{AMISE} is related to a *fixed density* f and not to a well defined class of densities. In Proposition 1.7, Tsybakov shows that for the fixed density f, it is possible to construct such a kernel estimate that the $\mathrm{MISE}(\hat{f}_n)$ diminishes, but this cannot be done uniformly over a sufficiently broad class of densities. Examples of such classes, e. g., Hölder, Sobolev or Nikol'ski classes, are presented in [17]. The Sobolev class is treated in Definition 3.4 below.

Based on this criticism, Tsybakov presents a different approach to the MISE analysis in Section 1.3 of [17]. The approach relies on Fourier analysis.

3.2. Fourier analysis

In the probability theory, Fourier analysis is intimately interconnected with the notion of the characteristic function. Let $\boldsymbol{X} : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be a d-variate real random vector with the joint distribution $\mu(\mathrm{d}\boldsymbol{x})$. The characteristic function $\phi_{\boldsymbol{X}}(\boldsymbol{\omega}) : \mathbb{R}^d \to \mathbb{C}$ of $\boldsymbol{X}$ is defined as the integral transform

$$\phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \mathbb{E}[e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{X}\rangle}] = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}\rangle}\,\mu(\mathrm{d}\boldsymbol{x}), \quad \boldsymbol{\omega} \in \mathbb{R}^d \tag{13}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard dot product in $\mathbb{R}^d$. It is well known that the transform provides the complete characterization of the distribution of $\boldsymbol{X}$; and we often speak about the Fourier transform of the random vector $\boldsymbol{X}$ or the distribution μ.

The other quite common view of the Fourier transform comes from the area of applied mathematics. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable function (a signal in electrical engineering), i. e., let $f \in L_1(\mathbb{R}^d)$, then its Fourier transform is specified as

$$\mathcal{F}[f](\boldsymbol{\omega}) = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}\rangle} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{14}$$

Formula (14) can be treated as the special case of formula (13) when the distribution of $\boldsymbol{X}$ is absolutely continuous with respect to the d-dimensional Lebesgue measure and has the density f, i. e., $\mu(d\boldsymbol{x}) = f(\boldsymbol{x})\,d\boldsymbol{x}$. On the other hand, in (14) f need not be necessarily a density.

Let $f, g \in L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$, i. e., we consider functions both L_1 and L_2 integrable over $\mathbb{R}^d$, then the following properties of the multivariate Fourier transform are relevant to our research:

- continuity: $\mathcal{F}[f]$ is uniformly continuous on $\mathbb{R}^d$
- linearity: $\mathcal{F}[af + bg](\boldsymbol{\omega}) = a\mathcal{F}[f](\boldsymbol{\omega}) + b\mathcal{F}[g](\boldsymbol{\omega}), \quad a, b \in \mathbb{R}$
- shifting: $\mathcal{F}[f(\boldsymbol{x} - \boldsymbol{s})](\boldsymbol{\omega}) = e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{s}\rangle}\mathcal{F}[f](\boldsymbol{\omega}), \quad \boldsymbol{s} \in \mathbb{R}^d$
- scaling: $\mathcal{F}[f(\boldsymbol{x}/h)/h^d](\boldsymbol{\omega}) = \mathcal{F}[f](h\boldsymbol{\omega}), \quad h > 0$
- shifting & scaling: $\mathcal{F}[f((\boldsymbol{x} - \boldsymbol{s})/h)/h^d] = e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{s}\rangle}\mathcal{F}[f](h\boldsymbol{\omega}), \quad \boldsymbol{s} \in \mathbb{R}^d$
- complex conjugate: $\overline{\mathcal{F}[f](\boldsymbol{\omega})} = \mathcal{F}[f](-\boldsymbol{\omega})$
- convolution: $\mathcal{F}[f * g](\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega})\mathcal{F}[g](\boldsymbol{\omega})$
- symmetry: if $f(-\boldsymbol{x}) = f(\boldsymbol{x})$, then $\mathcal{F}[f](-\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega})$
- isometry, due to the Plancheler's formula:

$$\int f^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{1}{(2\pi)^d}\int |\mathcal{F}[f](\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega}.$$

Now, the uniformly weighted sum of Dirac measures $\delta_n(\mathrm{d}\boldsymbol{x})$ introduced in formula (8) represents a probability distribution which does not have a density with respect to the corresponding Lebesgue measure. Its characteristic function is denoted $\phi_n(\boldsymbol{\omega})$ and specified as

$$\phi_n(\boldsymbol{\omega}) = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}\rangle}\delta_n(\mathrm{d}\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{X}_j\rangle}, \quad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{15}$$

Note that $\phi_n(\boldsymbol{\omega})$ constitutes a random variable for every $\boldsymbol{\omega} \in \mathbb{R}^d$ being fixed.

Under the assumption of $L_1(\mathbb{R}^d)$ integrability of the employed kernel K, we can consider the Fourier transform of the kernel density estimate (10). Using the linearity

and the shifting & scaling property of the Fourier transform, $\mathcal{F}[\hat{f}_n](\boldsymbol{\omega})$ is specified by formula

$$\mathcal{F}[\hat{f}_n](\boldsymbol{\omega}) = \frac{1}{n}\sum_{j=1}^{n}\mathcal{F}\left[\frac{1}{h^d}K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_j}{h}\right)\right] = \frac{1}{n}\sum_{j=1}^{n} e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{X}_j\rangle}\mathcal{F}[K](h\boldsymbol{\omega}). \tag{16}$$

Writing $K_{\mathcal{F}}(\boldsymbol{\omega})$ for $\mathcal{F}[K](\boldsymbol{\omega})$, we obtain the compact expression of $\mathcal{F}[\hat{f}_n](\boldsymbol{\omega})$ in the form

$$\mathcal{F}[\hat{f}_n](\boldsymbol{\omega}) = \phi_n(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega}). \tag{17}$$

This shows that the standard kernel estimator, which is based on an i.i.d. sample, is obtained by the convolution of the employed kernel with the uniformly weighted sum of Dirac measures corresponding to the sample.

