Kernel Density Estimation for Dynamical Systems

Hanyuan Hang HANYUAN.HANG@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Ingo Steinwart INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE
Institute for Stochastics and Applications
University of Stuttgart
70569 Stuttgart, Germany

Yunlong Feng YUNLONG.FENG@ESAT.KULEUVEN.BE
Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Abstract

We study the density estimation problem with observations generated by certain dynamical systems that admit a unique underlying invariant Lebesgue density. Observations drawn from dynamical systems are not independent and moreover, usual mixing concepts may not be appropriate for measuring the dependence among these observations. By employing the $\mathcal{C}$-mixing concept to measure the dependence, we conduct statistical analysis on the consistency and convergence of the kernel density estimator. Our main results are as follows: First, we show that with properly chosen bandwidth, the kernel density estimator is universally consistent under L_1-norm; Second, we establish convergence rates for the estimator with respect to several classes of dynamical systems under L_1-norm. In the analysis, the density function f is only assumed to be Hölder continuous which is a weak assumption in the literature of nonparametric density estimation and also more realistic in the dynamical system context. Last but not least, we prove that the same convergence rates of the estimator under L_∞-norm and L_1-norm can be achieved when the density function is Hölder continuous, compactly supported, and bounded. The bandwidth selection problem of the kernel density estimator for dynamical system is also discussed in our study via numerical simulations.

Keywords: Kernel density estimation, dynamical system, dependent observations, $\mathcal{C}$-mixing process, universal consistency, convergence rates, covering number, learning theory

1. Introduction

Dynamical systems are now ubiquitous and are vital in modeling complex systems, especially when they admit recurrence relations. Statistical inference for dynamical systems has drawn continuous attention across various fields, the topics of which include parameter estimation, invariant measure estimation, forecasting, noise detection, among others. For instance, in the statistics and machine learning community, the statistical inference for

© 2016 Hanyuan Hang, Ingo Steinwart, Yunlong Feng and Johan A.K. Suykens.

certain dynamical systems have been recently studied in Suykens et al. (1995); Suykens and Vandewalle (2000); Suykens et al. (2002); Zoeter and Heskes (2005); Anghel and Steinwart (2007); Steinwart and Anghel (2009); Deisenroth and Mohamed (2012); McGoff et al. (2015a); Hang and Steinwart (2016), just to name a few. We refer the reader to a recent survey in McGoff et al. (2015b) for a general depiction of this topic. The purpose of this study is to investigate the density estimation problem for dynamical systems via a classical nonparametric approach, i.e., kernel density estimation.

The commonly considered density estimation problem can be stated as follows. Let x_1, x_2, $\ldots, x_n$ be observations drawn independently from an unknown distribution P on $\mathbb{R}^d$ with the density f. Density estimation is concerned with the estimation of the underlying density f. Accurate estimation of the density is important for many machine learning tasks such as regression, classification, and clustering problems and also plays an important role in many real-world applications. Nonparametric density estimators are popular since weaker assumptions are applied to the underlying probability distribution. Typical nonparametric density estimators include the histogram and kernel density estimator. In this study, we are interested in the latter one, namely, *kernel density estimator*, which is also termed as *Parzen-Rosenblatt estimator* (Parzen, 1962; Rosenblatt, 1956) and takes the following form

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1}$$

Here, $h := h_n > 0$ is a bandwidth parameter and K is a smoothing kernel. In the literature, point-wise and uniform consistency and convergence rates of the estimator f_n to the unknown truth density f under various distance measurements, e.g., L_1, L_2, and L_∞, have been established by resorting to the regularity assumptions on the smoothing kernel $\mathcal{K}$, the density f, and the decay of the bandwidth sequence $\{h_n\}$. Besides the theoretical concerns on the consistency and convergence rates, another practical issue one usually needs to address is the choice of the bandwidth parameter h_n, which is also called the *smoothing parameter*. It plays a crucial role in the bias-variance trade-off in kernel density estimation. In the literature, approaches to choosing the smoothing parameter include least-squares cross-validation (Bowman, 1984; Rudemo, 1982), biased cross-validation (Scott and Terrell, 1987), plug-in method (Park and Marron, 1990; Sheather and Jones, 1991), the double kernel method (Devroye, 1989), and also the method based on a discrepancy principle (Eggermont and LaRiccia, 2001). We refer the reader to Jones et al. (1996a) for a general overview and to Wand and Jones (1994); Cao et al. (1994); Jones et al. (1996b); Devroye (1997) for more detailed reviews.

Note that studies on the kernel density estimator (1) mentioned above heavily rely on the assumption that the observations are drawn in an i.i.d fashion. In the literature of statistics and machine learning, it is commonly accepted that the i.i.d assumption on the given data can be very much restrictive in real-world applications. Having realized this, researchers turn to weaken this i.i.d assumption by assuming that the observations are weakly dependent under various notions of weakly dependence which include α-mixing, β-mixing, and ϕ-mixing (Bradley, 2005). There has been a flurry of work to attack this problem with theoretical and practical concerns, see e.g., Györfi (1981); Masry (1983, 1986); Robinson (1983); Masry and Györfi (1987); Tran (1989b); Györfi and Masry (1990); Tran (1989a); Hart and Vieu (1990); Yu (1993) and Hall et al. (1995), under the above notions of

dependence. These studies were conducted under various notions of sample dependence. In fact, as Györfi and Lugosi (1992) pointed out, for samples that are ergodic, kernel density estimation is not universally consistent under the usual conditions. A counter example was devised there showing the existence of an ergodic sequence of uniformly distributed random variables based on which the kernel density estimation almost surely does not tend to zero in the L_1 sense. On the other hand, the assumed correlation among the observations complicates the kernel density estimation problem from a technical as well as practical view and also brings inherent barriers. This is because, more frequently, the analysis on the consistency and convergence rates of the kernel density estimator (1) is proceeded by decomposing the error term into bias and variance terms, which correspond to data-free and data-dependent error terms, respectively. The data-free error term can be tackled by using techniques from the approximation theory while the data-dependent error term is usually dealt with by exploiting arguments from the empirical process theory such as concentration inequalities. As a result, due to the existence of dependence among observations and various notions of the dependence measurement, the techniques, and results concerning the data-dependent error term are in general not universally applicable. On the other hand, it has been also pointed out that the bandwidth selection in kernel density estimation under dependence also departures from the independent case, see e.g., Hart and Vieu (1990); Hall et al. (1995).

In fact, when the observations $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ are generated by certain ergodic measure-preserving dynamical systems, the problem of kernel density estimation can be even more involved. To explain, let us consider a discrete-time ergodic measure-preserving dynamical system described by the sequence $(T^n)_{n\geq 1}$ of iterates of an unknown map $T : \Omega \to \Omega$ with $\Omega \subset \mathbb{R}^d$ and a unique invariant measure P which possesses a density f with respect to the Lebesgue measure (rigorous definitions will be given in the sequel). That is, we have

$$x_i = T^i(x_0), \quad i = 1, 2, \ldots, n, \tag{2}$$

where x_0 is the initial state. It is noticed that in this case the usual mixing concepts are not general enough to characterize the dependence among observations generated by (2) (Maume-Deschamps, 2006; Hang and Steinwart, 2016). On the other hand, existing theoretical studies on the consistency and convergence rates of the kernel density estimator for i.i.d. observations frequently assume that the density function f is sufficiently smooth, e.g., first-order or even second-order smoothness. However, more often than not, this requirement can be stringent in the dynamical system context. For instance, the Lasota-Yorke map (Lasota and Yorke, 1973) admits a density f which only belongs to the space BV, i.e., functions of bounded variation. This is also the case for the β-map in Example 3 (see Subsection 2.2). Therefore, studies on kernel density estimation mentioned above with dependent observations, in general, may not be applicable.

In this study, the kernel density estimation problem with observations generated by dynamical systems (2) is approached by making use of a more general concept for measuring the dependence of observations, namely, the so-called $\mathcal{C}$-mixing process (refer to Section 2 for the definition). Proposed in Maume-Deschamps (2006) and recently investigated in Hang and Steinwart (2016) and Hang et al. (2016), the $\mathcal{C}$-mixing concept is shown to be more general and powerful in measuring dependence among observations generated by dynamical

systems and can accommodate a large class of dynamical systems. Recently, a Bernstein-type exponential inequality for $\mathcal{C}$-mixing processes was established in Hang and Steinwart (2016) and its applications to some learning schemes were explored in Hang and Steinwart (2016) and Hang et al. (2016).

Our main purpose in this paper is to conduct some theoretical analysis and practical implementations on the kernel density estimator for dynamical systems. The primary concern is the consistency and convergence rates of the kernel density estimator (1) with observations generated by dynamical systems (2). The consistency and convergence analysis is conducted under L_1-norm, and L_∞-norm, respectively. We show that under mild assumptions on the smoothing kernel, with properly chosen bandwidth, the estimator is universally consistent under L_1-norm. When the probability distribution P possesses a polynomial or exponential decay outside of a radius-r ball in its support, under the Hölder continuity assumptions on the kernel function and the density, we obtain almost optimal convergence rates under L_1-norm. Moreover, when the probability distribution P is compactly supported, which is a frequently encountered setting in the dynamical system context, we prove that stronger convergence results of the estimator can be developed, i.e., convergence results under L_∞-norm which are shown to be of the same order with its L_1-norm convergence rates. Finally, with regard to the practical implementation of the estimator, we also discuss the bandwidth selection problem by performing numerical comparisons among several typical existing selectors that include least squares cross-validation and its variants for dependent observations, and the double kernel method. We show that the double kernel bandwidth selector proposed in Devroye (1989) can in general work well. Moreover, according to our numerical experiments, we find that bandwidth selection for kernel density estimator of dynamical systems is usually ad-hoc in the sense that its performance may depend on the considered dynamical system.

The rest of this paper is organized as follows. Section 2 is a warm-up section for the introduction of some notations, definitions and assumptions that are related to the kernel density estimation problem and dynamical systems. Section 3 is concerned with the consistency and convergence of the kernel density estimator and presents the main theoretical results of this study. We discuss the bandwidth selection problem in Section 4. All the proofs of Section 3 can be found in Section 5. We end this paper in Section 6.

2. Preliminaries

2.1 Notations

Throughout this paper, λ^d is denoted as the Lebesgue measure on $\mathbb{R}^d$ and $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$. We denote B_r as the centered ball of $\mathbb{R}^d$ with radius r, that is,

$$B_r := \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : \|x\| \leq r\},$$

and its complement H_r as

$$H_r := \mathbb{R}^d \backslash B_r = \{x \in \mathbb{R}^d : \|x\| > r\}.$$

Recall that for $1 \leq p < \infty$, the ℓ_p^d-norm is defined as $\|x\|_{\ell_p^d} := (x_1^p + \cdots + x_d^p)^{1/p}$, and the ℓ_∞^d-norm is defined as $\|x\|_{\ell_\infty^d} := \max_{i=1,\ldots,d} |x_i|$. Let $(\Omega, \mathcal{A}, \mu)$ be a probability space. We

denote $L_p(\mu)$ as the space of (equivalence classes of) measurable functions $g : \Omega \to \mathbb{R}$ with finite L_p-norm $\|g\|_p$. Then $L_p(\mu)$ together with $\|g\|_p$ forms a Banach space. Moreover, if $\mathcal{A}' \subset \mathcal{A}$ is a sub-σ-algebra, then $L_p(\mathcal{A}', \mu)$ denotes the space of all $\mathcal{A}'$-measurable functions $g \in L_p(\mu)$. Finally, for a Banach space E, we write B_E for its closed unit ball.

In what follows, the notation $a_n \lesssim b_n$ means that there exists a positive constant c such that $a_n \leq c\, b_n$, for all $n \in \mathbb{N}$. With a slight abuse of notation, in this paper, c, c' and C are used interchangeably for positive constants while their values may vary across different lemmas, propositions, theorems, and corollaries.

2.2 Dynamical Systems and $\mathcal{C}$-mixing Processes

In this subsection, we first introduce the dynamical systems of interest, namely, ergodic measure-preserving dynamical systems. Mathematically, an *ergodic measure-preserving dynamical system* is a system $(\Omega, \mathcal{A}, \mu, T)$ with a mapping $T : \Omega \to \Omega$ that is measure-preserving, i.e., $\mu(A) = \mu(T^{-1}A)$ for all $A \in \mathcal{A}$, and ergodic, i.e., $T^{-1}A = A$ implies $\mu(A) = 0$ or 1. In this study, we are confined to the dynamical systems in which Ω is a subset of $\mathbb{R}^d$, μ is a probability measure that is absolutely continuous with respect to the Lebesgue measure λ and admits a unique invariant Lebesgue density f.

In our study, it is assumed that the observations $x_1, x_2, \cdots, x_n$ are generated by the discrete-time dynamical system (2). Below we list several typical examples of discrete-time dynamical systems that satisfy the above assumptions (Lasota and Mackey, 1985):

Example 1 (Logistic Map) *The Logistic map is defined by*

$$T(x) = \theta x(1-x), \;\; x \in (0,1), \;\; \theta \in [0,4],$$

with a unique invariant Lebesgue density

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \;\; 0 < x < 1.$$

Example 2 (Gauss Map) *The Gauss map is defined by*

$$T(x) = \frac{1}{x} \mod 1, \;\; x \in (0,1),$$

with a unique invariant Lebesgue density

$$f(x) = \frac{1}{\log 2} \cdot \frac{1}{1+x}, \;\; x \in (0,1).$$

Example 3 (β-Map) *For $\beta > 1$, the β-map is defined as*

$$T(x) = \beta x \mod 1, \;\; x \in (0,1),$$

with a unique invariant Lebesgue density given by

$$f(x) = c_\beta \sum_{i \geq 0} \beta^{-(i+1)} \mathbf{1}_{[0, T^i(1)]}(x),$$

where c_β is a constant chosen such that f has integral 1.

