Bernoulli **22**(3), 2016, 1839–1893
DOI: 10.3150/15-BEJ713

On the optimal estimation of probability measures in weak and strong topologies

BHARATH SRIPERUMBUDUR

Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.
E-mail: bks18@psu.edu

Given random samples drawn i.i.d. from a probability measure $\mathbb{P}$ (defined on say, $\mathbb{R}^d$), it is well-known that the empirical estimator is an optimal estimator of $\mathbb{P}$ in weak topology but not even a consistent estimator of its density (if it exists) in the strong topology (induced by the total variation distance). On the other hand, various popular density estimators such as kernel and wavelet density estimators are optimal in the strong topology in the sense of achieving the minimax rate over all estimators for a Sobolev ball of densities. Recently, it has been shown in a series of papers by Giné and Nickl that these density estimators on $\mathbb{R}$ that are optimal in strong topology are also optimal in $\|\cdot\|_{\mathcal{F}}$ for certain choices of $\mathcal{F}$ such that $\|\cdot\|_{\mathcal{F}}$ metrizes the weak topology, where $\|\mathbb{P}\|_{\mathcal{F}} := \sup\{\int f\,\mathrm{d}\mathbb{P}\colon\ f \in \mathcal{F}\}$. In this paper, we investigate this problem of optimal estimation in weak and strong topologies by choosing $\mathcal{F}$ to be a unit ball in a reproducing kernel Hilbert space (say $\mathcal{F}_H$ defined over $\mathbb{R}^d$), where this choice is both of theoretical and computational interest. Under some mild conditions on the reproducing kernel, we show that $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology and the kernel density estimator (with L^1 optimal bandwidth) estimates $\mathbb{P}$ at dimension independent optimal rate of $n^{-1/2}$ in $\|\cdot\|_{\mathcal{F}_H}$ along with providing a uniform central limit theorem for the kernel density estimator.

Keywords: adaptive estimation; bounded Lipschitz metric; exponential inequality; kernel density estimator; Rademacher chaos; reproducing kernel Hilbert space; smoothed empirical processes; total variation distance; two-sample test; uniform central limit theorem; U-processes

1. Introduction

Let $X_1, \ldots, X_n$ be independent random variables distributed according to a Borel probability measure $\mathbb{P}$ defined on a separable metric space $\mathcal{X}$ with $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ being the empirical measure induced by them. It is well known that $\mathbb{P}_n$ is a consistent estimator of $\mathbb{P}$ in weak sense as $n \to \infty$, that is, for every bounded continuous real-valued function f on $\mathcal{X}$, $\int f\,\mathrm{d}\mathbb{P}_n \stackrel{\text{a.s.}}{\to} \int f\,\mathrm{d}\mathbb{P}$ as $n \to \infty$, written as $\mathbb{P}_n \rightsquigarrow \mathbb{P}$. In fact, if nothing is known about $\mathbb{P}$, then $\mathbb{P}_n$ is probably the most appropriate estimator to use as it is asymptotically efficient and minimax in the sense of van der Vaart [37], Theorem 25.21, equation (25.22); also see Example 25.24. In addition, for any Donsker class of functions, $\mathcal{F}$, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \mathrm{O}_{\mathbb{P}}(n^{-1/2})$, where

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}\left|\int f\,\mathrm{d}\mathbb{P}_n - \int f\,\mathrm{d}\mathbb{P}\right|,$$

that is, $\mathbb{P}_n - \mathbb{P}$ is asymptotically of the order of $n^{-1/2}$ uniformly in $\mathcal{F}$ and the processes $f \mapsto \sqrt{n}\int f\,\mathrm{d}(\mathbb{P}_n - \mathbb{P})$, $f \in \mathcal{F}$ converge in law to a Gaussian process in $\ell^\infty(\mathcal{F})$, called the $\mathbb{P}$-Brownian bridge indexed by $\mathcal{F}$, where $\ell^\infty(\mathcal{F})$ denotes the Banach space of bounded real-valued

1350-7265

functions on $\mathcal{F}$. On the other hand, if $\mathbb{P}$ has a density p with respect to Lebesgue measure (assuming $\mathcal{X} = \mathbb{R}^d$), then $\mathbb{P}_n$, which is a random atomic measure, is not appropriate to estimate p. However, various estimators, p_n have been proposed in literature to estimate p, the popular ones being the kernel density estimator and wavelet estimator, which under suitable conditions have been shown to be optimal with respect to the L^r loss ($1 \le r \le \infty$) in the sense of achieving the minimax rate over all estimators for densities in certain classes (Devroye and Györfi [8], Hardle *et al.* [22], van der Vaart [37]). Therefore, depending on whether $\mathbb{P}$ has a density or not, there are two different estimators (i.e., $\mathbb{P}_n$ and p_n) that are optimal in two different performance measures, that is, $\|\cdot\|_{\mathcal{F}}$ and L^r. While $\mathbb{P}_n$ is not adequate to estimate p, the question arises as to whether $\mathbb{P}_n^\star$ defined as $\mathbb{P}_n^\star(A) := \int_A p_n(x)\,\mathrm{d}x$ for every Borel set $A \subset \mathbb{R}^d$, estimates $\mathbb{P}$ as good as $\mathbb{P}_n$ in the sense that $\|\mathbb{P}_n^\star - \mathbb{P}\|_{\mathcal{F}} = \mathrm{O}_{\mathbb{P}}(n^{-1/2})$, that is,