Let us assume that both density f and kernel K belong also to $L_2(\mathbb{R}^d)$. Then employing the Plancherel's theorem and (17), we get for the MISE of (12) the expression

$$\mathrm{MISE}(\hat{f}_n) = \frac{1}{(2\pi)^d}\,\mathbb{E}\int |\phi_n(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega}) - \phi(\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega}. \tag{18}$$

The next theorem provides the exact $\mathrm{MISE}(\hat{f}_n)$ for any fixed $n \in \mathbb{N}$.

Theorem 3.1. Let $f \in L_2(\mathbb{R}^d)$ be a density and $K \in L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$ a kernel. Then for all $n \geq 1$ and $h > 0$ the MISE of the kernel estimator $\hat{f}_n$ of (10) has the form

$$\begin{aligned}\mathrm{MISE}(\hat{f}_n) \;=\; & \frac{1}{(2\pi)^d}\left[\int |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2|\phi(\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega} + \frac{1}{n}\int |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega}\right] \\ & -\frac{1}{(2\pi)^d}\frac{1}{n}\int |\phi(\boldsymbol{\omega})|^2|K_{\mathcal{F}}(h\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega}.\end{aligned} \tag{19}$$

P r o o f . The proof is just a copy of the original univariate Tsybakov's proof, see [17], p. 22 (generally, we do not need the symmetry of the kernel here). It rests on developing the formula (18) using the facts that $|z|^2 = z\overline{z}$ for $z \in \mathbb{C}$ and $\mathbb{E}[\phi_n(\boldsymbol{\omega})] = \phi(\boldsymbol{\omega})$. □

Now, we are going to discuss the individual terms in the Fourier MISE formula (19). We start with the notion of the *order of a kernel*.

Definition 3.2. Let $\ell \geq 1$ be an integer. We say that the kernel $K\colon \mathbb{R}^d \to \mathbb{R}$ is of order ℓ, if K is in $L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$, its Fourier transform $K_{\mathcal{F}}(\boldsymbol{\omega})$ is real, satisfies $K_{\mathcal{F}}(\mathbf{0}) = 1$ and has continuous all partial derivatives $K^{(m)}_{\mathcal{F},i_1,\ldots i_d} = \partial^m K_{\mathcal{F}}/\partial_{i_1}\ldots\partial_{i_d}$, $m = i_1 + \cdots + i_d$, $m \in \mathbb{N}$, $i_j \in \{0,\ldots,m\}$, $j = 1,\ldots,d$ up to the ℓth order at point $\mathbf{0}$ such that $K^{(m)}_{\mathcal{F},i_1,\ldots i_d}(\mathbf{0}) = 0$ for all $m = 1,\ldots,\ell$.

Remark that the above definition imposes the following conditions on a multivariate kernel to be of order $\ell \geq 1$, $\ell \in \mathbb{N}$:

- $\int K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = 1$,
- $\int u_1^{i_1}\cdots u_d^{i_d}K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = 0$ for $m = 1,\ldots,\ell$.

Indeed, at the origin we have $K_{\mathcal{F}}(\mathbf{0}) = \int e^{\mathrm{i}\langle \mathbf{0},\boldsymbol{u}\rangle} K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \int K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = 1$. For the mth partial derivative, we get $K^{(m)}_{\mathcal{F},i_1,\ldots i_d}(\boldsymbol{\omega}) = \int (\mathrm{i}u_1)^{i_1}\cdots(\mathrm{i}u_d)^{i_d}\, e^{\mathrm{i}\langle \boldsymbol{\omega},\boldsymbol{u}\rangle} K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$, hence $0 = K^{(m)}_{\mathcal{F},i_1,\ldots i_d}(\mathbf{0}) = \mathrm{i}^m \int u_1^{i_1}\cdots u_d^{i_d} K(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$.

As an example, mention that the standard multivariate Gaussian kernel $K(\boldsymbol{u}) = (2/\pi)^{-d/2}\exp(-\frac{1}{2}||\boldsymbol{u}||^2)$ has the Fourier transform $K_{\mathcal{F}}(\boldsymbol{\omega}) = \exp(-\frac{1}{2}||\boldsymbol{\omega}||^2)$ and is of order $\ell = 1$.

3.2.1. The first term

For the first term in the Fourier MISE formula (19), we are able to say something more specific if we consider the order of the kernel involved in the estimate.

Lemma 3.3. Let $K\colon \mathbb{R}^d \to \mathbb{R}$ be a kernel of order $\ell \geq 1$, $\ell \in \mathbb{N}$ and $||\cdot||$ denote the Euclidean norm in $\mathbb{R}^d$. Then there exists a constant $A > 0$ such that

$$\sup_{\boldsymbol{\omega}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{|1 - K_{\mathcal{F}}(\boldsymbol{\omega})|}{||\boldsymbol{\omega}||^\ell} \leq A \tag{20}$$

and

$$\int |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 |\phi(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \leq A^2 h^{2\ell} \int ||\boldsymbol{\omega}||^{2\ell} |\phi(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \tag{21}$$

for any function f with the Fourier transform $\phi(\boldsymbol{\omega})$ and $h > 0$.

P r o o f . We employ the multidimensional Taylor's theorem [1]. Because the kernel K is of order $\ell \geq 1$, its Fourier transform $K_{\mathcal{F}}(\boldsymbol{\omega})$ is real and we have by the Taylor's theorem

$$K_{\mathcal{F}}(\boldsymbol{\omega}) = K_{\mathcal{F}}(\mathbf{0}) + \frac{1}{1!}\sum_{i=1}^{d} K^{(1)}_{\mathcal{F},i}(\mathbf{0})\,\omega_i + \cdots + \frac{1}{\ell!}\sum_{i_1,\ldots,i_d=1}^{d} K^{(\ell)}_{\mathcal{F},i_1,\ldots,i_d}(\mathbf{0})\,\omega_{i_1}\ldots\omega_{i_d} + R_\ell(\boldsymbol{\omega})$$

with $\lim_{\boldsymbol{\omega}\to\mathbf{0}} R_\ell(\boldsymbol{\omega})/||\boldsymbol{\omega}||^\ell = 0$ for the reminder, i. e., $R_\ell(\boldsymbol{\omega}) = o(||\boldsymbol{\omega}||^\ell)$.