We now introduce the notion for measuring the dependence among observations from dynamical systems, namely, $\mathcal{C}$-mixing process (Maume-Deschamps, 2006; Hang and Steinwart, 2016). To this end, let us assume that $(X, \mathcal{B})$ is a measurable space with $X \subset \mathbb{R}^d$. Let $\mathcal{X} := (X_n)_{n\geq 1}$ be an X-valued stochastic process on $(\Omega, \mathcal{A}, \mu)$, and for $1 \leq i \leq j \leq \infty$, denote by $\mathcal{A}_i^j$ the σ-algebra generated by $(X_i, \ldots, X_j)$. Let $\Gamma : \Omega \to X$ be a measurable map. μ_Γ is denoted as the Γ-image measure of μ, which is defined as $\mu_\Gamma(B) := \mu(\Gamma^{-1}(B))$, $B \subset X$ measurable. The process $\mathcal{X}$ is called *stationary* if $\mu_{(X_{i_1+j},\ldots,X_{i_n+j})} = \mu_{(X_{i_1},\ldots,X_{i_n})}$ for all $n, j, i_1, \ldots, i_n \geq 1$. Denote $P := \mu_{X_1}$. Moreover, for $\psi, \varphi \in L_1(\mu)$ satisfying $\psi\varphi \in L_1(\mu)$, we denote the correlation of ψ and φ by

$$\operatorname{cor}(\psi, \varphi) := \int_\Omega \psi\, \varphi \, \mathrm{d}\mu - \int_\Omega \psi \, \mathrm{d}\mu \cdot \int_\Omega \varphi \, \mathrm{d}\mu \,.$$

It is shown that several dependency coefficients for $\mathcal{X}$ can be expressed in terms of such correlations for restricted sets of functions ψ and φ (Hang and Steinwart, 2016). In order to introduce the notion, we also need to define a new norm, which is taken from Maume-Deschamps (2006) and introduces restrictions on ψ and φ considered here. Let us assume that $\mathcal{C}(X)$ is a subspace of bounded measurable functions $g : X \to \mathbb{R}$ and that we have a semi-norm $|||\cdot|||$ on $\mathcal{C}(X)$. For $g \in \mathcal{C}(X)$, we define the $\mathcal{C}$-norm $\|\cdot\|_\mathcal{C}$ by

$$\|g\|_\mathcal{C} := \|g\|_\infty + |||g|||. \tag{3}$$

Additionally, we need to introduce the following restrictions on the semi-norm $|||\cdot|||$.

Assumption 1 *We assume that the following two restrictions on the semi-norm $|||\cdot|||$ hold:*

(i) $|||g||| = 0$ *for all constant functions* $g \in \mathcal{C}(X)$*;*

(ii) $|||e^g||| \leq \left\|e^g\right\|_\infty |||g|||, \qquad g \in \mathcal{C}(X).$

Note that the first constraint on the semi-norm in Assumption 1 implies its shift invariance on $\mathbb{R}$ while the inequality constraint can be viewed as an abstract *chain rule* if one views the semi-norm as a norm describing aspects of the smoothness of g, as discussed in Hang and Steinwart (2016). In fact, it is easy to show that the following function classes, which are probably also the most frequently considered in the dynamical system context, satisfy Condition (i) in Assumption 1. Moreover, they also satisfy Condition (ii) in Assumption 1, as shown in (Hang and Steinwart, 2016):

- $L_\infty(X)$: The class of bounded functions on X;
- $BV(X)$: The class of bounded variation functions on X;
- $C_{b,\alpha}(X)$: The class of bounded and α-Hölder continuous functions on X;
- $\mathrm{Lip}(X)$: The class of Lipschitz continuous functions on X;
- $C^1(X)$: The class of continuously differentiable functions on X.

Definition 2 ($\mathcal{C}$-mixing Process) *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(X, \mathcal{B})$ be a measurable space, $\mathcal{X} := (X_i)_{i\geq 1}$ be an X-valued, stationary process on Ω, and $\|\cdot\|_{\mathcal{C}}$ be defined by (3) for some semi-norm $|||\cdot|||$. Then, for $n \geq 1$, we define the $\mathcal{C}$-mixing coefficients by*

$$\phi_{\mathcal{C}}(\mathcal{X}, n) := \sup\big\{\mathrm{cor}(\psi, g(X_{k+n})) : k \geq 1, \psi \in B_{L_1(\mathcal{A}_1^k,\mu)}, g \in B_{\mathcal{C}(X)}\big\},$$

and the time-reversed $\mathcal{C}$-mixing coefficients by

$$\phi_{\mathcal{C},rev}(\mathcal{X}, n) := \sup\big\{\mathrm{cor}(g(X_k), \varphi) : k \geq 1, g \in B_{\mathcal{C}(X)}, \varphi \in B_{L_1(\mathcal{A}_{k+n}^{\infty},\mu)}\big\}.$$

Let $(d_n)_{n\geq 1}$ be a strictly positive sequence converging to 0. We say that $\mathcal{X}$ ***is*** **(time-reversed) $\mathcal{C}$-mixing** *with rate $(d_n)_{n\geq 1}$, if we have $\phi_{\mathcal{C},rev}(\mathcal{X}, n) \leq d_n$ for all $n \geq 1$. Moreover, if $(d_n)_{n\geq 1}$ is of the form*

$$d_n := c_0 \exp\big(-bn^{\gamma}\big), \qquad n \geq 1,$$

for some constants $c_0 > 0$, $b > 0$, and $\gamma > 0$, then $\mathcal{X}$ is called **geometrically *(time-reversed) $\mathcal{C}$-mixing***.

From the above definition, we see that a $\mathcal{C}$-mixing process is defined in association with an underlying function space. For the above listed function spaces, i.e., $L_\infty(X)$, $BV(X)$, $C_{b,\alpha}(X)$, $\mathrm{Lip}(X)$, and $C^1(X)$, the increase of the smoothness enlarges the class of the associated stochastic processes, as illustrated in Hang and Steinwart (2016). Note that the classical ϕ-mixing process is essentially a $\mathcal{C}$-mixing process associated with the function space $L_\infty(X)$. Note also that not all α-mixing processes are $\mathcal{C}$-mixing, and vice versa. We refer the reader to Hang and Steinwart (2016) for the relations among α-, ϕ- and $\mathcal{C}$-mixing processes.

On the other hand, under the above notations and definitions, from Theorem 4.7 in Maume-Deschamps (2006), we know that Logistic map in Example 1 is geometrically time-reversed $\mathcal{C}$-mixing with $\mathcal{C} = \mathrm{Lip}(0, 1)$ while Theorem 4.4 in Maume-Deschamps (2006) (see also Chapter 3 in Baladi (2000)) indicates that Gauss map in Example 2 is geometrically time-reversed $\mathcal{C}$-mixing with $\mathcal{C} = BV(0, 1)$. Example 3 is also geometrically time-reversed $\mathcal{C}$-mixing with $\mathcal{C} = BV(0, 1)$ according to Maume-Deschamps (2006). For more examples of geometrically time-reversed $\mathcal{C}$-mixing dynamical systems, the reader is referred to Section 2 in Hang and Steinwart (2016).

2.3 Kernel Density Estimation: Assumptions and Formulations

For the smoothing kernel K in the kernel density estimator, in this paper we consider its following general form, namely, d-dimensional smoothing kernel:

Definition 3 *A bounded, monotonically decreasing function $K : [0, \infty) \to [0, \infty)$ is a* ***d-dimensional smoothing kernel*** *if*

$$\int_{\mathbb{R}^d} K(\|x\|)\,\mathrm{d}x =: \kappa \in (0, \infty). \tag{4}$$

The choice of the norm in Definition 3 does not matter since all norms on $\mathbb{R}^d$ are equivalent. To see this, let $\|\cdot\|'$ be another norm on $\mathbb{R}^d$ satisfying $\kappa \in (0,\infty)$. From the equivalence of the two norms on $\mathbb{R}^d$, one can find a positive constant c such that $\|x\| \le c\|x\|'$ holds for all $x \in \mathbb{R}$. Therefore, easily we have

$$\int_{\mathbb{R}^d} K(\|x\|')\,\mathrm{d}x \le \int_{\mathbb{R}^d} K\left(\|x\|/c\right)\,\mathrm{d}x = c^d \int_{\mathbb{R}^d} K(\|x\|)\,\mathrm{d}x < \infty.$$

In what follows, without loss of generality, we assume that the constant κ in Definition 3 equals to 1.

Lemma 4 *A bounded, monotonically decreasing function $K : [0,\infty) \to [0,\infty)$ is a d-dimensional smoothing kernel if and only if*

$$\int_0^\infty K(r) r^{d-1}\,\mathrm{d}r \in (0,\infty).$$

Proof From the above discussions, it suffices to consider the integration constraint for the kernel function K with respect to the Euclidean norm $\|\cdot\|_{\ell_2^d}$. We thus have

$$\int_{\mathbb{R}^d} K\big(\|x\|_{\ell_2^d}\big)\,\mathrm{d}x = d\tau_d \int_0^\infty K(r) r^{d-1}\,\mathrm{d}r,$$

where $\tau_d = \pi^{d/2}/\Gamma\big(\frac{d}{2}+1\big)$ is the volume of the unit ball $B_{\ell_2^d}$ of the Euclidean space ℓ_2^d. This completes the proof of Lemma 4. ■

Let $r \in [0,+\infty)$ and denote $\mathbf{1}_A$ as the indicator function. Several common examples of d-dimensional smoothing kernels $K(r)$ include the Naive kernel $\mathbf{1}_{[0,1]}(r)$, the Triangle kernel $(1-r)\mathbf{1}_{[0,1]}(r)$, the Epanechnikov kernel $(1-r^2)\mathbf{1}_{[0,1]}(r)$, and the Gaussian kernel e^{-r^2}. In this paper, we are interested in the kernels that satisfy the following restrictions on their shape and regularity:

Assumption 5 *For a fixed function space $\mathcal{C}(X)$, we make the following assumptions on the d-dimensional smoothing kernel K:*

(i) *K is Hölder continuous with exponent β with $\beta \in [0,1]$;*

(ii) *$\int_0^\infty K(r) r^{\beta+d-1}\,\mathrm{d}r < \infty$;*

(iii) *For all $x \in \mathbb{R}^d$, we have $|||K(\|x-\cdot\|/h)||| \in \mathcal{C}(X)$ and there exists a function $\varphi : (0,\infty) \to (0,\infty)$ such that*

$$\sup_{x\in\mathbb{R}^d} |||K(\|x-\cdot\|/h)||| \le \varphi(h).$$

It is easy to verify that for $\mathcal{C} = \mathrm{Lip}$, Assumption 5 is met for the Triangle kernel, the Epanechnikov kernel, and the Gaussian kernel. Particularly, Condition *(iii)* holds for all these kernels with $|||\cdot|||$ being the Lipschitz norm and $\varphi(h) \le \mathcal{O}(h^{-1})$. Moreover, as we shall see below, not all the conditions in Assumption 5 are required for the analysis conducted in this study and conditions assumed on the kernel will be specified explicitly.

We now show that given a d-dimensional smoothing kernel K as in Definition 3, one can easily construct a probability density on $\mathbb{R}^d$.

Definition 6 (K-Smoothing of a Measure) *Let K be a d-dimensional smoothing kernel and Q be a probability measure on $\mathbb{R}^d$. Then, for $h > 0$,*

$$f_{Q,h}(x) := f_{Q,K,h}(x) := h^{-d} \int_{\mathbb{R}^d} K\left(\|x - x'\|/h\right) \, \mathrm{d}Q(x'), \qquad x \in \mathbb{R}^d,$$

*is called a K-**smoothing of** Q.*

It is not difficult to see that $f_{Q,h}$ defines a probability density on $\mathbb{R}^d$ since Fubini's theorem yields that

$$\begin{aligned}\int_{\mathbb{R}^d} f_{Q,h}(x) \, \mathrm{d}x &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h^{-d} K\left(\|x - x'\|/h\right) \, \mathrm{d}Q(x') \, \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} K(\|x\|) \, \mathrm{d}x \, \mathrm{d}Q(x') = 1.\end{aligned}$$

Let us denote $K_h : \mathbb{R}^d \to [0, +\infty)$ as

$$K_h(x) := h^{-d} K\left(\|x\|/h\right), \; x \in \mathbb{R}^d. \tag{5}$$

Note that K_h also induces a density function on $\mathbb{R}^d$ since there holds $\|K_h\|_1 = 1$.

For the sake of notational simplification, in what follows, we introduce the convolution operator $*$. Under this notation, we then see that $f_{Q,h}$ is the density of the measure that is the convolution of the measure Q and $\nu_h = K_h \, \mathrm{d}\lambda^d$. Recalling that P is a probability measure on $\mathbb{R}^d$ with the corresponding density function f, by taking $Q := P$ with $\mathrm{d}P = f \, \mathrm{d}\lambda^d$, we have

$$f_{P,h} = K_h * f = f * K_h = K_h * \mathrm{d}P. \tag{6}$$

Since $K_h \in L_\infty(\mathbb{R}^d)$ and $f \in L_1(\mathbb{R}^d)$, from Proposition 8.8 in Folland (1999), we know that $f_{P,h}$ is uniformly continuous and bounded. Specifically, when Q is the empirical measure $D_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, the *kernel density estimator for dynamical systems* in this study can be expressed as

$$f_{D_n,h}(x) = K_h * \mathrm{d}D_n(x) = n^{-1} h^{-d} \sum_{i=1}^n K\left(\|x - x_i\|/h\right). \tag{7}$$

From now on, for notational simplicity, we will suppress the subscript n of D_n and denote $D := D_n$, e.g., $f_{D,h} := f_{D_n,h}$.

3. Consistency and Convergence Analysis

In this section, we study the consistency and convergence rates of $f_{D,h}$ to the true density f under L_1-norm and also L_∞-norm for some special cases. Recall that $f_{D,h}$ is a nonparametric density estimator and so the criterion that measures its goodness-of-fit matters, which, for instance, includes L_1-distance, L_2-distance, and L_∞-distance.