$$\sup_{f\in\mathcal{F}}\left|\int f(x)p_n(x)\,\mathrm{d}x - \int f(x)p(x)\,\mathrm{d}x\right| = \mathrm{O}_{\mathbb{P}}\bigl(n^{-1/2}\bigr), \tag{1.1}$$

and whether the processes $f \mapsto \sqrt{n}\int f(x)(p_n - p)(x)\,\mathrm{d}x$, $f \in \mathcal{F}$ converge in law to a Gaussian process in $\ell^\infty(\mathcal{F})$ for $\mathbb{P}$-Donsker class, $\mathcal{F}$. If p_n satisfies these properties, then it is a *plug-in* estimator in the sense of Bickel and Ritov [6], Definition 4.1, as it is simultaneously optimal in two different performance measures. The question of whether (1.1) holds has been addressed for the kernel density estimator (Yukich [41], van der Vaart [36], Giné and Nickl [15]) and wavelet density estimator (Giné and Nickl [17]) where a uniform central limit theorem as stated above has been proved for various $\mathbb{P}$-Donsker classes, $\mathcal{F}$ (and also for non-Donsker but pre-Gaussian classes in Radulović and Wegkamp [27] and Giné and Nickl [15], Section 4.2). For a $\mathbb{P}$-Donsker class $\mathcal{F}$, it easy to show that (1.1) and the corresponding uniform central limit theorem (UCLT) hold if $\|\mathbb{P}_n^\star - \mathbb{P}_n\|_{\mathcal{F}} = \mathrm{o}_{\mathbb{P}}(n^{-1/2})$, that is,

$$\sup_{f\in\mathcal{F}}\left|\int f(x)p_n(x)\,\mathrm{d}x - \int f(x)\,\mathrm{d}\mathbb{P}_n(x)\right| = \mathrm{o}_{\mathbb{P}}\bigl(n^{-1/2}\bigr). \tag{1.2}$$

Several recent works (Bickel and Ritov [6], Nickl [26], Giné and Nickl [15,17–19]) have shown that many popular density estimators on $\mathcal{X} = \mathbb{R}$, such as maximum likelihood estimator, kernel density estimator and wavelet estimator satisfy (1.2) if $\mathcal{F}$ is $\mathbb{P}$-Donsker – the Donsker classes that were considered in these works are: functions of bounded variation, $\{\mathbb{1}_{(-\infty,t]}: t \in \mathbb{R}\}$, Hölder, Lipschitz and Sobolev classes on $\mathbb{R}$. In other words, these works show that there exists estimators that are within a $\|\cdot\|_{\mathcal{F}}$-ball of size $\mathrm{o}_{\mathbb{P}}(n^{-1/2})$ around $\mathbb{P}_n$ such that they estimate $\mathbb{P}$ consistently in $\|\cdot\|_{\mathcal{F}}$ at the rate of $n^{-1/2}$, that is, they have a statistical behavior similar to that of $\mathbb{P}_n$.

The main contribution of this paper is to generalize the above behavior of kernel density estimators to any d by showing that $\mathbb{P}$ can be estimated optimally in $\|\cdot\|_{\mathcal{F}_H}$ using a kernel density estimator, p_n (with L^1 optimal bandwidth) on $\mathbb{R}^d$ where under certain conditions on $\mathcal{K}$, $\|\cdot\|_{\mathcal{F}_H}$ with

$$\mathcal{F}_H := \bigl\{f : \mathbb{R}^d \to \mathbb{R} \,|\, \|f\|_{\mathcal{H}_k} \le 1:\ f \in \mathcal{H}_k, k \in \mathcal{K}\bigr\} \tag{1.3}$$

metrizes the weak topology on the space of Borel probability measure on $\mathbb{R}^d$. Here, $\mathcal{H}_k$ denotes a reproducing kernel Hilbert space (RKHS) (Aronszajn [2]); also see Berlinet and Thomas-Agnan [5] and Steinwart and Christmann [35], Chapter 4, for a nice introduction to RKHS and its applications in probability, statistics and learning theory – with $k:\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as the reproducing kernel (and, therefore, positive definite) and $\mathcal{K}$ is a cone of positive definite kernels. To elaborate, the paper shows that the kernel density estimator on $\mathbb{R}^d$ with an appropriate choice of bandwidth is not only optimal in the strong topology (i.e., in total variation distance or L^1) but also optimal in the weak topology induced by $\|\cdot\|_{\mathcal{F}_H}$ (i.e., has a similar statistical behavior to that of $\mathbb{P}_n$). On the other hand, note that $\mathbb{P}_n$ is an optimal estimator of $\mathbb{P}$ only in the weak topology and is far from optimal in the strong topology as it is not even a consistent estimator of $\mathbb{P}$. A similar result – optimality of kernel density estimator in both weak and strong topologies – was shown by Giné and Nickl [16] for only $d = 1$ where $\mathcal{F}$ is chosen to be a unit ball of bounded Lipschitz functions, $\mathcal{F}_{\text{BL}}$ on $\mathbb{R}^d$, defined as

$$\mathcal{F}_{\text{BL}} := \left\{ f:\mathbb{R}^d \to \mathbb{R} \,\middle|\, \|f\|_{\text{BL}} := \sup_{x\in\mathbb{R}^d} \left|f(x)\right| + \sup_{x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2} \le 1 \right\}, \tag{1.4}$$

with $\|\cdot\|_2$ being the Euclidean norm. In comparison, our work generalizes the result of Giné and Nickl [16] to any d by working with $\mathcal{F}_H$.