As the partial derivatives vanish at origin, the remainder writes $R_\ell(\boldsymbol{\omega}) = K_{\mathcal{F}}(\boldsymbol{\omega}) - K_{\mathcal{F}}(\mathbf{0}) = K_{\mathcal{F}}(\boldsymbol{\omega}) - 1$ and $\lim_{\boldsymbol{\omega}\to\mathbf{0}} |1 - K_{\mathcal{F}}(\boldsymbol{\omega})|/||\boldsymbol{\omega}||^\ell = 0$ by the Taylor's theorem.

Let us define $A_\ell(\boldsymbol{\omega}) = |1 - K_{\mathcal{F}}(\boldsymbol{\omega})|/||\boldsymbol{\omega}||^\ell$ for $\boldsymbol{\omega} \neq \mathbf{0}$, and $A_\ell(\mathbf{0}) = 0$. The function $A_\ell : \mathbb{R}^d \to [0,\infty)$ is continuous on $\mathbb{R}^d$ and attains its maximum on the unit ball $||\boldsymbol{\omega}|| \leq 1$. We denote this maximum by M_1, i. e., $M_1 = \max_{\{\boldsymbol{\omega}:||\boldsymbol{\omega}||\,\leq 1\}}\{A_\ell(\boldsymbol{\omega})\}$. Because $K \in L_1(\mathbb{R}^d)$, we have $0 \leq |K_{\mathcal{F}}(\boldsymbol{\omega})| \leq M_2 < \infty$. Indeed, $|K_{\mathcal{F}}(\boldsymbol{\omega})| \leq \int |e^{\mathrm{i}\langle \boldsymbol{\omega},\boldsymbol{u}\rangle}|\,|K(\boldsymbol{u})|\,\mathrm{d}\boldsymbol{u} \leq \int |K(\boldsymbol{u})|\,\mathrm{d}\boldsymbol{u} = M_2 < \infty$. Therefore, $|1 - K_{\mathcal{F}}(\boldsymbol{\omega})|/||\boldsymbol{\omega}||^\ell \leq 1 + M_2$ for $||\boldsymbol{\omega}|| > 1$. Composing both cases one gets $A_\ell(\boldsymbol{\omega}) \leq \max\{M_1, 1 + M_2\} = A < \infty$ for $\boldsymbol{\omega} \in \mathbb{R}^d$.

The inequality (21) is implied by (20) as follows:

$$\begin{aligned}
\sup_{\boldsymbol{\omega}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|/||h\boldsymbol{\omega}||^\ell &\leq A,\\
|1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 &\leq A^2||h\boldsymbol{\omega}||^{2\ell},\\
\int |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 |\phi(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} &\leq A^2 h^{2\ell} \int ||\boldsymbol{\omega}||^{2\ell} |\phi(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

This concludes the proof. □

The other terms in formula (19) refer to properties of the kernel and density under considerations. We mention only two straightforward observations.

3.2.2. The second term

The second term can be translated from the frequency to the "time" domain using the Plancherel's theorem and the scaling property of the Fourier transform. Change of variables gives the final result:

$$\frac{1}{n}\int |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega} = \frac{(2\pi)^d}{nh^{2d}}\int K^2(\boldsymbol{x}/h)\,\mathrm{d}\boldsymbol{x} = \frac{(2\pi)^d}{nh^d}\int K^2(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}. \tag{22}$$

3.2.3. The third term

The third term is actually the correction term. We have the following inequality for it:

$$\frac{1}{(2\pi)^d}\frac{1}{n}\int |\phi(\boldsymbol{\omega})|^2|K(h\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega} \;\le\; \frac{||K_{\mathcal{F}}||_\infty^2}{n}\int f^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

3.3. The upper bound on the Fourier MISE formula

Concerning an upper bound on the Fourier MISE formula (19), we sum up the above results. First of all, to obtain the upper bound we can omit the correction (the third) term in (19). The second term is solely determined by the properties of the kernel, which is expressed by formula (22). Finally, to obtain a bound on the first term, the properties of the density the data are sampled from and the properties of the kernel have to be matched somehow. To do this we introduce the so-called Sobolev class of densities.

Definition 3.4. Let $\beta \ge 1$ be an integer and $L > 0$ a real. The Sobolev class of densities $\mathcal{P}_{S(\beta,L)}$ consists of all probability density functions $f : \mathbb{R}^d \to [0,\infty)$ satisfying

$$\int ||\boldsymbol{\omega}||^{2\beta}|\phi(\boldsymbol{\omega})|^2\,\mathrm{d}\boldsymbol{\omega} \le (2\pi)^d L^2 \tag{23}$$

where $\phi(\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega})$ and $||\cdot||$ is the Euclidean norm.

The condition (23) is related to the boundedness of partial derivatives of densities in the Sobolev class; e. g., it can be shown that if $\int(\partial f/\partial x_j)^2\,\mathrm{d}\boldsymbol{x} \le L_j < \infty$ for all $j = 1,\ldots,d$, then (23) holds for $\beta = 1$ and $L = ||(L_1,\ldots,L_d)||$. Furthermore, if $f \in \mathcal{P}_S(\beta,L)$, for some $\beta \in \mathbb{N}$ and $L > 0$, then $f \in L_2(\mathbb{R}^d)$.

If $f \in \mathcal{P}_S(\beta,L)$, we say that f is β-Sobolev for given β and L.

Now, the announced matching is provided by fitting the order of the kernel to the Sobolev character of the estimated density. The next theorem, which is the variant of Theorem 1.5 in [17], provides the final result.

Theorem 3.5. Let $n \in \mathbb{N}$ be the number of i.i.d. samples from a distribution with the density $f : \mathbb{R}^d \to [0,\infty)$ which is β-Sobolev for some $\beta \in \mathbb{N}$ and $L > 0$. Let K be a kernel of order β. Assume that the inequality (20) holds for some constant $A > 0$. Fix $\alpha > 0$ and set $h(n) = \alpha n^{-\frac{1}{2\beta+d}}$. Then for any $n \ge 1$ the kernel density estimate $\hat{f}_n$ satisfies

$$\sup_{f\in\mathcal{P}_S(\beta,L)} \mathbb{E}\int (\hat{f}_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2\,\mathrm{d}\boldsymbol{x} \le C\cdot n^{-\frac{2\beta}{2\beta+d}} \tag{24}$$

where $C > 0$ is a constant depending only on α, β, d, A, L and the kernel K.