In the literature of kernel density estimation, probably the most frequently employed criterion is the L_2-distance of the difference between $f_{D,h}$ and f, since it entails an exact

bias-variance decomposition and can be analyzed relatively easily by using Taylor expansion involved arguments. However, it is argued in Devroye and Györfi (1985) (see also Devroye and Lugosi (2001)) that L_1-distance could be a more reasonable choice since: it is invariant under monotone transformations; it is always well-defined as a metric on the space of density functions; it is also proportional to the total variation metric and so leads to better visualization of the closeness to the true density function than L_2-distance. The downside of using L_1-distance is that it does not admit an exact bias-variance decomposition and the usual Taylor expansion involved techniques for error estimation may not apply directly. Nonetheless, if we introduce the intermediate estimator $f_{P,h}$ in (6), obviously the following inequality holds

$$\|f_{D,h} - f\|_1 \leq \|f_{D,h} - f_{P,h}\|_1 + \|f_{P,h} - f\|_1. \tag{8}$$

The consistency and convergence analysis in our study will be mainly conducted in the L_1 sense with the help of inequality (8). Besides, for some specific case, i.e., when the density f is compactly supported, we are also concerned with the consistency and convergence of $f_{D,h}$ to f under L_∞-norm. In this case, there also holds the following inequality

$$\|f_{D,h} - f\|_\infty \leq \|f_{D,h} - f_{P,h}\|_\infty + \|f_{P,h} - f\|_\infty. \tag{9}$$

It is easy to see that the first error term on the right-hand side of (8) or (9) is stochastic due to the empirical measure D while the second one is deterministic because of its sampling-free nature. Loosely speaking, the first error term corresponds to the variance of the estimator $f_{D,h}$, while the second one can be treated as its bias although (8) or (9) is not an exact error decomposition. In our study, we proceed with the consistency and convergence analysis on $f_{D,h}$ by bounding the two error terms, respectively.

3.1 Bounding the Deterministic Error Term

Our first theoretical result on bounding the deterministic error term shows that, given a d-dimensional kernel K, the L_1-distance between its K-smooth of the measure P, i.e., $f_{P,h}$, and f can be arbitrarily small by choosing the bandwidth appropriately. Moreover, under mild assumptions on the regularity of f and K, the L_∞-distance between the two quantities possesses a polynomial decay with respect to the bandwidth h.

Theorem 7 *Let K be a d-dimensional smoothing kernel.*

(i) For any $\varepsilon > 0$, there exists $0 < h_\varepsilon \leq 1$ such that for any $h \in (0, h_\varepsilon]$ we have

$$\|f_{P,h} - f\|_1 \leq \varepsilon.$$

(ii) If K satisfies Condition (ii) in Assumption 5 and f is α-Hölder continuous with $\alpha \leq \beta$, then there holds

$$\|f_{P,h} - f\|_\infty \lesssim h^\alpha.$$

We now show that the L_1-distance between $f_{P,h}$ and f can be upper bounded by their difference (in the sense of L_∞-distance) on a compact domain of $\mathbb{R}^d$ together with their difference (in the sense of L_1-distance) outside this domain. As we shall see later, this observation will entail us to consider different classes of the true density f. The following result is crucial in our subsequent analysis on the consistency and convergence rates of $f_{D,h}$.

Theorem 8 *Assume that K is a d-dimensional smoothing kernel that satisfies Conditions (i) and (ii) in Assumption 5. For $h \leq 1$ and $r \geq 1$, we have*

$$\|f_{P,h} - f\|_1 \lesssim r^d \|f_{P,h} - f\|_\infty + P(H_{r/2}) + (h/r)^\beta.$$

3.2 Bounding the Stochastic Error Term

We now proceed with the estimation of the stochastic error term $\|f_{D,h} - f_{P,h}\|_1$ by establishing probabilistic oracle inequalities. For the sake of readability, let us start with an overview of the analysis conducted in this subsection for bounding the stochastic error term.

3.2.1 An Overview of the Analysis

In this study, the stochastic error term is tackled by using capacity-involved arguments and the Bernstein-type inequality established in Hang and Steinwart (2016). In the sequel, for any fixed $x \in \Omega \subset \mathbb{R}^d$, we write

$$k_{x,h} := h^{-d} K(\|x - \cdot\|/h), \tag{10}$$

and we further denote the centered random variable $\widetilde{k}_{x,h}$ on Ω as

$$\widetilde{k}_{x,h} := k_{x,h} - \mathbb{E}_P k_{x,h}. \tag{11}$$

It thus follows that

$$\mathbb{E}_D \widetilde{k}_{x,h} = \mathbb{E}_D k_{x,h} - \mathbb{E}_P k_{x,h} = f_{D,h}(x) - f_{P,h}(x),$$

and consequently we have

$$\|f_{D,h} - f_{P,h}\|_1 = \int_{\mathbb{R}^d} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x,$$

and

$$\|f_{D,h} - f_{P,h}\|_\infty = \sup_{x \in \Omega} |\mathbb{E}_D \widetilde{k}_{x,h}|.$$

As a result, in order to bound $\|f_{D,h} - f_{P,h}\|_1$, it suffices to bound the supremum of the empirical process $\mathbb{E}_D \widetilde{k}_{x,h}$ indexed by $x \in \mathbb{R}^d$. For any $r > 0$, there holds

$$\|f_{D,h} - f_{P,h}\|_1 = \int_{B_r} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x + \int_{H_r} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x.$$

The second term of the right-hand side of the above equality can be similarly dealt with as in the proof of Theorem 8. In order to bound the first term, we define $\widetilde{\mathcal{K}}_{h,r}$ as the function set of $\widetilde{k}_{x,h}$ that corresponds to x which lies on a radius-r ball of $\mathbb{R}^d$:

$$\widetilde{\mathcal{K}}_{h,r} := \left\{\widetilde{k}_{x,h} : x \in B_r\right\} \subset L_\infty(\mathbb{R}^d).$$

The idea here is to apply capacity-involved arguments and the Bernstein-type exponential inequality in Hang and Steinwart (2016) to the function set $\widetilde{\mathcal{K}}_{h,r}$ and the associated empirical process $\mathbb{E}_D\widetilde{k}_{x,h}$. The difference between $f_{D,h}$ and $f_{P,h}$ under the L_∞-norm can be bounded analogously. Therefore, to further our analysis, we first need to bound the capacity of $\widetilde{\mathcal{K}}_{h,r}$ in terms of covering numbers.

3.2.2 Bounding the Capacity of the Function Set $\widetilde{\mathcal{K}}_{h,r}$

Definition 9 (Covering Number) *Let (X, d) be a metric space and $A \subset X$. For $\varepsilon > 0$, the ε-**covering number** of A is denoted as*

$$\mathcal{N}(A, d, \varepsilon) := \min\left\{n \geq 1 : \exists\, x_1, \cdots, x_n \in X \text{ such that } A \subset \bigcup_{i=1}^{n} B_d(x_i, \varepsilon)\right\},$$

where $B_d(x, \varepsilon) := \{x' \in X : d(x, x') \leq \varepsilon\}$.

For any fixed $r \geq 1$, we consider the function set

$$\mathcal{K}_{h,r} := \{k_{x,h} : x \in B_r\} \subset L_\infty(\mathbb{R}^d).$$

The following proposition provides an estimate of the covering number of $\mathcal{K}_{h,r}$.

Proposition 10 *Let K be a d-dimensional smoothing kernel that satisfies Condition (i) in Assumption 5 and $h \in (0, 1]$. Then there exists a positive constant c' such that for all $\varepsilon \in (0, 1]$, we have*

$$\mathcal{N}(\mathcal{K}_{h,r}, \|\cdot\|_\infty, \varepsilon) \leq c' r^d h^{-d-\frac{d^2}{\beta}} \varepsilon^{-\frac{d}{\beta}}.$$

3.2.3 Oracle Inequalities under L_1-Norm, and L_∞-Norm

We now establish oracle inequalities for the kernel density estimator (7) under L_1-norm, and L_∞-norm, respectively. These oracle inequalities will be crucial in establishing the consistency and convergence results of the estimator. Recall that the considered kernel density estimation problem is based on samples from an X-valued $\mathcal{C}$-mixing process which is associated with an underlying function class $\mathcal{C}(X)$. As shown below, the established oracle inequality holds without further restrictions on the support of the density function.

Theorem 11 *Suppose that Assumption 5 holds. Let $\mathcal{X} := (X_n)_{n\geq 1}$ be an X-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with $\|\cdot\|_{\mathcal{C}}$ being defined for some semi-norm $|||\cdot|||$ that satisfies Assumption 1. Then for all $0 < h \leq 1$, $r \geq 1$ and*

$\tau \geq 1$, *there exists an* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$, *with probability* μ *at least* $1 - 3e^{-\tau}$, *there holds*

$$\|f_{D,h} - f_{P,h}\|_1 \lesssim \sqrt{\frac{(\log n)^{2/\gamma} r^d (\tau + \log \frac{nr}{h})}{h^d n}} + \frac{(\log n)^{2/\gamma} r^d (\tau + \log \frac{nr}{h})}{h^d n} + P(H_{r/4}) + \sqrt{\frac{32\tau (\log n)^{2/\gamma}}{n}} + \left(\frac{h}{r}\right)^{\beta}.$$

Here n_0 *will be given explicitly in the proof.*

Our next result shows that when the density function f is compactly supported and bounded, an oracle inequality under L_∞-norm can be also derived.

Theorem 12 *Let* K *be a* d*-dimensional kernel function that satisfies Conditions* (i) *and* (iii) *in Assumption 5. Let* $\mathcal{X} := (X_n)_{n \geq 1}$ *be an* X*-valued stationary geometrically (time-reversed)* $\mathcal{C}$*-mixing process on* $(\Omega, \mathcal{A}, \mu)$ *with* $\|\cdot\|_{\mathcal{C}}$ *being defined for some semi-norm* $|||\cdot|||$ *that satisfies Assumption 1. Assume that there exists a constant* $r_0 \geq 1$ *such that* $\Omega \subset B_{r_0} \subset \mathbb{R}^d$ *and the density function* f *satisfies* $\|f\|_\infty < \infty$. *Then for all* $0 < h \leq 1$ *and* $\tau > 0$, *there exists an* $n_0^* \in \mathbb{N}$ *such that for all* $n \geq n_0^*$, *with probability* μ *at least* $1 - e^{-\tau}$, *there holds*

$$\|f_{D,h} - f_{P,h}\|_\infty \lesssim \sqrt{\frac{\|f\|_\infty (\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)(\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n}.$$

Here n_0^* *will be given explicitly in the proof.*

In Theorem 12, the kernel K is only required to satisfy Conditions (i) and (iii) in Assumption 5 whereas the condition that $\int_0^\infty K(r) r^{\beta+d-1} \, \mathrm{d}r < \infty$ for some $\beta > 0$ is not needed. This is again due to the compact support assumption of the density function f as stated in Theorem 12.

3.3 Results on Universal Consistency

We now present results on the universal consistency property of the kernel density estimator $f_{D,h}$ in the sense of L_1-norm. A kernel density estimator $f_{D,h}$ is said to be *universally consistent* in the sense of L_1-norm if $f_{D,h}$ converges to f almost surely under L_1-norm without any further restrictions on the probability distribution P.

Theorem 13 *Let* K *be a* d*-dimensional smoothing kernel that satisfies Conditions* (i) *and* (iii) *in Assumption 5. Let* $\mathcal{X} := (X_n)_{n \geq 1}$ *be an* X*-valued stationary geometrically (time-reversed)* $\mathcal{C}$*-mixing process on* $(\Omega, \mathcal{A}, \mu)$ *with* $\|\cdot\|_{\mathcal{C}}$ *being defined for some semi-norm* $|||\cdot|||$ *that satisfies Assumption 1. If*

$$h_n \to 0 \quad \text{and} \quad \frac{n h_n^d}{(\log n)^{(2+\gamma)/\gamma}} \to \infty, \quad \text{as} \quad n \to \infty,$$

then the kernel density estimator f_{D,h_n} *is universally consistent in the sense of* L_1*-norm.*

3.4 Convergence Rates under L_1-Norm

The consistency result in Theorem 13 is independent of the probability distribution P and is therefore said to be universal. In this subsection, we will show that if certain tail assumptions on P are applied, convergence rates can be obtained under L_1-norm. Here, we consider three different situations, namely, the tail of the probability distribution P has a polynomial decay, exponential decay and disappears, respectively.

Theorem 14 *Let K be a d-dimensional smoothing kernel that satisfies Assumption 5. Assume that the density f is α-Hölder continuous with $\alpha \leq \beta$. Let $\mathcal{X} := (X_n)_{n\geq 1}$ be an X-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with $\|\cdot\|_{\mathcal{C}}$ being defined for some semi-norm $|||\cdot|||$ that satisfies Assumption 1. We consider the following cases:*

(i) $P(H_r) \lesssim r^{-\eta d}$ *for some* $\eta > 0$ *and for all* $r \geq 1$*;*

(ii) $P(H_r) \lesssim e^{-ar^{\eta}}$ *for some* $a > 0$*,* $\eta > 0$ *and for all* $r \geq 1$*;*

(iii) $P(H_{r_0}) = 0$ *for some* $r_0 \geq 1$*.*

For the above cases, if $n \geq n_0$ with n_0 the same as in Theorem 11, and the sequences h_n are of the following forms:

(i) $h_n = \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{1+\eta}{(1+\eta)(2\alpha+d)-\alpha}}$*;*

(ii) $h_n = \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{1}{2\alpha+d}} (\log n)^{-\frac{d}{\gamma}\cdot\frac{1}{2\alpha+d}}$*;*

(iii) $h_n = \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{1}{2\alpha+d}}$*;*

then with probability μ at least $1 - \frac{1}{n}$, there holds

$$\|f_{D,h_n} - f\|_1 \leq \varepsilon_n,$$

where the convergence rates

(i) $\varepsilon_n \lesssim \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{\alpha\eta}{(1+\eta)(2\alpha+d)-\alpha}}$*;*

(ii) $\varepsilon_n \lesssim \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{d}{\gamma}\cdot\frac{\alpha+d}{2\alpha+d}}$*;*

(iii) $\varepsilon_n \lesssim \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{\alpha}{2\alpha+d}}$*.*

3.5 Convergence Rates under L_∞-Norm

In Subsection 3.2.3, when the density function f is bounded and compactly supported, we establish oracle inequality of $f_{D,h}$ under L_∞-norm. Combining this with the estimate of the deterministic error term in Theorem 7 *(ii)* under L_∞-norm, we arrive at the following result that characterizes the convergence of $f_{D,h}$ to f under L_∞-norm.