Before presenting our results, in Section 3, we provide a brief introduction to reproducing kernel Hilbert spaces, discuss some relevant properties of $\|\cdot\|_{\mathcal{F}_H}$ and provide concrete examples for $\mathcal{F}_H$ through some concrete choices of $\mathcal{K}$. We then present our first main result in Theorem 3.2 which shows that under certain conditions on $\mathcal{K}$, $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology on the space of probability measures. Since $\mathbb{P}_n$ is a consistent estimator of $\mathbb{P}$ in weak sense, we then obtain a rate for this convergence by showing in Theorem 3.3 that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}(n^{-1/2})$ by bounding the expected suprema of U-processes – specifically, the homogeneous Rademacher chaos process of degree 2 – indexed by a uniformly bounded Vapnik-Červonenkis (VC)-subgraph class $\mathcal{K}$ (see de la Peña and Giné [7], Chapter 5 for details on U-processes). Since Theorems 3.2 and 3.3 are very general, we provide examples (see Example 2) to show that a large family of $\mathcal{K}$ satisfy the assumptions in these results and, therefore, yield a variety of probability metrics that metrize the weak convergence while ensuring a dimension independent rate of $n^{-1/2}$ for $\mathbb{P}_n$ converging to $\mathbb{P}$.

In Theorem 4.1, we present our second main result which provides an exponential inequality for the tail probabilities of $\|\mathbb{P}_n^{\star} - \mathbb{P}_n\|_{\mathcal{F}_H} = \|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H}$, where $\mathbb{P}_n * K_h$ is the kernel density estimator with bandwidth h, $*$ represents the convolution and $K_h = h^{-d}K(\cdot/h)$ with $K:\mathbb{R}^d \to \mathbb{R}$. The proof is based on an application of McDiarmid's inequality, together with expectation bounds on the suprema of homogeneous Rademacher chaos process of degree 2, indexed over VC-subgraph classes. For sufficiently smooth reproducing kernels (see Theorem 4.1 for details), this result shows that the kernel density estimator on $\mathbb{R}^d$ is within a $\|\cdot\|_{\mathcal{F}_H}$-ball of size $\mathrm{o}_{\mathbb{P}}(n^{-1/2})$ around $\mathbb{P}_n$ (which means $\mathcal{F}_H$ ensures (1.2)) and, therefore, combining Theorems 3.2 and 4.1 yields that the kernel density estimator with L^1 optimal bandwidth is a consistent estimator of $\mathbb{P}$ in weak sense with a convergence rate of $n^{-1/2}$ (and hence is optimal in both strong and weak topologies). We then provide concrete examples of $\mathcal{K}$ in Theorem 4.2 (also see Remark 4.3) that guarantee this behavior for the kernel density estimator. Giné and Nickl [16] proved a similar result for $\mathcal{F}_{\text{BL}}$ with $d = 1$ which can be generalized to any $d \ge 2$ using Corollary 3.5 in Sriperumbudur *et al.* [31]. However, for $d > 2$, it can only be shown that the kernel density estimator with L^1 optimal bandwidth is within $\|\cdot\|_{\mathcal{F}_{\text{BL}}}$-ball of size $\mathrm{o}_{\mathbb{P}}(n^{-1/d})$ – it is $\mathrm{o}_{\mathbb{P}}(\sqrt{\log n}/\sqrt{n})$ for $d = 2$ – around $\mathbb{P}_n$ instead of $\mathrm{o}_{\mathbb{P}}(n^{-1/2})$ as with $\|\cdot\|_{\mathcal{F}_H}$.

Now given that (1.1) holds for $\mathcal{F} = \mathcal{F}_H$ (see Theorem 4.1 for detailed conditions and Theorem 4.2 for examples), it is of interest to know whether the processes $f \mapsto \sqrt{n} \int f \, \mathrm{d}(\mathbb{P}_n * K_h - \mathbb{P})$, $f \in \mathcal{F}_H$ converge in law to a Gaussian process in $\ell^\infty(\mathcal{F}_H)$. While it is not easy to verify the $\mathbb{P}$-Donsker property of $\mathcal{F}_H$ or the conditions in Giné and Nickl ([15], Theorem 3) which ensure this UCLT in $\ell^\infty(\mathcal{F}_H)$ for any general $\mathcal{K}$ that induces $\mathcal{F}_H$, in Theorem 4.3, we present concrete examples of $\mathcal{K}$ for which $\mathcal{F}_H$ is $\mathbb{P}$-Donsker so that the following UCLTs are obtained:

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P} \quad \text{and} \quad \sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P},$$

where $\mathbb{G}_\mathbb{P}$ denotes the $\mathbb{P}$-Brownian bridge indexed by $\mathcal{F}_H$ and $\rightsquigarrow_{\ell^\infty(\mathcal{F}_H)}$ denotes the convergence in law of random elements in $\ell^\infty(\mathcal{F}_H)$. A similar result was presented in Giné and Nickl ([16], Theorem 1) for $\mathcal{F}_{\mathrm{BL}}$ with $d = 1$ under the condition that $\mathbb{P}$ satisfies $\int_\mathbb{R} |x|^{2\gamma} \, \mathrm{d}\mathbb{P}(x) < \infty$ for some $\gamma > 1/2$, which shows that additional conditions are required on $\mathbb{P}$ to obtain a UCLT while working with $\mathcal{F}_{\mathrm{BL}}$ in contrast to $\mathcal{F}_H$ where no such conditions are needed.