P r o o f . The proof is a multidimensional version of the original univariate proof presented in [17]. Nevertheless, let us give a sketch of it. By Lemma 3.3 and from the definition of the Sobolev class of densities, one has

$$\int |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 |\phi(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega} \leq (2\pi)^d A^2 h^{2\beta} L^2.$$

Plugging this into the Fourier MISE formula (19) with the correction term omitted, employing $\frac{1}{(2\pi)^d n} \int |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} = \frac{1}{nh^d} \int K^2(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{u}$ and using $h^{2\beta} = \alpha^{2\beta} n^{-\frac{2\beta}{2\beta+d}}$ and $(nh^d)^{-1} = n^{-1}\alpha^{-d} n^{\frac{d}{2\beta+d}} = \alpha^{-d} n^{-\frac{2\beta}{2\beta+d}}$ we get the assertion of the theorem. □

4. PARTICLE FILTER AND KERNEL METHODS

This section presents our own research in the area of combination of the particle filter and kernel methods. The main question here is if the kernel density estimates constructed on the basis of empirical measures approximate the related filtering densities reasonably well. The main obstacle to a direct application of the presented kernel estimation methodology is the fact that the generated empirical measures are not based on i.i.d. samples due to the resampling step of the filter.

Our results are twofold. First, we show that despite the mentioned obstacle the standard kernel density estimates still converge to the related filtering densities. The proof of the assertion is based on Fourier analysis of the convergence result for the particle filter.

The second result concerns a deeper analysis of the obtained convergence formula. The convergence result is based on the assumption on the Sobolev character of the filtering densities. We present a sufficient condition for the persistency of this Sobolev character over time.

4.1. Convergence of kernel density estimates

To start, we recall that the particle filter generates at each time step $t = 1, \ldots, T$, $T \in \mathbb{N}$ the empirical measure $\pi_t^n(\mathrm{d}\boldsymbol{x}_t) = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_t^i}(\mathrm{d}\boldsymbol{x}_t)$. This measure approximates the related filtering distribution π_t that is assumed to have the density $p_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ with respect to the d-dimensional Lebesgue measure, i. e., $\pi_t(\mathrm{d}\boldsymbol{x}_t) = p_t(\boldsymbol{x}_t)\,\mathrm{d}\boldsymbol{x}_t$.

A carrier of the empirical measure π_t^n is the set of particles $\{\boldsymbol{x}_t^i\}_{i=1}^n$, $n \in \mathbb{N}$. This set does not constitute an i.i.d. sample from π_t. If one constructs the standard kernel density estimate on the basis of $\{\boldsymbol{x}_t^i\}_{i=1}^n$ and the selected kernel K, i. e., the estimate

$$\hat{p}_t^n(\boldsymbol{x}_t) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\boldsymbol{x}_t - \boldsymbol{x}_t^i}{h}\right), \tag{25}$$

we ask if $\hat{p}_t^n$ converges in the MISE to the filtering density p_t, provided that the number of particles goes to infinity.

Theorem 4.1. In the filtering problem, let $\{\pi_t\}_{t=0}^T$, $\{p_t\}_{t=0}^T$, $T \in \mathbb{N}$ be the sequences of filtering distributions and corresponding filtering densities. Let p_t, $t \in \{0, 1, \ldots, T\}$ be β-Sobolev for some $\beta \in \mathbb{N}$ and $L_t > 0$, i. e., $p_t \in \mathcal{P}_{S(\beta, L_t)}$. Let $\{\pi_t^n\}_{t=1}^T$, $\{\hat{p}_t^n\}_{t=1}^T$, $n \in \mathbb{N}$

be the sequences of the empirical measures generated by the particle filter and related kernel density estimates (25) with the bandwidth varying as $h(n) = \alpha n^{-\frac{1}{2\beta+d}}$ for some $\alpha > 0$. Let the kernel K employed in the estimates be of order β. Then we have the following upper bound on the MISE of $\hat{p}_t^n$ for $t \in \{1, \ldots, T\}$:

$$\mathbb{E}\left[\int (\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t\right] \leq C_t^2 \cdot n^{-\frac{2\beta}{2\beta+d}} \tag{26}$$

where

$$C_t = AL_t\alpha^\beta + c_t\alpha^{-d/2}||K||. \tag{27}$$

In (27), A is the constant of Lemma 3.3, c_t, $t \in \{1, \ldots, T\}$ are the constants of Theorem 2.3 and $||K||$ is the L_2 norm of the kernel K.

P r o o f . The proof employs the Fourier transform. We start by the assertion of Theorem 2.3:

$$\mathbb{E}[|\pi_t^n f - \pi_t f|^2] \leq \frac{c_t^2||f||_\infty^2}{n} \tag{28}$$

where we replace a general function $f \in B^{\mathbb{C}}(\mathbb{R}^{d_x})$ by the complex exponential on $\mathbb{R}^d$. Note that $d_x = d$.