Theorem 15 *Let K be a d-dimensional smoothing kernel that satisfies Conditions (i) and (iii) in Assumption 5. Let $\mathcal{X} := (X_n)_{n\geq 1}$ be an X-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with $\|\cdot\|_{\mathcal{C}}$ being defined for some semi-norm $|||\cdot|||$ that satisfies Assumption 1. Assume that there exists a constant $r_0 \geq 1$ such that $\Omega \subset B_{r_0} \subset \mathbb{R}^d$ and the density function f is α-Hölder continuous with $\alpha \leq \beta$ and $\|f\|_\infty < \infty$. Then for all $n \geq n_0^*$ with n_0^* as in Theorem 12, by choosing*

$$h_n = \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{1}{2\alpha+d}},$$

with probability μ at least $1 - \frac{1}{n}$, there holds

$$\|f_{D,h_n} - f\|_\infty \lesssim \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{\alpha}{2\alpha+d}}. \tag{12}$$

In Theorems 14 and 15, one needs to ensure that $n \geq n_0$ with n_0 as in Theorem 11 and $n \geq n_0^*$ with n_0^* as in Theorem 12, respectively. One may also note that due to the involvement of the term $\varphi(h_n)$, the numbers n_0 and n_0^* depend on the h_n. However, recalling that for the Triangle kernel, the Epanechnikov kernel, and the Gaussian kernel, we have $\varphi(h_n) \leq \mathcal{O}(h_n^{-1})$, which, together with the choices of h_n in Theorems 14 and 15, implies that n_0 and n_0^* are well-defined. It should be also remarked that in the scenario where the density function f is compactly supported and bounded, the convergence rate of $f_{D,h}$ to f is not only obtainable, but also the same with that derived under L_1-norm. This is indeed an interesting observation since convergence under L_∞-norm implies convergence under L_1-norm.

3.6 Comments and Discussions

This section presents some comments on the obtained theoretical results on the consistency and convergence rates of $f_{D,h}$ and compares them with related findings in the literature.

We highlight that in our analysis the density function f is only assumed to be Hölder continuous. As pointed out in the introduction, in the context of dynamical systems, this seems to be more than a reasonable assumption. On the other hand, the consistency and the convergence results obtained in our study, are of type "with high probability" due to the use of the Bernstein-type exponential inequality that takes into account the variance information of the random variables. From our analysis and the obtained theoretical results, one can also easily observe the influence of the dependence among observations. For instance, from Theorem 13 we see that with increasing dependence among observations (corresponding to smaller γ), in order to ensure the universal consistency of f_{D,h_n}, the decay of h_n (with respect to n^{-1}) is required to be faster. This is in fact also the case if we look at results on the convergence rates in Theorems 14 and 15. Moreover, the influence of the dependence among observations is also indicated there. That is, an increase of the dependence among observations may slow down the convergence of $f_{D,h}$ in the sense of both L_1-norm and L_∞-norm. It is also interesting to note that when γ tends to infinity, which corresponds to the case where observations can be roughly treated as independent ones, meaningful convergence rates can be also deduced. It turns out that, up to a logarithmic factor, the established convergence rates (12) under L_∞-norm, namely, $\mathcal{O}(((\log n)^{(2+\gamma)/\gamma}/n)^{\alpha/(2\alpha+d)})$, match the optimal rates in the i.i.d. case, see, e.g., Khas′minskii (1979) and Stone (1983).

As mentioned in the introduction, there exist several studies in the literature that address the kernel density estimation problem for dynamical systems. For example, Bosq and Guégan (1995) conducted some first studies and showed the point-wise consistency and the convergence (in expectation) of the kernel density estimator. The convergence rates obtained in their study are of the type $\mathcal{O}(n^{-4/(4+2d)})$, which are conducted in terms of the variance of $f_{D,h}$. The notion they used for measuring the dependence among observations is α-mixing coefficient (see A_3 in Bosq and Guégan (1995)). Considering the density estimation problem for one-dimensional dynamical systems, Prieur (2001) presented some studies on the kernel density estimator $f_{D,h}$ by developing a central limit theorem and apply it to bound the variance of the estimator. Further some studies on the kernel density estimation of the invariant Lebesgue density for dynamical systems were conducted in Blanke et al. (2003). By considering both dynamical noise and observational noise, point-wise convergence of the estimator $f_{D,h}$ in expectation was established, i.e., the convergence of $\mathbb{E}f_{D,h}(x) - f(x)$ for any $x \in \mathbb{R}^d$. Note further that these results rely on the second-order smoothness and boundedness of f. Therefore, the second-order smoothness assumption on the density function together with the point-wise convergence in expectation makes it different from our work. In particular, under the additional assumption on the tail of the noise distribution, the convergence of $\mathbb{E}(f_{D,h}(x) - f(x))^2$ for any fixed $x \in \mathbb{R}^d$ is of the order $\mathcal{O}(n^{-2/(2+\beta d)})$ with $\beta \geq 1$. Concerning the convergence of $f_{D,h}$ in a dynamical system setup, Maume-Deschamps (2006) also presented some interesting studies which in some sense also motivated our work here. By using also the $\mathcal{C}$-mixing concept as adopted in our study to measure the dependence among observations from dynamical systems, she presented the point-wise convergence of $f_{D,h}$ with the help of Hoeffding-type exponential inequality (see Proposition 3.1 in Maume-Deschamps (2006)). The assumption applied on f is that it is bounded from below and also α-Hölder continuous (more precisely, f is assumed to be α-*regular*, see Assumption 2.3 in Maume-Deschamps (2006)). Hence, from the above discussions, we suggest that the work we present in this study is essentially different from that in Maume-Deschamps (2006).

4. Bandwidth Selection and Simulation Studies

This section discusses the model selection problem of the kernel density estimator (7) by performing numerical simulation studies. In the context of kernel density estimation, model selection is mainly referred to the choice of the smoothing kernel K and the selection of the kernel bandwidth h, which are of crucial importance for the practical implementation of the data-driven density estimator. According to our experimental experience and also the empirical observations reported in Maume-Deschamps (2006), it seems that the choice of the kernel or the noise does not have a significant influence on the performance of the estimator. Therefore, our emphasis will be placed on the bandwidth selection problem in our simulation studies.

4.1 Several Bandwidth Selectors

In the literature of kernel density estimation, various bandwidth selectors have been proposed, several typical examples of which have been alluded to in the introduction. When turning to the case with dependent observations, the bandwidth selection problem has been

also drawing much attention, see e.g., Hart and Vieu (1990); Chu and Marron (1991); Hall et al. (1995); Yao and Tong (1998). Among existing bandwidth selectors, probably the most frequently employed ones are based on the cross-validation ideas. For cross-validation bandwidth selectors, one tries to minimize the integrated squared error (ISE) of the empirical estimator $f_{D,h}$ where

$$\mathrm{ISE}(h) := \int (f_{D,h} - f)^2 = \int f_{D,h}^2 - 2\int f_{D,h} \cdot \int f + \int f^2.$$

Note that on the right-hand side of the above equality, the last term $\int f^2$ is independent of h and so the minimization of $\mathrm{ISE}(h)$ is equivalent to minimize

$$\int f_{D,h}^2 - 2\int f_{D,h} \cdot \int f.$$

It is shown that with i.i.d observations, an unbiased estimator of the above quantity, which is termed as least squares cross-validation (LSCV), is given as follows:

$$\mathrm{LSCV}(h) := \int f_{D,h}^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,h}(x_i), \tag{13}$$

where the leave-one-out density estimator $\hat{f}_{-i,h}$ is defined as

$$\hat{f}_{-i,h}(x) := \frac{1}{n-1}\sum_{j\neq i}^{n} K_h(x - x_j).$$

When the observations are dependent, it is shown that cross-validation can produce much under-smoothed estimates, see e.g., Hart and Wehrly (1986); Hart and Vieu (1990). Observing this, Hart and Vieu (1990) proposed the modified least squares cross-validation (MLSCV), which is defined as follows

$$\mathrm{MLSCV}(h) := \int f_{D,h}^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,h,l_n}(x_i), \tag{14}$$

where l_n is set to 1 or 2 as suggested in Hart and Vieu (1990) and

$$\hat{f}_{-i,h,l_n}(x) := \frac{1}{\#\{j : |j-i| > l_n\}} \sum_{|j-i|>l_n} K_h(x - x_j).$$

The underlying intuition of proposing MLSCV is that when estimating the density of a fixed point, ignoring observations in the vicinity of this point may be help in reducing the influence of dependence among observations. However, when turning to the L_1 point of view, the above bandwidth selectors may not work well due to the use of the least squares criterion. Alternatively, Devroye (1989) proposed the double kernel bandwidth selector that minimizes the following quantity

$$\mathrm{DKM}(h) := \int |f_{D,h,K} - f_{D,h,L}|, \tag{15}$$

where $f_{D,h,K}$ and $f_{D,h,L}$ are kernel density estimators based on the kernels K and L, respectively. Some rigorous theoretical treatments on the effectiveness of the above bandwidth selector were made in Devroye (1989).

Our purpose in simulation studies is to conduct empirical comparisons among the above bandwidth selectors in the dynamical system context instead of proposing new approaches.

4.2 Experimental Setup

In our experiments, observations $x_1, \cdots, x_n$ are generated from the following model[1]

$$\begin{cases} \tilde{x}_i = T^i(x_0), \\ x_i = \tilde{x}_i + \varepsilon_i, \end{cases} \quad i = 1, \cdots, n, \tag{16}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ is set to 0.01 and the initial state x_0 is randomly generated based on the density f. For the map T in (16), we choose Logistic map in Example 1 and Gauss map in Example 2. We vary the sample size among $\{5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$, implement bandwidth selection procedures over 20 replications and select the bandwidth from a grid of values in the interval $[h_L, h_U]$ with 100 equispaced points. Here, h_L is set as the minimum distance between consecutive points x_i, $i = 1, \cdots, n$ (Devroye and Lugosi, 1997), while h_U is chosen according to the *maximal smoothing principle* proposed in Terrell (1990). Throughout our experiments, we use the Gaussian kernel for the kernel density estimators.

In our experiments, we conduct comparisons among the above-mentioned bandwidth selectors which are, respectively, denoted as follows:

- LSCV: the least squares cross-validation given in (13);
- MLSCV-1: the modified least squares cross-validation in (14) with $l_n = 1$;
- MLSCV-2: the modified least squares cross-validation in (14) with $l_n = 2$;
- DKM: the double kernel method defined in (15) where the two kernels used here are the Epanechnikov kernel and the Triangle kernel, respectively.

In the experiments, due to the known density functions for Logistic map and Gauss map, and in accordance with our previous analysis from the L_1 point of view, the criterion of comparing different selected bandwidths is the following absolute mean error (AME):

$$\text{AME}(h) = \frac{1}{m} \sum_{i=1}^{m} |f_{D,h}(u_i) - f(u_i)|,$$

where $u_1, \cdots, u_m$ are m equispaced points in the interval $[0, 1]$ and m is set to 10000. We also compare the selected bandwidth with the one that has the minimum absolute mean error which serves as a ***baseline*** method in our experiments.

1. Note that here the observational noise is assumed for the considered dynamical system (16), which differs from (2) and can be a more realistic setup from an empirical and experimental viewpoint. In fact, it is observed also in Maume-Deschamps (2006) that the influence of low SNR noise is not obvious in density estimation. We therefore adopt this setup in our experiments. All the observations reported in this experimental section apply to the noiseless case (2).

4.3 Simulation Results and Observations

The AMEs of the above bandwidth selectors for Logistic map in Example 1 and Gauss map in Example 2 over 20 replications are averaged and recorded in Tables 1 and 2 below.

In Figs. 1 and 2, we also plot the kernel density estimators for Logistic map in Example 1 and Gauss map in Example 2 with different bandwidths and their true density functions with different sample sizes. The sample size of each panel, in Figs. 1 and 2, from up to bottom, is 10^3, 10^4, and 10^5, respectively. In each panel, the densely dashed black curve represents the true density, the dotted blue curve is the estimated density function with the bandwidth selected by the baseline method while the solid red curve stands for the estimated density with the bandwidth selected by the double kernel method. All density functions in Figs. 1 and 2 are plotted with 100 equispaced points in the interval $(0, 1)$.

Table 1: The AMEs of Different Bandwidth Selectors for Logistic Map in Example 1

sample size	LSCV	MLSCV-1	MLSCV-2	DKM	Baseline
5×10^2	.3372	.3369	.3372	.3117	.3013
1×10^3	.2994	.2994	.2994	.2804	.2770
5×10^3	.2422	.2422	.2422	.2340	.2326
1×10^4	.2235	.2235	.2235	.2220	.2192

Table 2: The AMEs of Different Bandwidth Selectors for Gauss Map in Example 2

sample size	LSCV	MLSCV-1	MLSCV-2	DKM	Baseline
5×10^2	.1027	.1026	.1059	.1181	.0941
1×10^3	.0925	.0933	.0926	.0925	.0878
5×10^3	.0626	.0626	.0626	.0586	.0585
1×10^4	.0454	.0454	.0454	.0440	.0439

From Tables 1 and 2, and Figs. 1 and 2, we see that the true density functions of Logistic map and Gauss map can be approximated well with enough observations and the double kernel method works slightly better than the other three methods for the two dynamical systems. In fact, according to our experimental experience, we find that the bandwidth selector of the kernel density estimator for a dynamical system is usually ad-hoc. That is, for existing bandwidth selectors, there seems no a universal optimal one that can be applicable to all dynamical systems and outperforms the others. Therefore, further exploration and insights on the bandwidth selection problem in the dynamical system context certainly deserve future study. On the other hand, we also notice that due to the presence of dependence among observations generated by dynamical systems, the sample size usually needs to be large enough to approximate the density function well. This can be also seen from the plotted density functions in Figs. 1 and 2 with varying sample sizes.

Aside from the above observations, not surprisingly, from Figs. 1 and 2, we also observe the *boundary effect* (Gasser et al., 1985) from the kernel density estimators for dynamical systems, which seems to be even more significant than the i.i.d case. From a practical implementation view, some special studies are arguably called for addressing this problem.