While the choice of $\mathcal{F}_H$ is abstract, there are significant computational advantages associated with this choice (over say $\mathcal{F}_{\mathrm{BL}}$), which we discuss in Section 5, where we show that for certain $\mathcal{K}$, it is very easy to compute $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H}$ compared to $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_{\mathrm{BL}}}$ as in the former case, the problem reduces to a maximization problem over $\mathbb{R}$ in contrast to an infinite dimensional optimization problem in $\mathcal{F}_{\mathrm{BL}}$. The need to compute $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_\mathcal{F}$ occurs while constructing adaptive estimators that estimate $\mathbb{P}$ efficiently in $\mathcal{F}$ and at the same time estimates the density of $\mathbb{P}$ (if it exists, but without a priori assuming its existence) at the best possible convergence rate in some relevant loss over prescribed class of densities, for example, sup-norm loss over the Hölder balls and L^1-loss over Sobolev balls. The construction of these adaptive estimators involves applying Lepski's method (Lepski, Mammen and Spokoiny [23]) to kernel density estimators that are within a $\|\cdot\|_\mathcal{F}$-ball of size smaller than $n^{-1/2}$ around $\mathbb{P}_n$, which in turn involves computing $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_\mathcal{F}$ (see Giné and Nickl [16], Theorem 1, [18], Theorem 2 and [19], Theorem 3). Along the lines of Giné and Nickl [16], Theorem 1, in Section 5, we also discuss the optimal adaptive estimation of $\mathbb{P}$ in weak and strong topologies.

Various notation and definitions that are used throughout the paper are collected in Section 2. The missing proofs of the results are provided in Section 6 and supplementary results are collected in the Appendix.

2. Definitions and notation

Let $\mathcal{X}$ be a topological space. $\ell^\infty(\mathcal{X})$ denotes the Banach space of bounded real-valued functions F on $\mathcal{X}$ normed by $\|F\|_\mathcal{X} := \sup_{x \in \mathcal{X}} |F(x)|$. $C(\mathcal{X})$ denotes the space of all continuous real-valued functions on $\mathcal{X}$. $C_b(\mathcal{X})$ is the space of all bounded, continuous real-valued functions on $\mathcal{X}$. For a locally compact Hausdorff space, $\mathcal{X}$, $f \in C(\mathcal{X})$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x \in \mathcal{X}: |f(x)| \geq \epsilon\}$ is compact. The class of all continuous f on $\mathcal{X}$ which vanish at infinity is denoted as $C_0(\mathcal{X})$. The spaces $C_b(\mathcal{X})$ and $C_0(\mathcal{X})$ are endowed with the uniform norm, $\|\cdot\|_\mathcal{X}$, which we alternately denote as $\|\cdot\|_\infty$. $M_+^1(\mathcal{X})$ denotes the space of all Borel probability measures defined on $\mathcal{X}$ while $M_b(\mathcal{X})$ denotes the space of all finite signed Borel measures on $\mathcal{X}$. $L^r(\mathcal{X}, \mu)$ denotes the Banach space of r-power ($r \geq 1$) μ-integrable functions where μ is a Borel measure defined on $\mathcal{X}$. We will write $L^r(\mathcal{X})$ for $L^r(\mathcal{X},\mu)$ if μ is a Lebesgue measure on $\mathcal{X}\subset\mathbb{R}^d$. $W_1^s(\mathbb{R}^d)$ denotes the space of functions $f\in L^1(\mathbb{R}^d)$ whose partial derivatives up to order $s\in\mathbb{N}$ exist and are in $L^1(\mathbb{R}^d)$. $\mathcal{F}_{\mathrm{BL}}$ denotes the unit ball of bounded Lipschitz functions on $\mathcal{X}=\mathbb{R}^d$ as shown in (1.4). A function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{C}$ is called a positive definite (p.d.) kernel if, for all $n\in\mathbb{N}$, $(\alpha_1,\dots,\alpha_n)\in\mathbb{C}^n$ and all $(x_1,\dots,x_n)\in\mathcal{X}^n$, we have

$$\sum_{i,j=1}^{n}\alpha_i\overline{\alpha}_j k(x_i,x_j)\geq 0,$$

where $\overline{\alpha}$ is the complex conjugate of $\alpha\in\mathbb{C}$. $\mathcal{H}_k$ denotes a reproducing kernel Hilbert space (RKHS) (see Definition 1) of functions with a positive definite k as the reproducing kernel and $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ denotes the inner product on $\mathcal{H}_k$. $\mathcal{F}_H$ denotes the unit ball of RKHS functions indexed by a cone of positive definite kernels, $\mathcal{K}$ as shown in (1.3). The convolution $f*g$ of two measurable functions f and g on $\mathbb{R}^d$ is defined as $(f*g)(x):=\int_{\mathbb{R}^d}f(y)g(x-y)\,\mathrm{d}y$, provided the integral exists for all $x\in\mathbb{R}^d$. Similarly, the convolution of $\mu\in M_b(\mathbb{R}^d)$ and measurable f is defined as

$$(f*\mu)(x):=\int_{\mathbb{R}^d}f(x-y)\,\mathrm{d}\mu(y)$$

if the integral exists for all $x\in\mathbb{R}^d$. For $f\in L^1(\mathbb{R}^d)$, its Fourier transform is defined as

$$\widehat{f}(y)=(2\pi)^{-d/2}\int_{\mathbb{R}^d}f(x)\mathrm{e}^{-\sqrt{-1}\langle y,x\rangle}\,\mathrm{d}x.$$