Let $f(\boldsymbol{x}_t) = e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle}$, then $||f||_\infty = 1$. Denoting ψ_t^n and ψ_t the characteristic functions of π_t^n and π_t, respectively, which are defined according to formula (13), we get from (28)

$$\begin{aligned}
\mathbb{E}[|\psi_t^n(\boldsymbol{\omega}) - \psi_t(\boldsymbol{\omega})|^2] &\leq \frac{c_t^2}{n}, \\
|K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \cdot \mathbb{E}[|\psi_t^n(\boldsymbol{\omega}) - \psi_t(\boldsymbol{\omega})|^2] &\leq |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \cdot \frac{c_t^2}{n}, \\
\mathbb{E}\left[|\psi_t^n(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega}) - \psi_t(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega})|^2\right] &\leq |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \cdot \frac{c_t^2}{n}, \\
\mathbb{E}\left[\int |\psi_t^n(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega}) - \psi_t(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega}\right] &\leq \frac{c_t^2}{n}\int |K_{\mathcal{F}}(h\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega}, \\
\mathbb{E}\left[\int (\hat{p}_t^n(\boldsymbol{x}_t) - p_t^*(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t\right] &\leq \frac{c_t^2}{nh^d}\int K^2(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{u}.
\end{aligned} \tag{29}$$

For any density p_t and its convolution $p_t^* = p_t * (h^{-d}K(\cdot/h))$,

$$\begin{aligned}
\int (p_t^*(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t &= \frac{1}{(2\pi)^d}\int |\psi_t(\boldsymbol{\omega})K_{\mathcal{F}}(h\boldsymbol{\omega}) - \psi_t(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega} \\
&= \frac{1}{(2\pi)^d}\int |1 - K_{\mathcal{F}}(h\boldsymbol{\omega})|^2|\psi_t(\boldsymbol{\omega})|^2 \,\mathrm{d}\boldsymbol{\omega}.
\end{aligned} \tag{30}$$

We assume that the employed kernel has order β and $p_t \in \mathcal{P}_{S(\beta,L_t)}$. Therefore the right-hand side of (30) is bounded according to Lemma 3.3. Further, there is nothing random here and we can apply the expectation with no effect to obtain

$$\mathbb{E}\left[\int (p_t^*(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t\right] \leq A^2h^{2\beta}L_t^2. \tag{31}$$

To proceed, let us consider the product measure $\lambda^d \otimes P$ with the corresponding L_2 norm $||\cdot||_{\lambda^d \otimes P} = [\int\int |\cdot|^2 \mathrm{d}(\lambda^d \otimes P)]^{1/2}$. We have

$$||\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t)||_{\lambda^d \otimes P} \leq Ah^\beta L_t + \frac{c_t}{(nh^d)^{1/2}}||K|| \tag{32}$$

by (29), (31) and the triangle inequality for $||\cdot||_{\lambda^d \otimes P}$.

Let the bandwidth h develop with n as $h(n) = \alpha n^{-\frac{1}{2\beta+d}}$ for some $\alpha > 0$. We have $h^\beta = \alpha^\beta n^{-\frac{\beta}{2\beta+d}}$. Further, $(nh^d)^{-1} = n^{-1}\alpha^{-d}n^{\frac{d}{2\beta+d}} = \alpha^{-d}n^{-\frac{2\beta}{2\beta+d}}$, thus $(nh^d)^{-1/2} = \alpha^{-d/2}n^{-\frac{\beta}{2\beta+d}}$. Inequality (32) then reads as

$$\begin{aligned} ||\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t)||_{\lambda^d \otimes P} &\leq AL_t\alpha^\beta n^{-\frac{\beta}{2\beta+d}} + c_t\alpha^{-d/2}n^{-\frac{\beta}{2\beta+d}}||K|| \\ &\leq (AL_t\alpha^\beta + c_t\alpha^{-d/2}||K||) \cdot n^{-\frac{\beta}{2\beta+d}}. \end{aligned}$$

Squaring to obtain the MISE we get

$$\mathbb{E}\int (\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t \leq (AL_t\alpha^\beta + c_t\alpha^{-d/2}||K||)^2 \cdot n^{-\frac{2\beta}{2\beta+d}}$$

or in the more compact form

$$\mathbb{E}\int (\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t \leq C_t^2 \cdot n^{-\frac{2\beta}{2\beta+d}}$$

where $C_t = AL_t\alpha^\beta + c_t\alpha^{-d/2}||K||$. □

Let us discuss the theorem.

1) First of all, the theorem is proved without any assumption on the i.i.d. character of particles constituting the empirical measures π_t^n. This is the crucial observation, as we know that due to the resampling step the generated particles are not i.i.d.

2) Convergence. For $t \in \mathbb{N}$ fixed, we immediately see from (26) that the MISE of kernel estimates goes to zero as the number of particles increases and the bandwidth decreases accordingly, i. e., $\lim_{n\to\infty} \mathbb{E}\int (\hat{p}_t^n(\boldsymbol{x}_t) - p_t(\boldsymbol{x}_t))^2 \,\mathrm{d}\boldsymbol{x}_t = 0$.

3) Consistency. The theorem proposes that the bandwidth develops with the number of particles n as $h(n) = \alpha n^{-\frac{1}{2\beta+d}}$ for some $\alpha > 0, \beta, d \in \mathbb{N}$. Obviously, $\lim_{n\to\infty} h(n) = 0$, and $\lim_{n\to\infty} nh(n) = \lim_{n\to\infty} \alpha n^{\frac{2\beta+d-1}{2\beta+d}} = \infty$.

4) The dimension matters. We have $n^{-\frac{2\beta}{2\beta+d_1}} < n^{-\frac{2\beta}{2\beta+d_2}}$ for $d_1 < d_2$, and therefore we must increase the number of particles in order to assure a given accuracy as the dimension increases.

5) The order helps. Contrary to the previous result, we have $n^{-\frac{2\beta_1}{2\beta_1+d}} > n^{-\frac{2\beta_2}{2\beta_2+d}}$ for $\beta_1 < \beta_2$. Hence the greater is the order of the employed kernel, the tighter is the bound on the related MISE, in fact, it tends towards n^{-1}. There are techniques available for

constructing kernels of arbitrary orders [17], however, the order of the employed kernel is primarily driven by the Sobolev character of the filtering densities.

6) The theorem assumes that the filtering densities p_t are β-Sobolev for some $L_t > 0$, $t \in \{0, \ldots, T\}$, $T \in \mathbb{N}$ and $\beta \in \mathbb{N}$ being constant over time. It is the question when this assumption holds. In Section 5, we show that the Sobolev character of the filtering densities is retained over time, if a certain condition holds on the transition kernels of the signal process.

7) For $\alpha = 1$, the specification of C_t simplifies to $C_t = AL_t + c_t||K||$ and C_t consists of four terms. Two of them, A and $||K|| = [\int K^2(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}]^{1/2}$ are the constants determined by the employed kernel. The other two, L_t and c_t, develop with time. The L_t term is discussed in Section 5.