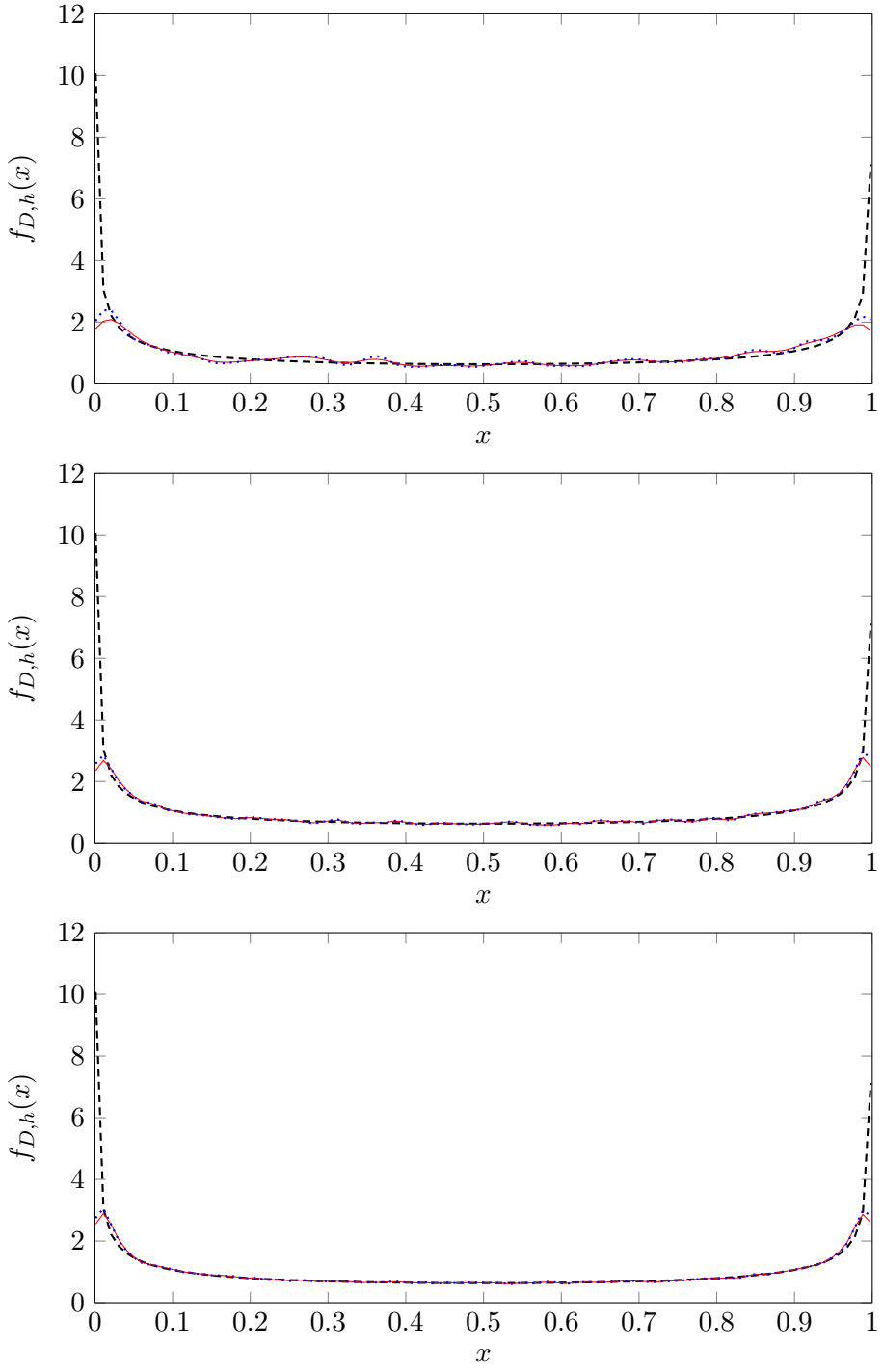

Figure 1: Plots of the kernel density estimators $f_{D,h}$ for Logistic map in Example 1 with different bandwidths and its true density with different sample sizes. The sample size of each panel, from up to bottom, is 10^3, 10^4, and 10^5, respectively. In each panel, the dashed black curve represents the true density of Logistic map, the dotted blue curve is the estimated density of Logistic map with the bandwidth selected by the baseline method while the solid red curve stands for the estimated density of Logistic map with the bandwidth selected by the double kernel method. All curves are plotted with 100 equispaced points in the interval $(0, 1)$.

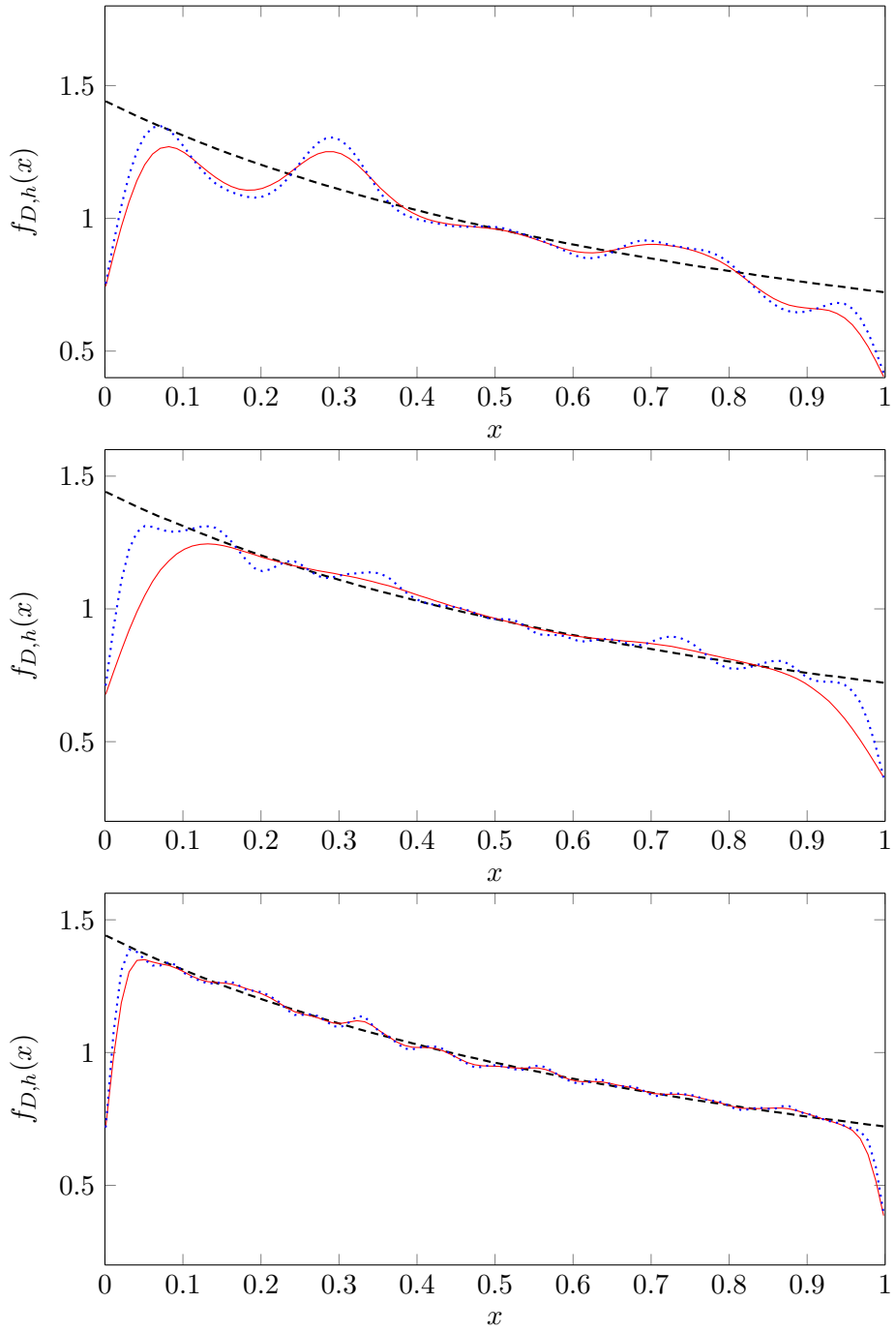

Figure 2: Plots of the kernel density estimators $f_{D,h}$ for Gauss map in Example 2 with different bandwidths and its true density with different sample sizes. The sample size of each panel, from up to bottom, is 10^3, 10^4, and 10^5, respectively. In each panel, the dashed black curve represents the true density of Gauss map, the dotted blue curve is the estimated density of Gauss map with the bandwidth selected by the baseline method while the solid red curve stands for the estimated density of Gauss map with the bandwidth selected by the double kernel method. All curves are plotted with 100 equispaced points in the interval $(0, 1)$.

5. Proofs of Section 3

Proof [of Theorem 7] *(i)* Since the space of continuous and compactly supported functions $C_c(\mathbb{R}^d)$ is dense in $L_1(\mathbb{R}^d)$, we can find $\bar{f} \in C_c(\mathbb{R}^d)$ such that

$$\|f - \bar{f}\|_1 \leq \varepsilon/3, \ \ \forall \varepsilon > 0.$$

Therefore, for any $\varepsilon > 0$, we have

$$\begin{aligned} \|f_{P,h} - f\|_1 &= \int_{\mathbb{R}^d} |f * K_h - f| \, \mathrm{d}x \\ &\leq \int_{\mathbb{R}^d} |f * K_h - \bar{f} * K_h| \, \mathrm{d}x + \int_{\mathbb{R}^d} |\bar{f} * K_h - \bar{f}| \, \mathrm{d}x + \int_{\mathbb{R}^d} |f - \bar{f}| \, \mathrm{d}x \qquad (17) \\ &\leq \frac{2\varepsilon}{3} + \int_{\mathbb{R}^d} |\bar{f} * K_h - \bar{f}| \, \mathrm{d}x, \end{aligned}$$

where K_h is defined in (5) and the last inequality follows from the fact that

$$\|f * K_h - \bar{f} * K_h\|_1 \leq \|f - \bar{f}\|_1 \leq \varepsilon/3.$$

The above inequality is due to Young's inequality (8.7) in Folland (1999). Moreover, there exist a constant $M > 0$ such that $\mathrm{supp}(\bar{f}) \subset B_M$ and a constant $r > 0$ such that

$$\int_{H_r} K(\|x\|) \, \mathrm{d}x \leq \frac{\varepsilon}{9\|\bar{f}\|_1}.$$

Now we define $L : \mathbb{R}^d \to [0, \infty)$ by

$$L(x) := \mathbf{1}_{[-r,r]}(\|x\|) K(\|x\|)$$

and $L_h : \mathbb{R}^d \to [0, \infty)$ by

$$L_h(x) := h^{-d} L(x/h).$$

Then we have

$$\begin{aligned} \int_{\mathbb{R}^d} |\bar{f} * K_h - \bar{f}| \, \mathrm{d}x &\leq \int_{\mathbb{R}^d} |\bar{f} * K_h - \bar{f} * L_h| \, \mathrm{d}x + \int_{\mathbb{R}^d} |\bar{f} * L_h - \bar{f}| \, \mathrm{d}x \\ &\leq \|\bar{f}\|_1 \|K_h - L_h\|_1 + \int_{\mathbb{R}^d} \left| \bar{f} * L_h - \bar{f} \int_{\mathbb{R}^d} L_h \, \mathrm{d}x \right| \mathrm{d}x \\ &\quad + \int_{\mathbb{R}^d} \left| \bar{f} * \int_{\mathbb{R}^d} (L_h - K_h) \, \mathrm{d}x \right| \mathrm{d}x \\ &\leq 2\|\bar{f}\|_1 \|K_h - L_h\|_1 + \int_{\mathbb{R}^d} \left| \bar{f} * L_h - \bar{f} \int_{\mathbb{R}^d} L_h \, \mathrm{d}x \right| \mathrm{d}x. \end{aligned}$$

Moreover, we have

$$\begin{aligned} \|K_h - L_h\|_1 &= \int_{\mathbb{R}^d} \frac{1}{h^d} \left| \mathbf{1}_{[-r,r]}\left(\frac{\|x\|}{h}\right) K\left(\frac{\|x\|}{h}\right) - K\left(\frac{\|x\|}{h}\right) \right| \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \left| \mathbf{1}_{[-r,r]}(\|x\|) K(\|x\|) - K(\|x\|) \right| \mathrm{d}x \\ &= \int_{H_r} K(\|x\|) \, \mathrm{d}x \leq \frac{\varepsilon}{9\|\bar{f}\|_1}. \end{aligned}$$

Finally, for $h \le 1$, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\left|\bar f * L_h - \bar f \int_{\mathbb{R}^d} L_h \,\mathrm{d}x\right| \mathrm{d}x &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} \big(\bar f(x-x') - \bar f(x)\big) L_h(x')\,\mathrm{d}x'\right| \mathrm{d}x \\ &\le \int_{B_{r+M}} \int_{\mathbb{R}^d} |\bar f(x-x') - \bar f(x)| L_h(x')\,\mathrm{d}x'\,\mathrm{d}x.\end{aligned}$$

Since $\bar f$ is uniformly continuous, there exists a constant $h_\varepsilon > 0$ such that for all $h \le h_\varepsilon$ and $\|x'\| \le rh$, we have

$$|\bar f(x-x') - \bar f(x)| \le \varepsilon' := \frac{\varepsilon}{9(r+M)^d \lambda^d(B_1)}.$$

Consequently we obtain

$$\int_{\mathbb{R}^d} |\bar f(x-x') - \bar f(x)| L_h(x')\,\mathrm{d}x' \le \varepsilon' \int_{B_{rh}} L_h(x')\,\mathrm{d}x' \le \varepsilon' \int_{\mathbb{R}^d} K_h \,\mathrm{d}x = \varepsilon'.$$

Therefore, we obtain

$$\int_{\mathbb{R}^d}\left|\bar f * L_h - \bar f \int_{\mathbb{R}^d} L_h \,\mathrm{d}x\right| \mathrm{d}x \le \int_{B_{r+M}} \varepsilon' \,\mathrm{d}x = \frac{\varepsilon}{9} \tag{18}$$

and consequently the assertion can be proved by combining estimates in (17) and (18).