A sequence of probability measures, $(\mathbb{P}_{(n)})_{n\in\mathbb{N}}$ is said to converge weakly to $\mathbb{P}$ (denoted as $\mathbb{P}_{(n)}\rightsquigarrow\mathbb{P}$) if and only if $\int f\,\mathrm{d}\mathbb{P}_{(n)}\to\int f\,\mathrm{d}\mathbb{P}$ for all $f\in C_b(\mathcal{X})$ as $n\to\infty$. For a Borel-measurable real-valued function f on $\mathcal{X}$ and $\mu\in M_b(\mathcal{X})$, we define $\mu f:=\int_{\mathcal{X}}f\,\mathrm{d}\mu$. The empirical process indexed by $\mathcal{F}\subset L^2(\mathcal{X},\mathbb{P})$ is given by $f\mapsto\sqrt{n}(\mathbb{P}_n-\mathbb{P})f=n^{-1/2}\sum_{i=1}^n(f(X_i)-\mathbb{P}f)$, where $\mathbb{P}_n:=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}$ with $(X_i)_{i=1}^n$ being random samples drawn i.i.d. from $\mathbb{P}$ and δ_x represents the Dirac measure at x. $\mathcal{F}$ is said to be $\mathbb{P}$-*Donsker* if $\sqrt{n}(\mathbb{P}_n-\mathbb{P})\rightsquigarrow_{\ell^\infty(\mathcal{F})}\mathbb{G}_{\mathbb{P}}$, where $\mathbb{G}_{\mathbb{P}}$ is the Brownian bridge indexed by $\mathcal{F}$, that is, a centered Gaussian process with covariance $\mathbb{E}\mathbb{G}_{\mathbb{P}}(f)\mathbb{G}_{\mathbb{P}}(g)=\mathbb{P}((f-\mathbb{P}f)(g-\mathbb{P}g))$ and if $\mathbb{G}_{\mathbb{P}}$ is sample-bounded and sample-continuous w.r.t. the covariance metric. $\rightsquigarrow_{\ell^\infty(\mathcal{F})}$ denotes the convergence in law (or weak convergence) of random elements in $\ell^\infty(\mathcal{F})$. $\mathcal{F}$ is said to be *universal Donsker* if it is $\mathbb{P}$-Donsker for all $\mathbb{P}\in M_+^1(\mathcal{X})$.

Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. The collection $\mathcal{C}$ is said to *shatter* an arbitrary set of n points, $\{x_1,\dots,x_n\}$, if for each of its 2^n subsets, there exists $C\in\mathcal{C}$ such that $C\cap\{x_1,\dots,x_n\}$ yields the subset. The *Vapnik-Červonenkis (VC)-index*, $VC(\mathcal{C})$ of the class $\mathcal{C}$ is the maximal n for which an n-point set is shattered by $\mathcal{C}$. If $VC(\mathcal{C})$ is finite, then $\mathcal{C}$ is said to be a *VC-class*. A collection $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ is called a *VC-subgraph class* if the collection of all subgraphs of the functions in $\mathcal{F}$, that is, $\{\{(x,t)\colon t<f(x)\}\colon f\in\mathcal{F}\}$ forms a VC-class of sets in $\mathcal{X}\times\mathbb{R}$. The *covering number* $\mathcal{N}(\mathcal{F},\rho,\epsilon)$ is the minimal number of balls $\{g\colon\rho(f,g)<\epsilon\}$ of radius ϵ needed to cover $\mathcal{F}$, where ρ is a metric on $\mathcal{F}$.

Given random samples $(X_i)_{i=1}^n \subset \mathbb{R}^d$ drawn i.i.d. from $\mathbb{P}$, the kernel density estimator is defined as

$$(\mathbb{P}_n * K_h)(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right), \qquad x \in \mathbb{R}^d,$$

where $K:\mathbb{R}^d \to \mathbb{R}$ is the smoothing kernel that satisfies $K(x) = K(-x), x \in \mathbb{R}^d$, $K \in L^1(\mathbb{R}^d)$ and $\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = 1$ with $K_h(x) := h^{-d}K(x/h)$ and $0 < h := h_n \to 0$ as $n \to \infty$. K is said to be of order $r > 0$ if

$$\int_{\mathbb{R}^d} \prod_{i=1}^d y_i^{\alpha_i} K(y)\,\mathrm{d}y = 0 \qquad \text{for } 0 < |\alpha| \le r-1 \quad \text{and}$$

$$\int_{\mathbb{R}^d} \prod_{i=1}^d |y_i|^{\alpha_i} \big|K(y)\big|\,\mathrm{d}y < \infty \qquad \text{for } |\alpha| = r,$$

where $y = (y_1,\dots,y_d)$, $\alpha = (\alpha_1,\dots,\alpha_d)$, $\alpha_i \ge 0, \forall i = 1,\dots,d$ and $|\alpha| := \sum_{i=1}^d \alpha_i$. We refer the reader to Berlinet and Thomas-Agnan ([5], Chapter 3, Section 8) for details about the construction of kernels of arbitrary order, r.