8) The c_t constant (with respect to the number of particles) comes from Theorem 2.3. It can be computed recursively as $c_t = c_{t-1}\left(1 + 4||g_t^v||_\infty / \overline{\pi}_t g_t\right), c_0 = 1$. The integral $\overline{\pi}_t g_t$ depends on the values of the observation process and c_t generally develops exponentially with time, see the remark in concluding Section 7.

5. SOBOLEV CHARACTER OF FILTERING DENSITIES

In Theorem 4.1, we have assumed that the filtering densities p_t, $t \in \{0, \ldots, T\}$, $T \in \mathbb{N}$ are β-Sobolev over time. This assumption can be verified for p_0, but for other time instants $t > 0$ a direct verification is typically impossible. That is why we are interested in a practical tool for performing the verification indirectly so that the assumptions for the convergence result of Theorem 4.1 were fulfilled. As a result, we present a sufficient condition on the densities of transition kernels of the signal process such that the Sobolev character of the filtering densities is retained over time.

In the statement below, we work with the prediction and update formulas, (6) and (7), respectively, of Section 2.3. We rewrite these formulas in the more compact form using the following shortcuts: $\overline{p}_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$, $p_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ (in fact, this shortcut was already used in Theorem 4.1) and $g_t(\boldsymbol{x}_t) = g_t(\boldsymbol{y}_t|\boldsymbol{x}_t) = g_t^v(\boldsymbol{y}_t - h_t(\boldsymbol{x}_t))$ for the respective densities; and $\overline{\pi}_t g_t = \int g_t(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})\,\mathrm{d}\boldsymbol{x}_t = \int g_t(\boldsymbol{x}_t)\overline{p}_t(\boldsymbol{x}_t)\,\mathrm{d}\boldsymbol{x}_t$ for the normalizing integrals. As we assume that g_t^v are bounded and strictly positive we have $\overline{\pi}_t g_t$ finite and $\overline{\pi}_t g_t > 0$. Using the introduced shortcuts, (6) and (7) write as

$$\overline{p}_t(\boldsymbol{x}_t) = \int K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) p_{t-1}(\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}, \tag{33}$$

$$p_t(\boldsymbol{x}_t) = \frac{g_t(\boldsymbol{x}_t)\overline{p}_t(\boldsymbol{x}_t)}{\overline{\pi}_t g_t}. \tag{34}$$

Let us recall explicitly, that we assume that the transition kernel has a density, i. e., $K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_t$. With a slight abuse of notation we use the same symbol K_{t-1} for denoting the kernel, the corresponding conditional measure and its density. However, the density is always indicated as a function including its argument $\boldsymbol{x}_t$, i. e., as $K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. The conditional distribution induced by the kernel K_{t-1} is then denoted as $K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$.

Definition 5.1. Let K_{t-1} be the transition kernel in the filtering problem for time $t-1$, $t-1 \in \mathbb{N}_0$. As the conditional characteristic function $\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1})$ of the transition kernel K_{t-1} we denote the characteristic function of the conditional distribution induced by this kernel, i. e.,

$$\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1}) = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle} K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_t = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle} K_{t-1}(\mathrm{d}\boldsymbol{x}_t|\boldsymbol{x}_{t-1}).$$

Theorem 5.2. In the filtering problem, let $p_0 \in \mathcal{P}_{S(\beta,L_0)}$. Let $\{K_{t-1},\ t \in \mathbb{N}\}$ be the set of the transition kernels and $\{\mathcal{F}[K_{t-1}],\ t \in \mathbb{N}\}$ be the set of the corresponding conditional characteristic functions. For all $t \in \mathbb{N}$, let $\mathcal{F}[K_{t-1}]$ be bounded by a function $K_b\colon \mathbb{R}^d \to \mathbb{C}$ in such a way that for any $\boldsymbol{x}_{t-1} \in \mathbb{R}^d$ and $\boldsymbol{\omega} \in \mathbb{R}^d$

$$|\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1})| \le |K_b(\boldsymbol{\omega})|. \tag{35}$$

Let the function K_b satisfy (23) for some $\beta \in \mathbb{N}$ and $L_{K_b} > 0$. Then the filtering densities p_t are β-Sobolev for all $t \in \mathbb{N}$, i. e., $p_t \in \mathcal{P}_{S(\beta,L_t)}$, with the recurrence for L_t written as

$$L_t = ||g_t^v||_\infty L_{K_b}/\overline{\pi}_t g_t \tag{36}$$

where $||g_t^v||_\infty = \sup_{\boldsymbol{u}}\{|g_t^v(\boldsymbol{u})|\}$.

P r o o f. The theorem holds for p_0 by the assumption. Let $t \in \mathbb{N}$, then multiplying both sides of (33) by the complex exponential we get from the prediction formula

$$e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle}\,\overline{p}_t(\boldsymbol{x}_t) = e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle}\int K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p_{t-1}(\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}.$$

By integration, the left-hand side gives the characteristic function $\overline{\psi}_t(\boldsymbol{\omega})$ of $\overline{p}_t(\boldsymbol{x}_t)$, i. e.,

$$\overline{\psi}_t(\boldsymbol{\omega}) = \int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle}\overline{p}_t(\boldsymbol{x}_t)\,\mathrm{d}\boldsymbol{x}_t.$$

The right-hand side has then form

$$\begin{aligned}
&\int\int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle} K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p_{t-1}(\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}\,\mathrm{d}\boldsymbol{x}_t\\
&= \int p_{t-1}(\boldsymbol{x}_{t-1})\left(\int e^{\mathrm{i}\langle\boldsymbol{\omega},\boldsymbol{x}_t\rangle} K_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_t\right)\mathrm{d}\boldsymbol{x}_{t-1},\\
&= \int p_{t-1}(\boldsymbol{x}_{t-1})\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}.
\end{aligned}$$

The equality of two complex numbers is equivalent to the equality of their complex conjugates. Hence we can multiply both sides by their complex conjugates with the equality retained. This gives us the expression

$$|\overline{\psi}_t(\boldsymbol{\omega})|^2 = \left|\int p_{t-1}(\boldsymbol{x}_{t-1})\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1})\,\mathrm{d}\boldsymbol{x}_{t-1}\right|^2.$$