(ii) The α-Hölder continuity of f tells us that for any $x \in \mathbb{R}^d$, there holds

$$\begin{aligned}|f_{P,h}(x) - f(x)| &= \left|\frac{1}{h^d}\int_{\mathbb{R}^d} K\left(\frac{\|x-x'\|}{h}\right) f(x')\,\mathrm{d}x' - f(x)\right| \\ &= \left|\int_{\mathbb{R}^d} K(\|x'\|) f(x+hx')\,\mathrm{d}x' - f(x)\right| \\ &= \left|\int_{\mathbb{R}^d} K(\|x'\|)\big(f(x+hx') - f(x)\big)\,\mathrm{d}x'\right| \\ &\lesssim \int_{\mathbb{R}^d} K(\|x'\|)\big(h\|x'\|\big)^\alpha \,\mathrm{d}x' \\ &\lesssim \int_{\mathbb{R}^d} K\big(\|x'\|_{\ell_2^d}\big) h^\alpha \|x'\|_{\ell_2^d}^\alpha \,\mathrm{d}x' \\ &\lesssim h^\alpha \int_0^\infty K(r) r^{\alpha+d-1}\,\mathrm{d}r \lesssim h^\alpha.\end{aligned}$$

We thus have completed the proof of Theorem 7. ■

The following lemma, which will be used several times in the sequel, supplies the key to the proof of Theorem 8.

Lemma 16 *Let the assumptions of Theorem 8 hold and $k_{x,h}$ be defined in (10). Then, for an arbitrary probability measure Q on $\mathbb{R}^d$, we have*

$$\int_{H_r} \mathbb{E}_Q k_{x,h}\,\mathrm{d}x \lesssim Q(H_{r/2}) + (h/r)^\beta.$$

Proof [of Lemma 16] For a positive constant t_0, we have

$$\begin{aligned}\int_{H_r} \mathbb{E}_P k_{x,h}\,\mathrm{d}x &= \int_{H_r}\int_{\mathbb{R}^d} h^{-d}K\left(\|x-x'\|/h\right)\,\mathrm{d}P(x')\,\mathrm{d}x\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(\|x\|)\mathbf{1}_{H_r}(hx+x')\,\mathrm{d}x\,\mathrm{d}P(x')\\ &= \int_{\mathbb{R}^d} K(\|x\|)\int_{\mathbb{R}^d}\mathbf{1}_{H_r}(hx+x')\,\mathrm{d}P(x')\,\mathrm{d}x\\ &\le \int_{B_{t_0}} K(\|x\|)\int_{\mathbb{R}^d}\mathbf{1}_{H_r}(hx+x')\,\mathrm{d}P(x')\mathrm{d}x + \int_{H_{t_0}} K(\|x\|)\,\mathrm{d}x.\end{aligned}$$

On the other hand, it is easy to see that $\mathbf{1}_{H_r}(hx+x')=1$ if and only if $\|hx+x'\|\ge r$. Now we set $t_0 := \frac{r}{2h}$. In this case, if we additionally have $x\in B_{t_0}$, then $\|x'\|\ge r-h\|x\|\ge r-ht_0 = r/2$. Therefore, we come to the following estimate

$$\begin{aligned}\int_{H_r}\mathbb{E}_P k_{x,h}\,\mathrm{d}x &\le \int_{B_{t_0}} K(\|x\|)P(H_{r/2})\,\mathrm{d}x + \int_{H_{t_0}} K(\|x\|)\,\mathrm{d}x\\ &\lesssim P(H_{r/2}) + \int_{t_0}^{\infty} K(t)t^{d-1}\,\mathrm{d}t\\ &\lesssim P(H_{r/2}) + \int_{t_0}^{\infty} K(t)t_0^{-\beta}t^{d+\beta-1}\,\mathrm{d}t\\ &\lesssim P(H_{r/2}) + t_0^{-\beta}\\ &\lesssim P(H_{r/2}) + (h/r)^{\beta}.\end{aligned} \tag{19}$$

We thus have shown the assertion of Lemma 16. ■

Proof [of Theorem 8] We decompose $\|f_{P,h}-f\|_1$ as follows

$$\begin{aligned}\|f_{P,h}-f\|_1 &= \int_{B_r}|f_{P,h}-f|\,\mathrm{d}x + \int_{H_r}|f_{P,h}-f|\,\mathrm{d}x\\ &\le \lambda^d(B_r)\|f_{P,h}-f\|_\infty + \int_{H_r}\mathbb{E}_P k_{x,h}\,\mathrm{d}x + \int_{H_r} f\,\mathrm{d}x\\ &\le r^d\|f_{P,h}-f\|_\infty + \int_{H_r}\mathbb{E}_P k_{x,h}\,\mathrm{d}x + P\left(H_r\right).\end{aligned} \tag{20}$$

Combining the two estimates in (20) and (19), we obtain the desired conclusion. ■

To prove Proposition 10, we need the following lemmas.

Lemma 17 *Let (X,d) and (Y,e) be metric spaces and $T: X\to Y$ be an α-Hölder continuous function with constant c. Then, for $A\subset X$ and all $\varepsilon>0$ we have*

$$\mathcal{N}(T(A),e,c\varepsilon^{\alpha})\le\mathcal{N}(A,d,\varepsilon).$$

Proof [of Lemma 17] Let $x_1, \ldots, x_n$ be an ε-net of A, that is, $A \subset \bigcup_{i=1}^n B_d(x_i, \varepsilon)$. For $i = 1, \cdots, n$, we set $y_i := T(x_i)$. Now, it only suffices to show that this gives a $c\varepsilon^\alpha$-net of $T(A)$.

In fact, supposing that $y \in T(B_d(x_i, \varepsilon))$, then there exists $x \in B_d(x_i, \varepsilon)$ such that $T(x) = y$. This implies

$$e(T(x), T(x_i)) \leq c d^\alpha(x, x_i) \leq c\varepsilon^\alpha.$$

Therefore, we have $T(B_d(x_i, \varepsilon)) \subset B_e(y_i, c\varepsilon^\alpha)$. That is, $y_1, \ldots, y_n$ is a $c\varepsilon^\alpha$-net of $T(A)$. This completes the proof of Lemma 17. ■

Remark 18 *We remark that when X is a Banach space with the norm $\|\cdot\|$, then for any $c > 0$ there holds*

$$\mathcal{N}(cA, \|\cdot\|, \varepsilon) = \mathcal{N}(A, \|\cdot\|, \varepsilon/c).$$

Lemma 19 *Let $\|\cdot\|'$ be another norm on $\mathbb{R}^d$. Then for all $\varepsilon \in (0, 1]$ we have*

$$\mathcal{N}(B_1, \|\cdot\|', \varepsilon) \lesssim \varepsilon^{-d}.$$

Proof [of Lemma 19] It is a straightforward conclusion of Proposition 1.3.1 in Carl and Stephani (1990) and Lemma 6.21 in Steinwart and Christmann (2008). ■

Lemma 20 *Let K be a d-dimensional smoothing kernel that satisfies Conditions (i) in Assumption 5. Let $h > 0$ be the bandwidth parameter, and $k_{x,h}$ be defined in (10) for any $x \in \mathbb{R}^d$. Then we have*

$$\sup_{y \in \mathbb{R}^d} |k_{x,h}(y) - k_{x',h}(y)| \leq \frac{c}{h^{\beta+d}} \|x - x'\|^\beta, \quad x, x' \in \mathbb{R}^d,$$

where c is a positive constant.

Proof [of Lemma 20] From the definition of $k_{x,h}$ and the fact that K is a d-dimensional β-Hölder continuous kernel, we have

$$\begin{aligned} |k_{x,h}(y) - k_{x',h}(y)| &= \frac{1}{h^d} \left| K\left(\frac{\|x - y\|}{h}\right) - K\left(\frac{\|x' - y\|}{h}\right) \right| \\ &\leq \frac{c}{h^d} \left| \frac{\|x - y\|}{h} - \frac{\|x' - y\|}{h} \right|^\beta \\ &\leq c h^{-(\beta+d)} \|x - x'\|^\beta, \end{aligned}$$

where c is a positive constant. The desired conclusion is thus obtained. ■

Proof [of Proposition 10] Lemma 20 reveals that $\mathcal{K}_{h,r}$ is the image of Hölder continuous map $B_r \to L_\infty(\mathbb{R}^d)$ with the constant $ch^{-(\beta+d)}$. By Lemmas 17 and 19 we obtain

$$\begin{aligned}\mathcal{N}(\mathcal{K}_{h,r}, \|\cdot\|_\infty, \varepsilon) &\le \mathcal{N}\left(B_r, \|\cdot\|, \left(\frac{\varepsilon h^{\beta+d}}{c}\right)^{1/\beta}\right) \\ &= \mathcal{N}\left(B_1, \|\cdot\|, \left(\frac{\varepsilon h^{\beta+d}}{cr^\beta}\right)^{1/\beta}\right) \\ &\le c'\left(\frac{\varepsilon h^{\beta+d}}{r^\beta}\right)^{-d/\beta},\end{aligned}$$

where c' is a constant independent of ε. This completes the proof of Proposition 10. ■

The following Bernstein-type exponential inequality, which was developed recently in Hang and Steinwart (2016), will serve as one of the main ingredients in the consistency and convergence analysis of the kernel density estimator (7). It can be stated in the following general form:

Theorem 21 (Bernstein Inequality (Hang and Steinwart, 2016)) *Assume that $\mathcal{X} := (X_n)_{n\ge1}$ is an X-valued stationary geometrically (time-reversed) $\mathcal{C}$-mixing process on $(\Omega, \mathcal{A}, \mu)$ with $\|\cdot\|_{\mathcal{C}}$ be defined by (3) for some semi-norm $|||\cdot|||$ satisfying Condition (ii) in Assumption 1, and $P := \mu_{X_1}$. Moreover, let $g : X \to \mathbb{R}$ be a function such that $g \in \mathcal{C}(X)$ with $\mathbb{E}_P g = 0$ and assume that there exist some $A > 0$, $B > 0$, and $\sigma \ge 0$ such that $|||g||| \le A$, $\|g\|_\infty \le B$, and $\mathbb{E}_P g^2 \le \sigma^2$. Then, for all $\tau > 0$, $k \in \mathbb{N}$, and*

$$n \ge n_0 := \max\left\{\min\left\{m \ge 3 : m \ge \left(\frac{808c_0(3A+B)}{B}\right)^{\frac{1}{k}} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4\right\}, e^{\frac{k+1}{b}}\right\},$$

with probability μ at least $1 - 4e^{-\tau}$, there holds

$$\left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right| \le \sqrt{\frac{8(\log n)^{\frac{2}{\gamma}}\sigma^2\tau}{n}} + \frac{8(\log n)^{\frac{2}{\gamma}}B\tau}{3n}.$$

Proof [of Theorem 11] Let the notations $k_{x,h}$ and $\widetilde{k}_{x,h}$ be defined in (10) and (11), respectively, that is, $k_{x,h} := h^{-d}K(\|x - \cdot\|/h)$, and $\widetilde{k}_{x,h} := k_{x,h} - \mathbb{E}_P k_{x,h}$. We first assume that $x \in \mathbb{R}^d$ is fixed and then estimate $\mathbb{E}_D f_{x,h}$ by using Bernstein's inequality in Theorem 21. For this purpose, we shall verify the following conditions: Obviously, we have $\mathbb{E}_P \widetilde{k}_{x,h} = 0$. Moreover, simple estimates yield

$$\|\widetilde{k}_{x,h}\|_\infty \le 2\|k_{x,h}\|_\infty \le 2h^{-d}\|K\|_\infty \le 2h^{-d}K(0)$$

and

$$\mathbb{E}_P \widetilde{k}_{x,h}^2 \le \mathbb{E}_P k_{x,h}^2 = \int_{\mathbb{R}^d} k_{x,h}^2(x')\mathrm{d}P(x').$$

Finally, the first condition in Assumption 1 and Condition (iii) in Assumption 5 imply

$$\|\!|\widetilde{k}_{x,h}|\!\| \le \|\!|k_{x,h}|\!\| \le h^{-d} \sup_{x\in\mathbb{R}^d} \|\!|K\left(\|x-\cdot\|/h\right)|\!\| \le h^{-d}\varphi(h).$$

Now we can apply the Bernstein-type inequality in Theorem 21 and obtain that for $n \ge n_1$, for any fixed $x \in \mathbb{R}^d$, with probability μ at most $4e^{-\tau}$, there holds

$$|\mathbb{E}_D \widetilde{k}_{x,h}| \ge \sqrt{\frac{8\tau(\log n)^{2/\gamma}\int_{\mathbb{R}^d} k_{x,h}^2(x')\mathrm{d}P(x')}{n}} + \frac{16\tau(\log n)^{2/\gamma}K(0)}{3h^d n}, \tag{21}$$

where

$$n_1 := \max\left\{\min\left\{m \ge 3 : m \ge \left(\frac{808c_0(3h^{-d}\varphi(h)+K(0))}{2K(0)}\right)^{\frac{1}{d+1}} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4\right\}, e^{\frac{d+1}{b}}\right\}. \tag{22}$$

Consider the function set $\widetilde{\mathcal{K}}_{h,r} := \{\widetilde{k}_{x,h} : x \in B_r\}$. We choose $y_1, \ldots, y_m \in B_r$ such that $\{k_{y_1,h}, \ldots, k_{y_m,h}\}$ is a minimal $\varepsilon/2$-net of $\mathcal{K}_{h,r} = \{k_{x,h} : x \in B_r\}$ with respect to $\|\cdot\|_\infty$. Noticing the following relation

$$\|\widetilde{k}_{x,h} - \widetilde{k}_{y_j,h}\|_\infty \le 2\|k_{x,h} - k_{y_j,h}\|_\infty \le \varepsilon,$$

we know that $\{\widetilde{k}_{y_1,h}, \ldots, \widetilde{k}_{y_m,h}\}$ is an ε-net of $\widetilde{\mathcal{K}}_{h,r}$ with respect to $\|\cdot\|_\infty$. Note that here we have $m = \mathcal{N}(\mathcal{K}_{h,r}, \|\cdot\|_\infty, \frac{\varepsilon}{2})$, since the net is minimal. From Proposition 10, we know that there exists a positive constant c independent of ε such that $\log m \le c \log \frac{r}{h\varepsilon}$. From the estimate in (21) and a union bound argument, with probability μ at least $1 - 4me^{-\tau}$, the following estimate holds

$$\sup_{j=1,\ldots,m} |\mathbb{E}_D \widetilde{k}_{y_j,h}| \le \sqrt{\frac{8\tau(\log n)^{2/\gamma}\int_{\mathbb{R}^d} k_{y_j,h}^2(x')\mathrm{d}P(x')}{n}} + \frac{16\tau(\log n)^{2/\gamma}K(0)}{h^d n}.$$