We would like to mention that throughout the paper, we ignore the measurability issues that are associated with the suprema of an empirical process (or a U-process) and therefore the probabilistic statements about these objects should be considered in the outer measure.

3. Reproducing kernel Hilbert spaces and $\|\cdot\|_{\mathcal{F}_H}$

In this section, we present a brief overview of RKHS along with some properties of $\|\cdot\|_{\mathcal{F}_H}$ with a goal to provide an intuitive understanding of, otherwise an abstract class $\mathcal{F}_H$ and its associated distance, $\|\cdot\|_{\mathcal{F}_H}$. Throughout this section, we assume that $\mathcal{X}$ is a topological space.

3.1. Preliminaries

We start with the definition of an RKHS, which we quote from Berlinet and Thomas-Agnan [5]. For the purposes of this paper, we deal with real-valued RKHS though the following definition can be extended to the complex-valued case (see Berlinet and Thomas-Agnan [5], Chapter 1, Definition 1).

Definition 1 (Reproducing kernel Hilbert space). *Let $(\mathcal{H}_k, \langle\cdot,\cdot\rangle_{\mathcal{H}_k})$ be a Hilbert space of real-valued functions on $\mathcal{X}$. A function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}, (x,y)\mapsto k(x,y)$ is called a* reproducing kernel *of the Hilbert space $\mathcal{H}_k$ if and only if the following hold*:

(i) $\forall y \in \mathcal{X}, k(\cdot,y) \in \mathcal{H}_k$;
(ii) $\forall y \in \mathcal{X}, \forall f \in \mathcal{H}_k, \langle f, k(\cdot,y)\rangle_{\mathcal{H}_k} = f(y)$.

If such a k exists, then $\mathcal{H}_k$ is called a reproducing kernel Hilbert space.

Using the Riesz representation theorem, the above definition can be shown to be equivalent to defining $\mathcal{H}_k$ as an RKHS if for all $x \in \mathcal{X}$, the evaluation functional, $\delta_x : \mathcal{H}_k \to \mathbb{R}$, $\delta_x(f) := f(x)$, $f \in \mathcal{H}_k$ is continuous (Berlinet and Thomas-Agnan [5], Chapter 1, Theorem 1). Starting from Definition 1, it can be shown that $\mathcal{H}_k = \overline{\text{span}}\{k(\cdot, x)\colon\ x \in \mathcal{X}\}$ where the closure is taken w.r.t. the RKHS norm (see Berlinet and Thomas-Agnan [5], Chapter 1, Theorem 3), which means the kernel function, k generates the RKHS. Since $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y), \forall x, y \in \mathcal{X}$, it is easy to show that every reproducing kernel (r.k.), k is symmetric and positive definite. More interestingly, the converse is also true, that is, the Moore–Aronszajn theorem (Aronszajn [2]) states that for every positive definite kernel, k, there exists a unique RKHS, $\mathcal{H}_k$ with k as the r.k. Since k is a reproducing kernel if and only if it is positive definite, usually it might be simpler to verify for the positive definiteness of k rather than explicitly constructing $\mathcal{H}_k$ and verifying whether k satisfies the properties in Definition 1. An important characterization for positive definiteness on $\mathbb{R}^d$ (more generally on locally compact Abelian groups) is given by Bochner's theorem (Wendland [39], Theorem 6.6): a bounded continuous translation invariant kernel $k(x, y) = \psi(x - y)$ on $\mathbb{R}^d$ is positive definite if and only if ψ is the Fourier transform of a nonnegative finite Borel measure, Υ, that is,

$$k(x, y) = \psi(x - y) = \int \mathrm{e}^{-\sqrt{-1}(x-y)^T\omega}\, \mathrm{d}\Upsilon(\omega), \qquad x, y \in \mathbb{R}^d. \tag{3.1}$$

In addition, if $\psi \in L^1(\mathbb{R}^d)$, then the corresponding RKHS is given by Wendland (see [39], Theorem 10.12),

$$\mathcal{H}_k = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)\colon \int \frac{|\widehat{f}(\omega)|^2}{\widehat{\psi}(\omega)}\, \mathrm{d}\omega < \infty \right\},$$

where $\widehat{f}$ and $\widehat{\psi}$ denote the Fourier transforms of f and ψ, respectively. Note that since $\psi \in L^1(\mathbb{R}^d)$, we have $\mathrm{d}\Upsilon(\omega) = (2\pi)^{-d/2}\widehat{\psi}(\omega)\, \mathrm{d}\omega$. Another characterization of positive definiteness (which we will use later in our results) is due to Schönberg for radially symmetric positive definite kernels (Wendland [39], Theorems 7.13 and 7.14): A function $k(x, y) = \phi(\|x - y\|_2^2)$, $x, y \in \mathbb{R}^d$ is positive definite if and only if ϕ is the Laplace transform of a nonnegative finite Borel measure, ν on $[0, \infty)$, that is,

$$k(x, y) = \phi(\|x - y\|_2^2) = \int_0^\infty \mathrm{e}^{-t\|x-y\|_2^2}\, \mathrm{d}\nu(t), \qquad x, y \in \mathbb{R}^d. \tag{3.2}$$