Now, by the Jensen's inequality for the absolute value and assumed boundedness of $\mathcal{F}[K_{t-1}]$, we have

$$\begin{aligned} |\overline{\psi}_t(\boldsymbol{\omega})|^2 &\leq \left(\int |\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1})|\, p_{t-1}(\boldsymbol{x}_{t-1})\, \mathrm{d}\boldsymbol{x}_{t-1} \right)^2 \\ &\leq \left(|K_b(\boldsymbol{\omega})| \int p_{t-1}(\boldsymbol{x}_{t-1})\, \mathrm{d}\boldsymbol{x}_{t-1} \right)^2 = |K_b(\boldsymbol{\omega})|^2. \end{aligned}$$

Thus,

$$\int ||\boldsymbol{\omega}||^{2\beta} |\overline{\psi}_t(\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega} \leq \int ||\boldsymbol{\omega}||^{2\beta} |K_b(\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega} \leq (2\pi)^d L_{K_b}^2. \tag{37}$$

The above formula shows that $\overline{p}_t \in \mathcal{P}_{(\beta, L_{K_b})}$ for any $t \in \mathbb{N}$. We proceed with the specification of the Sobolev constant L_t of the update (filtering) density p_t.

In Section 2.2 in formula (3), there was stated that the function $g_t(\boldsymbol{x}_t)$ of the update formula (34) has the form $g_t(\boldsymbol{x}_t) = g_t^v(\boldsymbol{y}_t - h(\boldsymbol{x}_t))$. Function g_t^v is the density of the noise term in the observation process and is assumed to be bounded and strictly positive. Thus, regardless of the form of h_t, we have $\sup_{\boldsymbol{x}_t, \boldsymbol{y}_t} \{|g_t^v(\boldsymbol{y}_t - h(\boldsymbol{x}_t))|\} = ||g_t^v||_\infty < \infty$ and $0 < \overline{\pi}_t g_t < \infty$.

Again, multiplying the update formula (34) by the complex exponential, integrating and multiplying by the respective conjugates gives us

$$\begin{aligned} (\overline{\pi}_t g_t) \int e^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}_t \rangle} p_t(\boldsymbol{x}_t)\, \mathrm{d}\boldsymbol{x}_t &= \int e^{\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x}_t \rangle} g_t(\boldsymbol{x}_t)\, \overline{p}_t(\boldsymbol{x}_t)\, \mathrm{d}\boldsymbol{x}_t, \\ ||\boldsymbol{\omega}||^{2\beta} |\psi_t(\boldsymbol{\omega})|^2 &\leq \frac{||g_t^v||_\infty^2}{(\overline{\pi}_t g_t)^2} \, ||\boldsymbol{\omega}||^{2\beta} |\overline{\psi}_t(\boldsymbol{\omega})|^2, \\ (2\pi)^{-d} \int ||\boldsymbol{\omega}||^{2\beta} |\psi_t(\boldsymbol{\omega})|^2 \, \mathrm{d}\boldsymbol{\omega} &\leq \frac{||g_t^v||_\infty^2 L_{K_b}^2}{(\overline{\pi}_t g_t)^2} = L_t^2, \end{aligned}$$

which concludes the proof. □

The theorem tells us that, in the particle filter, the β-Sobolev character of the filtering densities is retained over time if the set of the conditional characteristic functions of transition kernels $\{\mathcal{F}[K_{t-1}](\boldsymbol{\omega}|\boldsymbol{x}_{t-1}),\, t \in \mathbb{N}\}$ is uniformly bounded.

6. DISCUSSION

When finalizing our paper we became aware about publishing the paper [3] (available also at *arXiv:1111.5866*). This paper is particularly important for us as its topic significantly overlaps with the one presented herein. Given this fact, it is worth to discuss explicitly the differences between the results of the two works.

The results of [3] in Section 4 split into two groups. The first comprises a.s. convergence results – Theorems CM-4.1, CM-4.2, CM-4.3 and CM-4.5; all drawing on Lemma CM-4.1[1]. The second group comprises the results for integrated versions (w.r.t.

[1]Coincidentally, the concerned theorems of [3] happen to be numbered as 4.x, which is also the case of the main theorems of this paper. To make a clear distinction between them, we denote the theorems of [3] as CM-4.x and the theorems of this paper as DC-4.x and DC-5.x.

the probability and Lebesgue measures) – Theorems CM-4.4 and CM-4.6. Note that Theorem CM-4.5 falls into the first group as integration is provided only w.r.t. the Lebesgue measure. Results of Section 5 are aimed at applications and will not be discussed here.

Our theorems relates mainly to the second group. In fact, we have presented a stronger version of Theorem CM-4.4 due to the diferent assumption on the estimated density: the Sobolev character instead of the Lipschitz continuity. Let us discuss this in more details.

First of all, Theorems CM-4.4 and CM-4.6 are restricted to densities and kernels with compact supports, thus, for example, they do not cover the basic Gaussian case. It is clear that the constants $c_{\alpha,\mathcal{K},t}$ and $c_{\mathcal{K},t}$ grow to infinity as the volume of $\mathcal{K}$ does. The reason for introducing the compact support requirement is that (4.13) of [3] cannot be simply integrated w.r.t. Lebesgue measure on $\mathbb{R}^d$ as the right-hand side would turn to an uninformative unlimited upper bound. In our Theorem DC-4.1, we are not restricted by these limitations. In our approach, the behavior of p_t with respect to integration over $\mathbb{R}^d$ is induced by the requirement on its Sobolev character. Similarly, this is also the case for the employed convolution kernel when its behavior is determined by its order.

Further, for $\beta = 1$, the bound in our Theorem DC-4.1 is tighter than that of Theorem CM-4.4 and equals the one presented in Theorem CM-4.6. Indeed, in Theorem CM-4.6 it is required that the filtering density has bounded partial derivatives up to order 2, which implies that the density is 1-Sobolev. The bound in Theorem CM-4.6 writes $b_2 = n^{-4/2(d_x+2)}$; see the discussion in paragraph 4.4 of [3] for transforming k to the number of particles n. Our bound for $\beta = 1$ then writes $b_1 = n^{-2/(2+d)}$, so $b_2 = b_1$ (in both cases the constants are omitted).