By a simple variable transformation, we see that with probability μ at least $1 - e^{-\tau}$, there holds

$$\begin{aligned}\sup_{j=1,\ldots,m} |\mathbb{E}_D \widetilde{k}_{y_j,h}| \le & \sqrt{\frac{8(\log n)^{2/\gamma}\int_{\mathbb{R}^d} k_{y_j,h}^2(x')\mathrm{d}P(x')(\tau+\log(4m))}{n}} \\ & + \frac{16(\log n)^{2/\gamma}K(0)(\tau+\log(4m))}{h^d n}.\end{aligned}$$

Recalling that $\{k_{y_1,h}, \ldots, k_{y_m,h}\}$ is an $\varepsilon/2$-net of $\mathcal{K}_{h,r}$, this implies that, for any $x \in B_r$, there exists y_j such that $\|k_{x,h} - k_{y_j,h}\|_\infty \le \varepsilon/2$. Then we have

$$\begin{aligned}\left||\mathbb{E}_D \widetilde{k}_{x,h}| - |\mathbb{E}_D \widetilde{k}_{y_j,h}|\right| &\le \left|\mathbb{E}_D \widetilde{k}_{x,h} - \mathbb{E}_D \widetilde{k}_{y_j,h}\right| \\ &\le |\mathbb{E}_D k_{x,h} - \mathbb{E}_D k_{y_j,h}| + |\mathbb{E}_P k_{x,h} - \mathbb{E}_P k_{y_j,h}| \\ &\le \|k_{x,h} - k_{y_j,h}\|_{L_1(D)} + \|k_{x,h} - k_{y_j,h}\|_{L_1(P)} \\ &\le \varepsilon,\end{aligned}$$

and consequently

$$|\mathbb{E}_D \widetilde{k}_{x,h}| \le |\mathbb{E}_D \widetilde{k}_{y_j,h}| + \varepsilon. \tag{23}$$

By setting $a := 8(\log n)^{2/\gamma}(\tau + \log(4m))/n$, we have

$$\begin{aligned}\left|\sqrt{a\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')} - \sqrt{a\int_{\mathbb{R}^d} k_{y_j,h}^2(x')\,\mathrm{d}P(x')}\right| &= \left|\|\sqrt{a}k_{x,h}\|_{L_2(P)} - \|\sqrt{a}k_{y_j,h}\|_{L_2(P)}\right| \\ &\le \sqrt{a}\|k_{x,h} - k_{y_j,h}\|_{L_2(P)} \\ &\le \sqrt{a}\varepsilon/2.\end{aligned}$$

This together with inequality (23) implies that for any $x \in B_r$, there holds

$$\begin{aligned}|\mathbb{E}_D \widetilde{k}_{x,h}| &\le |\mathbb{E}_D \widetilde{k}_{y_j,h}| + 2\varepsilon \\ &\le \sqrt{a\int_{\mathbb{R}^d} k_{y_j,h}^2(x')\,\mathrm{d}P(x')} + \frac{2aK(0)}{h^d} + \varepsilon \\ &\le \sqrt{a\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')} + \frac{\sqrt{a}\varepsilon}{2} + \frac{2aK(0)}{h^d} + \varepsilon.\end{aligned}$$

Consequently we have

$$\begin{aligned}\int_{B_r} |\mathbb{E}_D \widetilde{k}_{x,h}|\,\mathrm{d}x \le \int_{B_r} \sqrt{a\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')}\,\mathrm{d}x \\ + r^d\lambda^d(B_1)\cdot\frac{2aK(0)}{h^d} + r^d\lambda^d(B_1)(\sqrt{a}/2+1)\varepsilon.\end{aligned}$$

Now recall that for $E \subset \mathbb{R}^d$ and $g : E \to \mathbb{R}$, Hölder's inequality implies

$$\|g\|_{\frac{1}{2}} = \left(\int_{\mathbb{R}^d} |\mathbf{1}_E|^{\frac{1}{2}}|g|^{\frac{1}{2}}\,\mathrm{d}x\right)^2 \le \int_{\mathbb{R}^d} |\mathbf{1}_E|\,\mathrm{d}x \int_{\mathbb{R}^d} |g|\,\mathrm{d}x = \mu(E)\cdot\|g\|_1.$$

This tells us that

$$\int_{B_r} \sqrt{a\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')}\,\mathrm{d}x \le \sqrt{\mu(B_r)}\cdot\sqrt{\int_{B_r} a\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')\,\mathrm{d}x}.$$

Moreover, there holds

$$\begin{aligned}\int_{B_r}\int_{\mathbb{R}^d} k_{x,h}^2(x')\,\mathrm{d}P(x')\,\mu(\mathrm{d}x) &= \int_{\mathbb{R}^d}\int_{B_r} h^{-2d}K^2\left(\|x-x'\|/h\right)\,\mathrm{d}x\,\mathrm{d}P(x') \\ &\le \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h^{-2d}K^2\left(\|x\|/h\right)\,\mathrm{d}x\,\mathrm{d}P(x') \\ &= h^{-d}\int_{\mathbb{R}^d} K^2(\|x\|)\,\mathrm{d}x \\ &\le K(0)h^{-d}.\end{aligned}$$

We now set $\varepsilon = \frac{1}{n}$ and obtain $\log(4m) \le c \log \frac{nr}{h}$. Thus we have

$$\begin{aligned}\int_{B_r} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x &\lesssim \sqrt{\frac{(\log n)^{2/\gamma} r^d (\tau + \log(4m))}{h^d n}} + \frac{(\log n)^{2/\gamma} r^d (\tau + \log(4m))}{h^d n} \\ &\quad + \sqrt{\frac{(\log n)^{2/\gamma} (\tau + \log(4m))}{n}} \cdot \frac{r^d}{n} \\ &\lesssim \sqrt{\frac{(\log n)^{2/\gamma} r^d (\tau + \log \frac{nr}{h})}{h^d n}} + \frac{(\log n)^{2/\gamma} r^d (\tau + \log \frac{nr}{h})}{h^d n}.\end{aligned} \tag{24}$$

Now we need to estimate the corresponding integral over H_r. By definition we have

$$\int_{H_r} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x \le \int_{H_r} \mathbb{E}_D k_{x,h} \, \mathrm{d}x + \int_{H_r} \mathbb{E}_P k_{x,h} \, \mathrm{d}x.$$

From Lemma 16 we obtain

$$\int_{H_r} \mathbb{E}_D k_{x,h} \, \mathrm{d}x \lesssim D(H_{r/2}) + \left(\frac{h}{r}\right)^\beta = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{H_{r/2}}(x_i) + \left(\frac{h}{r}\right)^\beta,$$

and

$$\int_{H_r} \mathbb{E}_P k_{x,h} \, \mathrm{d}x \lesssim P(H_{r/2}) + \left(\frac{h}{r}\right)^\beta.$$

Since $r \ge 1$, we can construct a function g with $\mathbf{1}_{H_{r/2}} \le g \le \mathbf{1}_{H_{r/4}}$ and there exists a function $\psi(r)$ such that $\|\|g\|\| \le \psi(r)$. Applying Bernstein inequality in Theorem 21 with respect to this function g, it is easy to see that when $n \ge n_2$, with probability μ at least $1 - 2e^{-\tau}$, there holds

$$\mathbb{E}_D g - \mathbb{E}_P g \le \sqrt{\frac{8\tau(\log n)^{2/\gamma}}{n}} + \frac{8\tau(\log n)^{2/\gamma}}{3n},$$

where

$$n_2 := \max\left\{ \min\left\{ m \ge 3 : m^2 \ge 808 c_0 (3\psi(r) + 1) \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \ge 4 \right\}, e^{\frac{3}{b}} \right\}.$$

This implies that with probability μ at least $1 - 2e^{-\tau}$, there holds

$$\begin{aligned} D(H_{r/2}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{H_{r/2}}(x_i) &\le \mathbb{E}_D g \\ &\le \mathbb{E}_P g + \sqrt{\frac{8\tau(\log n)^{2/\gamma}}{n}} + \frac{8\tau(\log n)^{2/\gamma}}{3n} \\ &\le \mathbb{E}_P \mathbf{1}_{H_{r/4}}(x_i) + \sqrt{\frac{8\tau(\log n)^{2/\gamma}}{n}} + \frac{8\tau(\log n)^{2/\gamma}}{3n} \end{aligned}$$

and consequently we obtain

$$\int_{H_r} |\mathbb{E}_D \widetilde{k}_{x,h}| \, \mathrm{d}x \lesssim P(H_{r/4}) + \sqrt{\frac{32\tau(\log n)^{2/\gamma}}{n}} + \left(\frac{h}{r}\right)^{\beta}. \tag{25}$$

By combining estimates in (24) and (25), and taking $n_0 = \max\{n_1, n_2\}$, we have accomplished the proof of Theorem 11. ■

Remark 22 *Let us briefly discuss the choice of the function $\psi(r)$ in the proof of Theorem 11. For example, in the case $\mathcal{C}(X) = \mathrm{Lip}(\mathbb{R})$, we can choose*

$$g(x) := \begin{cases} 1, & \text{for } |x| > r, \\ 0, & \text{for } |x| < r/4, \\ -\frac{4x}{3r} - \frac{1}{3}, & \text{for } -r \le x \le -r/4, \\ \frac{4x}{3r} - \frac{1}{3}, & \text{for } r/4 \le x \le r. \end{cases}$$

Then we have $|||g||| \le \frac{4}{3r} \le 4/3$ and therefore, n_2 is well-defined. Moreover, it is easily seen that even for smoother underlying functions classes like C^1 we can construct a function g such that $|||g||| < \infty$.

Proof [of Theorem 12] Recalling the definitions of $k_{x,h}$ and $\widetilde{k}_{x,h}$ given in (10) and (11), we have

$$\|f_{D,h} - f_{P,h}\|_\infty = \sup_{x\in\Omega} |\mathbb{E}_D \widetilde{k}_{x,h}|.$$

To prove the assertion, we first estimate $\mathbb{E}_D f_{x,h}$ for fixed $x \in \mathbb{R}^d$ using the Bernstein inequality in Theorem 21. For this purpose, we first verify the following conditions: Obviously, we have $\mathbb{E}_P \widetilde{k}_{x,h} = 0$. Then, simple estimates imply

$$\|\widetilde{k}_{x,h}\|_\infty \le 2\|k_{x,h}\|_\infty \le 2h^{-d}\|K\|_\infty \le 2h^{-d}K(0)$$

and

$$\mathbb{E}_P \widetilde{k}^2_{x,h} \le \mathbb{E}_P k^2_{x,h} = h^{-d} \int_{\mathbb{R}^d} K^2\left(\|x - x'\|/h\right) f(x') h^{-d} \, \mathrm{d}x' \lesssim \|f\|_\infty h^{-d}.$$

Finally, the first condition in Assumption 1 and Condition (iii) in Assumption 5 yield

$$|||\widetilde{k}_{x,h}||| \le |||k_{x,h}||| \le h^{-d} \sup_{x\in\mathbb{R}^d} |||K\left(\|x - \cdot\|/h\right)||| \le h^{-d}\varphi(h).$$

Therefore, we can apply the Bernstein inequality in Theorem 21 and obtain that for $n \ge n_0^*$, for any fixed $x \in \mathbb{R}^d$, with probability μ at least $1 - 4e^{-\tau}$, there holds

$$|\mathbb{E}_D \widetilde{k}_{x,h}| \lesssim \sqrt{\frac{\tau\|f\|_\infty(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)\tau(\log n)^{2/\gamma}}{3h^d n}, \tag{26}$$

where

$$n_0^* := \max\left\{\min\left\{m \geq 3 : m \geq \left(\frac{808c_0(3h^{-d}\varphi(h) + K(0))}{2K(0)}\right)^{\frac{1}{d+1}} \text{ and } \frac{m}{(\log m)^{\frac{2}{\gamma}}} \geq 4\right\}, e^{\frac{d+1}{b}}\right\}. \tag{27}$$

Let us consider the following function set

$$\mathcal{K}'_{h,r_0} := \{\widetilde{k}_{x,h} : x \in B_{r_0}\}$$

and choose $y_1, \ldots, y_m \in B_{r_0}$ such that $\{k_{y_1,h}, \ldots, k_{y_m,h}\}$ is a minimal $\varepsilon/2$-net of $\mathcal{K}_{h,r_0}$ with respect to $\|\cdot\|_\infty$ and $m = \mathcal{N}(\mathcal{K}_{h,r_0}, \|\cdot\|_\infty, \frac{\varepsilon}{2})$. As in the proof of Theorem 11, one can show that $\widetilde{k}_{y_1,h}, \ldots, \widetilde{k}_{y_m,h}$ is an ε-net of $\mathcal{K}'_{h,r_0}$. Again from Proposition 10 we know that there holds $\log(4m) \lesssim \log \frac{r_0}{h\varepsilon}$. This in connection with (26) implies that the following union bound

$$\sup_{j=1,\ldots,m} |\mathbb{E}_D \widetilde{k}_{y_j,h}| \lesssim \sqrt{\frac{\|f\|_\infty(\tau + \log(4m))(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)(\tau + \log(4m))(\log n)^{2/\gamma}}{h^d n}$$

holds with probability μ at least $1 - e^{-\tau}$. For any $x \in B_{r_0}$, there exists a y_j such that $\|k_{x,h} - k_{y_j,h}\|_\infty \leq \varepsilon$. Then we have

$$\begin{aligned}
\big||\mathbb{E}_D \widetilde{k}_{x,h}| - |\mathbb{E}_D \widetilde{k}_{y_j,h}|\big| &\leq |\mathbb{E}_D \widetilde{k}_{x,h} - \mathbb{E}_D \widetilde{k}_{y_j,h}| \\
&\leq |\mathbb{E}_D k_{x,h} - \mathbb{E}_D k_{y_j,h}| + |\mathbb{E}_P k_{x,h} - \mathbb{E}_P k_{y_j,h}| \\
&\leq \|k_{x,h} - k_{y_j,h}\|_{L_1(D)} + \|k_{x,h} - k_{y_j,h}\|_{L_1(P)} \\
&\leq \varepsilon,
\end{aligned}$$

and consequently with probability μ at least $1 - e^{-\tau}$, there holds

$$\begin{aligned}
|\mathbb{E}_D \widetilde{k}_{x,h}| &\leq |\mathbb{E}_D \widetilde{k}_{y_j,h}| + \varepsilon \\
&\lesssim \sqrt{\frac{\|f\|_\infty(\tau + \log(4m))(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)(\tau + \log(4m))(\log n)^{2/\gamma}}{h^d n} + \varepsilon
\end{aligned}$$

for any $x \in B_{r_0}$. By setting $\varepsilon = \frac{1}{n}$, we obtain $\log(4m) \lesssim \log \frac{nr_0}{h}$. Thus, with probability μ at least $1 - e^{-\tau}$, we have

$$\begin{aligned}
|\mathbb{E}_D \widetilde{k}_{x,h}| &\lesssim \sqrt{\frac{\|f\|_\infty(\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)(\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n} + \frac{1}{n} \\
&\lesssim \sqrt{\frac{\|f\|_\infty(\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n}} + \frac{K(0)(\tau + \log(\frac{nr_0}{h}))(\log n)^{2/\gamma}}{h^d n}.
\end{aligned}$$