Note that the Bochner's characterization in (3.1) is also valid for k in (3.2) as it is also translation invariant. Two important examples of positive definite kernels and their corresponding RKHSs that appear throughout the paper are: *Gaussian kernel*, $k(x, y) = \exp(-\sigma\|x - y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$, which induces the *Gaussian RKHS*,

$$\mathcal{H}_k = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)\colon \int |\widehat{f}(\omega)|^2 \mathrm{e}^{\|\omega\|_2^2/4\sigma}\, \mathrm{d}\omega < \infty \right\} \tag{3.3}$$

and the *Matérn kernel*, $k(x,y)=\frac{2^{1-\beta}}{\Gamma(\beta)}\|x-y\|_2^{\beta-d/2}\mathfrak{K}_{d/2-\beta}(\|x-y\|_2), x,y\in\mathbb{R}^d, \beta>d/2$, which induces the Sobolev space, H_2^β,

$$\mathcal{H}_k=H_2^\beta=\left\{f\in L^2(\mathbb{R}^d)\cap C(\mathbb{R}^d)\colon \int\left(1+\|\omega\|_2^2\right)^\beta\left|\widehat{f}(\omega)\right|^2\,\mathrm{d}\omega<\infty\right\}. \tag{3.4}$$

Here, Γ is the Gamma function and $\mathfrak{K}_v$ is the modified Bessel function of the third kind of order v, where v controls the smoothness of k. Note that $L^2(\mathbb{R}^d)$ is not an RKHS as it does not consist of functions.

3.2. Properties of $\|\cdot\|_{\mathcal{F}_H}$

In the following, we present various properties which are not only helpful to intuitively understand $\|\cdot\|_{\mathcal{F}_H}$ but are also useful to derive our main results in Section 4. First, in Proposition 3.1, we provide an alternate expression for $\|\cdot\|_{\mathcal{F}_H}$ to obtain a better interpretation, using which we discuss the relation of $\|\cdot\|_{\mathcal{F}_H}$ to other classical distances on probabilities. This alternate representation will be particularly helpful in studying the convergence of $\mathbb{P}_n$ to $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$ and also in deriving our main result (Theorem 4.1) in Section 4. Second, we discuss the question of the metric property of $\|\cdot\|_{\mathcal{F}_H}$ – it is easy to verify that $\|\cdot\|_{\mathcal{F}_H}$ is a pseudometric – and highlight some of the results that we obtained in our earlier works along with some examples in Example 1. Third, we present a new result in Theorem 3.2 about the topology induced by $\|\cdot\|_{\mathcal{F}_H}$ wherein we show that under certain mild conditions on $\mathcal{K}$, $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology on probability measures. We also present some examples of $\mathcal{K}$ in Example 2 that satisfy these conditions thereby ensuring the metrization of weak topology by the corresponding metric, $\|\cdot\|_{\mathcal{F}_H}$. Finally, in Theorem 3.3, we present an exponential concentration inequality for the tail probabilities of $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}_H}$ and show that for various families of $\mathcal{K}$ (see Remark 3.1(i) and Theorem 4.2), $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}_H}=\mathrm{O}_{\mathrm{a.s.}}(n^{-1/2})$, which when combined with Theorem 3.2 provides a rate of convergence of $n^{-1/2}$ for $\mathbb{P}_n$ converging to $\mathbb{P}$ in weak sense.

Alternate representation for $\|\cdot\|_{\mathcal{F}_H}$: The following result (a similar result is proved in Sriperumbudur *et al.* [34], Theorem 1, where $\mathcal{K}$ is chosen to be a singleton set but we provide a proof in Section 6.1 for completeness) presents an alternate representation to $\|\cdot\|_{\mathcal{F}_H}$. This representation is particularly useful as it shows that $\|\cdot\|_{\mathcal{F}_H}$ is completely determined by the kernels, $k\in\mathcal{K}$ and does not depend on the individual functions in the corresponding RKHSs.

Proposition 3.1. *Define* $\mathscr{P}_{\mathcal{K}}:=\{\mathbb{P}\in M_+^1(\mathcal{X})\colon \sup_{k\in\mathcal{K}}\int\sqrt{k(x,x)}\,\mathrm{d}\mathbb{P}(x)<\infty\}$ *where every* $k\in\mathcal{K}$, $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ *is measurable. Then for any* $\mathbb{P},\mathbb{Q}\in\mathscr{P}_{\mathcal{K}}$,

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H}=\sup_{k\in\mathcal{K}}\mathfrak{D}_k(\mathbb{P},\mathbb{Q}), \tag{3.5}$$

where

$$\mathfrak{D}_k(\mathbb{P},\mathbb{Q}) := \left\| \int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x) - \int k(\cdot,x)\,\mathrm{d}\mathbb{Q}(x) \right\|_{\mathcal{H}_k} \tag{3.6}$$

$$= \sqrt{\int\int k(x,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)}, \tag{3.7}$$

with $\int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x)$ *and* $\int k(\cdot,x)\,\mathrm{d}\mathbb{Q}(x)$ *being defined in Bochner sense* (*Diestel and Uhl* [9], *Definition* 1).