Our Theorem DC-5.1 corresponds to Remark 3.4 of [3]. The difference is that we are more specific. In Remark 3.4, it is required that $g_t^{y_t}$ (g_t^v in our notation) is bounded similarly as in our case, but we do not have any requirement on derivatives of g_t^v. Speaking about the transition kernels, our requirement is that they are uniformly bounded by a Sobolev function of the same order as the convolution kernel.

To sum up, due to our assumptions we are able to obtain stronger results for MISE in terms of a general integration domain. Moreover, in [3] the transition to a.s. versions comes from the integrated versions via Lemma CM-4.1. Thus, using this lemma we might obtain the a.s. version for ISE (the counterpart of Theorem CM-4.5) without further restrictions on filtering densities and the used convolution kernel.

7. CONCLUSION

In the paper, we have demonstrated that the standard methodology of kernel density estimation can be successfully applied in the area of particle filtering. We have proved that the kernel density estimates, which are constructed on the basis of particles generated by the particle filter, converge in the MISE to the theoretical filtering density at each time instant of the operation of the filter. In fact, we presented an upper bound on the MISE convergence rate which then implies the result, provided that the number of particles goes to infinity. Moreover, we have stated the sufficient condition for retaining the Sobolev character of the filtering densities over time.

In Theorem 2.3, the constant c_t is known that it typically grows exponentially with time, see e. g., [4] p. 87, therefore C_t of (27) does so; and, if one wants to assure a given accuracy of the filtering density approximation, then the number of generated particles must increase exponentially too. This is an unpleasant property of the particle filter. On the other hand, there are results available, e. g., [12] or [8], that under additional conditions, uniformly convergent particle filters can be constructed, which means that c_t of (9) is bounded over time.

The constant C_t depends on L_t. Under the conditions of Theorem 5.2, we know how L_t develops with time. In fact, this development is similar to the development of c_t constant and we face again the risk of an exponential growth of L_t. The study of the conditions when L_t evolves uniformly over time is an issue for the future research in this field.

Finally, let us add that in the Fourier domain, the presented convergence result can be straightforwardly extended on the convergence of partial derivatives of the kernel density estimates to the related partial derivatives of the filtering densities.

ACKNOWLEDGEMENTS

This work was supported by grant GA16-03708S of Czech Science Foundation, institutional support RVO:67985807 and grant SVV-2016-260334 of Charles University in Prague.

(Received September 15, 2015)

REFERENCES

[1] J. Brabec and B. Hrůza: Matematická analýza II (Mathematical Analysis II, in Czech). SNTL/ALFA, 1986.

[2] D. Crisan and A. Doucet: A survey of convergence results on particle filtering methods for practitioners. IEEE Trans. Signal Processing *50* (2002), 3, 736–746. DOI:10.1109/78.984773

[3] D. Crisan and J. Míguez: Particle-kernel estimation of the filter density in state-space models. Bernoulli *20* (2014), 4, 1879–1929. DOI:10.3150/13-bej545

[4] A. Doucet, N. de Freitas, and N. Gordon (Eds.): Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York 2001. DOI:10.1007/978-1-4757-3437-9

[5] A. Doucet and A. M. Johansen: A tutorial on particle filtering and smoothing: fifteen years later. In: The Oxford Handbook of Nonlinear Filtering (D. Crisan and B. Rozovskii, eds.), Oxford University Press, 2011.

[6] B. Fristedt, N. Jain, and N. Krylov: Filtering and Prediction: A Primer. American Mathematical Society, 2007. DOI:10.1090/stml/038

[7] G. H. Givens: Consistency of the local kernel density estimator. Statist. Probab. Lett. *25* (1995), 55–61. DOI:10.1016/0167-7152(94)00205-m

[8] K. Heine and D. Crisan: Uniform approximations of discrete-time filters. Adv. Appl. Probab. *40* (2008), 4, 979–1001. DOI:10.1239/aap/1231340161

[9] M. Hürzeler and H. R. Künsch: Monte Carlo approximations for general state-space models. J. Comput. Graph. Statist. *7* (1998), 2, 175–193. DOI:10.2307/1390812

[10] H. R. Künsch: Recursive Monte Carlo filters: Algorithms and theoretical bounds. Ann. Statist. *33* (2005), 5, 1983–2021. DOI:10.1214/009053605000000426

[11] F. Le Gland and N. Oudjane: Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. Ann. Appl. Probab. *14* (2004), 1, 144–187. DOI:10.1214/aoap/1075828050

[12] P. Del Moral and A. Guionnet: On the stability of interacting processes with applications to filtering and genetic algorithms. Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques *37* (2001), 2, 155–194. DOI:10.1016/s0246-0203(00)01064-5

[13] C. Musso, N. Oudjane, and F. Le Gland: Improving regularised particle filters. In: Sequential Monte Carlo Methods in Practice (A. Doucet, N. Freitas, and N. Gordon, eds.), Chapter 12, Springer 2001, pp. 247–272. DOI:10.1007/978-1-4757-3437-9_12

[14] E. Parzen: On estimation of a probability density function and mode. Ann. Math. Statist. *33* (1962), 3, 1065–1076. DOI:10.1214/aoms/1177704472

[15] S. Särkkä: Bayesian Filtering and Smoothing. Cambridge University Press, 2013.

[16] B. W. Silverman: Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, London, New York 1986. DOI:10.1007/978-1-4899-3324-9

[17] A. B. Tsybakov: Introduction to Nonparametric Estimation. Springer, 2009. DOI:10.1007/b13794

[18] M. P. Wand and M. C. Jones: Kernel Smoothing. Chapman and Hall/CRC, London, New York 1995. DOI:10.1007/978-1-4899-4493-1

David Coufal, The Czech Academy of Sciences, Institute of Computer Science AS CR, Pod Vodárenskou věží 2, 182 07 Praha 8, and Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8. Czech Republic.

e-mail: david.coufal@cs.cas.cz