By taking the supremum of the left hand side of the above inequality over x, we complete the proof of Theorem 12. ■

Proof [of Theorem 13] Without loss of generality, we assume that $h_n \le 1$. Since $h_n \to 0$, Theorem 7 implies that $\|f_{P,h} - f\|_1 \le \varepsilon$. We set

$$r_n := \left(\frac{nh_n^d}{(\log n)^{(2+2\gamma)/\gamma}}\right)^{1/d} \to \infty \tag{28}$$

and we can also assume w.l.o.g that $r_n \ge 2$. Moreover, there exists a constant n_1' such that

$$P(H_{r_n/2}) \le \varepsilon, \ \forall\, n \ge n_1'.$$

For any $0 < \delta < 1$, we select $\tau := \log(1/\delta)$. Then there exists a constant n_2' such that $\log \frac{nr_n}{h_n} \ge \tau$ for all $n \ge n_2'$. On the other hand, with the above choice of r_n, we have

$$\log \frac{nr_n}{h_n} \le \log\left(\frac{n^{1/d}h_n}{(\log n)^{(2+2\gamma)/\gamma}} \cdot \frac{n}{h_n}\right) \le (1 + d^{-1}) \log n \lesssim \log n.$$

Thus, for all $n \ge \max\{n_1', n_2'\}$, we have

$$\frac{(\log n)^{2/\gamma} r_n^d \log(\frac{nr_n}{h_n})}{nh_n^d} \lesssim \frac{(\log n)^{2/\gamma} r_n^d \log n}{nh_n^d} = \frac{1}{\log n} \to 0.$$

Thus, following from Theorem 11, when n is sufficient large, for any $\varepsilon > 0$, with probability μ at least $1 - 3\delta$, there holds

$$\|f_{D,h_n} - f\|_1 \lesssim \varepsilon.$$

Therefore, with properly chosen δ, one can show that f_{D,h_n} converges to f under L_1-norm almost surely. We have completed the proof of Theorem 13. ■

Proof [of Theorem 14] *(i)* Combining the estimates in Theorem 11 and Theorem 8, we know that with probability μ at least $1 - 2e^{-\tau}$, there holds

$$\begin{aligned}
\|f_{D,h} - f\|_1 &\lesssim \sqrt{\frac{(\log n)^{2/\gamma} r^d (\tau + \log(\frac{nr}{h_n}))}{h_n^d n}} + \frac{(\log n)^{2/\gamma} r^d (\tau + \log(\frac{nr}{h_n}))}{h_n^d n} \\
&\quad + \frac{\tau(\log n)^{2/\gamma}}{n} + P\big(H_r\big) + r^d h_n^\alpha + \left(\frac{h_n}{r}\right)^\beta \\
&\lesssim \sqrt{\frac{(\log n)^{2/\gamma} r^d (\tau + \log(\frac{nr}{h_n}))}{h_n^d n}} + \frac{\tau(\log n)^{2/\gamma}}{n} + P\big(H_r\big) + r^d h_n^\alpha + \left(\frac{h_n}{r}\right)^\beta.
\end{aligned}$$

Let $\tau := \log n$ and later we will see from the choices of h_n and r_n that there exists some constant c such that $\log(\frac{nr}{h})$ can be bounded by $c \log n$. Therefore, with probability μ at least $1 - \frac{1}{n}$ there holds

$$\begin{aligned}
\|f_{D,h} - f\|_1 &\lesssim \sqrt{\frac{r^d (\log n)^{(2+\gamma)/\gamma}}{h_n^d n}} + r^{-\eta d} + r^d h_n^\alpha \\
&\lesssim r^d \left(\frac{\log n}{nr^d}\right)^{\frac{\alpha}{2\alpha+d}} + r^{-\eta d} \\
&\lesssim \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{\alpha\eta}{(1+\eta)(2\alpha+d)-\alpha}},
\end{aligned}$$

by choosing

$$h_n = \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{1+\eta}{(1+\eta)(2\alpha+d)-\alpha}} \quad \text{and} \quad r := r_n = \left(\frac{n}{(\log n)^{(2+\gamma)/\gamma}}\right)^{\frac{\alpha}{d(1+\eta)(2\alpha+d)-\alpha d}}.$$

(ii) Similar to case *(i)*, one can show that with probability μ at least $1 - \frac{1}{n}$ there holds

$$\begin{aligned}\|f_{D,h} - f\|_1 &\lesssim \sqrt{\frac{r^d(\log n)^{(2+\gamma)/\gamma}}{h_n^d n}} + e^{-ar^\eta} + r^d h_n^\alpha \\ &\lesssim r^d\left(\frac{(\log n)^{(2+\gamma)/\gamma}}{nr^d}\right)^{\frac{\alpha}{2\alpha+d}} + e^{-ar^\eta} \\ &\lesssim \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{d}{\eta}\cdot\frac{\alpha+d}{2\alpha+d}},\end{aligned}$$

by choosing

$$h_n = \left(\frac{(\log n)^{(2+\gamma)/\gamma}}{n}\right)^{\frac{1}{2\alpha+d}} (\log n)^{-\frac{d}{\eta}\cdot\frac{1}{2\alpha+d}} \quad \text{and} \quad r_n = (\log n)^{\frac{1}{\eta}}.$$

(iii) From Theorem 8 we see that with confidence $1 - \frac{1}{n}$, there holds

$$\|f_{D,h} - f_{P,h}\|_1 \lesssim \sqrt{\frac{r_0^d(\log n)^{(2+\gamma)/\gamma}}{h_n^d n}} + h_n^\alpha \lesssim \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{\alpha}{2\alpha+d}},$$

where h_n is chosen as

$$h_n = \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{1}{2\alpha+d}}.$$

The proof of Theorem 14 is completed. ■

Proof [of Theorem 15] The desired estimate is an easy consequence if we combine the estimates in Theorem 12 and Theorem 7 *(ii)* and choose

$$h_n = \left((\log n)^{(2+\gamma)/\gamma}/n\right)^{\frac{1}{2\alpha+d}}.$$

We omit the details of the proof here. ■

6. Conclusion

In the present paper, we studied the kernel density estimation problem for dynamical systems admitting a unique invariant Lebesgue density by using the $\mathcal{C}$-mixing coefficient to measure the dependence among observations. The main results presented in this paper are

the consistency and convergence rates of the kernel density estimator in the sense of L_1-norm and L_∞-norm. With properly chosen bandwidth, we showed that the kernel density estimator is universally consistent. Under mild assumptions on the kernel function and the density function, we established convergence rates for the estimator. For instance, when the density function is bounded and compactly supported, both L_1-norm and L_∞-norm convergence rates with the same order can be achieved for general geometrically time-reversed $\mathcal{C}$-mixing dynamical systems. The convergence mentioned here is of type "with high probability" due to the use of a Bernstein-type exponential inequality and this makes the present study different from the existing related studies. We also discussed the model selection problem of the kernel density estimation in the dynamical system context by carrying out numerical experiments.

Acknowledgments

The authors are grateful to Professor László Györfi for helpful comments. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants. Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants. IWT: projects: SBO POM (100031); PhD/Postdoc grants. iMinds Medical Information Technologies SBO 2014. Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). The corresponding author is Yunlong Feng.

References

Marian Anghel and Ingo Steinwart. Forecasting the evolution of dynamical systems from noisy observations. *arXiv preprint arXiv:0707.4146*, 2007.

Viviane Baladi. *Positive Transfer Operators and Decay of Correlations*, volume 16. World Scientific, 2000.

Delphine Blanke, Denis Bosq, and Dominique Guégan. Modelization and nonparametric estimation for dynamical systems with noise. *Statistical Inference for Stochastic Processes*, 6(3):267–290, 2003.

Denis Bosq and Dominique Guégan. Nonparametric estimation of the chaotic function and the invariant measure of a dynamical system. *Statistics & Probability Letters*, 25(3):201–212, 1995.

Adrian W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2(2):107–144, 2005.

Ricardo Cao, Antonio Cuevas, and Wensceslao González Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, 17(2):153–176, 1994.

Bernd Carl and Irmtraud Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

Chih-Kang Chu and James S. Marron. Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, 19(4):1906–1918, 1991.

Marc Deisenroth and Shakir Mohamed. Expectation propagation in Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2609–2617. NIPS Foundation, 2012.

Luc Devroye. The double kernel method in density estimation. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 25(4):533–580, 1989.

Luc Devroye. Universal smoothing factor selection in density estimation: theory and practice. *Test*, 6(2):223–320, 1997.

Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L_1 View*, volume 119. John Wiley & Sons Incorporated, 1985.

Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *The Annals of Statistics*, 25(6):2626–2637, 1997.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media, 2001.

Paul Eggermont and Vince LaRiccia. *Maximum Penalized Likelihood Estimation: Volume I: Density Estimation*. Springer, New York, 2001.

Gerald B. Folland. *Real Analysis*. John Wiley & Sons, New York, 1999.

Theo Gasser, Hans-Georg Müller, and Volker Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 47(2):238–252, 1985.

László Györfi. Strong consistent density estimate from ergodic sample. *Journal of Multivariate Analysis*, 11(1):81–84, 1981.

László Györfi and Gábor Lugosi. Kernel density estimation from ergodic sample is not universally consistent. *Computational statistics & data analysis*, 14(4):437–442, 1992.

László Györfi and Elias Masry. The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples. *IEEE Transactions on Information Theory*, 36(3):531–539, 1990.

Peter Hall, Soumendra Nath Lahiri, and Jörg Polzehl. On bandwidth choice in nonparametric regression with both short-and long-range dependent errors. *The Annals of Statistics*, 23(6): 1921–1936, 1995.

Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics, in press*, 2016. URL http://www.e-publications.org/ims/submission/AOS/user/submissionFile/22219?confirm=ec9efb84.

Hanyuan Hang, Yunlong Feng, Ingo Steinwart, and Johan A.K. Suykens. Learning theory estimates with observations from general stationary stochastic processes. *Neural Computation, in press*, 2016. URL http://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00870#.V4ZTVtx95pi.

Jeffrey D. Hart and Philippe Vieu. Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, 18(2):873–890, 1990.

Jeffrey D. Hart and Thomas E. Wehrly. Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81(396):1080–1088, 1986.

Michael C. Jones, James S. Marron, and Simon J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996a.

Michael C. Jones, James S. Marron, and Simon J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11(3):337–381, 1996b.

Rafail Z. Khas′minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

Andrzej Lasota and Michael C. Mackey. *Probabilistic Properties of Deterministic Systems*. Cambridge University Press, 1985.

Andrzej Lasota and James A. Yorke. On the existence of invariant measures for piecewise monotonic transformations. *Transactions of the American Mathematical Society*, 186:481–488, 1973.

Elias Masry. Probability density estimation from sampled data. *Information Theory, IEEE Transactions on*, 29(5):696–709, 1983.

Elias Masry. Recursive probability density estimation for weakly dependent stationary processes. *Information Theory, IEEE Transactions on*, 32(2):254–267, 1986.

Elias Masry and László Györfi. Strong consistency and rates for recursive probability density estimators of stationary processes. *Journal of Multivariate Analysis*, 22(1):79–93, 1987.

Véronique Maume-Deschamps. Exponential inequalities and functional estimation for weak dependent data: applications to dynamical systems. *Stochastics and Dynamics*, 6(4):535–560, 2006.

Kevin McGoff, Sayan Mukherjee, Andrew Nobel, and Natesh Pillai. Consistency of maximum likelihood estimation for some dynamical systems. *The Annals of Statistics*, 43(1):1–29, 2015a.

Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: a review. *Statistics Surveys*, 9:209–252, 2015b.

Byeong U. Park and James S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Clémentine Prieur. Density estimation for one-dimensional dynamical systems. *ESAIM: Probability and Statistics*, 5:51–76, 2001.

Peter M. Robinson. Nonparametric estimators for time series. *Journal of Time Series Analysis*, 4 (3):185–207, 1983.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

David W. Scott and George R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.

Simon J. Sheather and Michael C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):683–690, 1991.

Ingo Steinwart and Marian Anghel. Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *The Annals of Statistics*, 37(2):841–875, 2009.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Charles J. Stone. Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics*, pages 393–406. Academic Press, New York, 1983.

Johan A.K. Suykens and Joos Vandewalle. Recurrent least squares support vector machines. *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, 47(7):1109–1114, 2000.

Johan A.K. Suykens, Joos Vandewalle, and Bart De Moor. *Artificial Neural Networks for Modelling and Control of Non-Linear Systems*. Springer Science & Business Media, 1995.

Johan A.K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

George R. Terrell. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477, 1990.

Lanh Tat Tran. The L_1 convergence of kernel density estimates under dependence. *The Canadian Journal of Statistics*, 17(2):197–208, 1989a.

Lanh Tat Tran. Recursive density estimation under dependence. *Information Theory, IEEE Transactions on*, 35(5):1103–1108, 1989b.

Matt P. Wand and Chris M. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1994.

Qiwei Yao and Howell Tong. Cross-validatory bandwidth selections for regression estimation based on dependent data. *Journal of Statistical Planning and Inference*, 68(2):387–415, 1998.

Bin Yu. Density estimation in the L^∞-norm for dependent data with applications to the Gibbs sampler. *The Annals of Statistics*, 21(2):711–735, 1993.

Onno Zoeter and Tom Heskes. Change point problems in linear dynamical systems. *The Journal of Machine Learning Research*, 6:1999–2026, 2005.