Since $\sqrt{k(x,x)} = \|k(\cdot,x)\|_{\mathcal{H}_k}$, it is easy to verify from Proposition 3.1 that for any $\mathbb{P},\mathbb{Q} \in \mathscr{P}_{\mathcal{K}}$, $\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H} < \infty$. It also follows from (3.5) and (3.6) that $\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H}$ can be interpreted as the supremum distance between the embeddings $\mathbb{P} \mapsto \int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x)$ and $\mathbb{Q} \mapsto \int k(\cdot,x)\,\mathrm{d}\mathbb{Q}(x)$, indexed by $k \in \mathcal{K}$. Choosing $k(\cdot,x)$ as

$$\frac{1}{(2\pi)^{d/2}}\mathrm{e}^{-\sqrt{-1}\langle\cdot,x\rangle_2}, \qquad \mathrm{e}^{\langle\cdot,x\rangle_2} \quad \text{and} \quad \frac{1}{(4\pi)^{d/2}}\mathrm{e}^{-\|\cdot-x\|_2^2/4}, \qquad x \in \mathbb{R}^d, \tag{3.8}$$

the embedding $\Phi : M_+^1(\mathcal{X}) \to \mathcal{H}_k$, $\mathbb{P} \mapsto \int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x)$ reduces to the characteristic function, moment generating function (if it exists) and Weierstrass transform of $\mathbb{P}$, respectively. In this sense, Φ can be seen as a generalization of these notions (which are all defined on $\mathbb{R}^d$) to an arbitrary topological space $\mathcal{X}$ (in fact, it holds for any arbitrary measurable space).

When is $\|\cdot\|_{\mathcal{F}_H}$ *a metric on* $\mathscr{P}_{\mathcal{K}}$? While $\|\cdot\|_{\mathcal{F}_H}$ is a pseudo-metric on $\mathscr{P}_{\mathcal{K}}$, it is in general not a metric as $\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H} = 0 \not\Longrightarrow \mathbb{P} = \mathbb{Q}$ as shown by the choice $\mathcal{K} = \{k\}$ where $k(x,y) = \langle x,y\rangle_2$, $x,y \in \mathbb{R}^d$. For this choice, it is easy to check that $\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H}$ is the Euclidean distance between the means of $\mathbb{P}$ and $\mathbb{Q}$ and, therefore, is not a metric on $\{\mathbb{P} \in M_+^1(\mathbb{R}^d): \int \|x\|\,\mathrm{d}\mathbb{P}(x) < \infty\}$ (and hence on $M_+^1(\mathbb{R}^d)$). The question of when is $\mathfrak{D}_k$ a metric on $M_+^1(\mathcal{X})$ is addressed in Fukumizu *et al.* [13,14], Gretton *et al.* [21] and Sriperumbudur *et al.* [34]. By defining any kernel for which $\mathfrak{D}_k$ is a metric as the *characteristic kernel*, it is easy to see that if any $k \in \mathcal{K}$ is characteristic, then $\|\cdot\|_{\mathcal{F}_H}$ is a metric on $\mathscr{P}_{\mathcal{K}}$. Sriperumbudur *et al.* ([34], Theorem 7) showed that k is characteristic if and only if

$$\int\int k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) > 0 \qquad \forall \mu \in M_b(\mathcal{X}) \setminus \{0\} \text{ with } \mu(\mathcal{X}) = 0. \tag{3.9}$$

Combining this with the Bochner characterization for positive definiteness (see (3.1)), Sriperumbudur *et al.* ([34], Corollary 4) showed that

$$\mathfrak{D}_k(\mathbb{P},\mathbb{Q}) = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d,\Upsilon)}, \qquad \mathbb{P},\mathbb{Q} \in M_+^1(\mathbb{R}^d),$$

using which k is shown to be characteristic if and only if $\operatorname{supp}(\Upsilon) = \mathbb{R}^d$ (Sriperumbudur *et al.* [34], Theorem 9) – Fukumizu *et al.* [14] generalized this result to locally compact Abelian groups, compact non-Abelian groups and the semigroup $\mathbb{R}_+^d$. Here, $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ represent the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Another interesting characterization for the characteristic property of k is obtained by Fukumizu *et al.* [13,14] and Gretton *et al.* [21], which relates it to the richness of $\mathcal{H}_k$ in the sense of approximating certain classes of functions by functions in $\mathcal{H}_k$. We refer the reader to Sriperumbudur, Fukumizu and Lanckriet [33] for more details on the relation between the characteristic property of k and the richness of $\mathcal{H}_k$.

Example 1. The following are some examples of $\mathcal{K}$ for which $\|\cdot\|_{\mathcal{F}_H}$ is a metric on $M_+^1(\mathbb{R}^d)$:

1. Gaussian: $\mathcal{K} = \{\mathrm{e}^{-\sigma\|x-y\|_2^2}, x, y \in \mathbb{R}^d \colon \sigma \in (0,\infty)\}$;
2. Laplacian: $\mathcal{K} = \{\mathrm{e}^{-\sigma\|x-y\|_1}, x, y \in \mathbb{R}^d \colon \sigma \in (0,\infty)\}$;
3. Matérn:
$$\mathcal{K} = \left\{\frac{2(c/2)^{\beta-d/2}}{\Gamma(\beta-d/2)}\|x-y\|_2^{\beta-d/2}\mathfrak{K}_{d/2-\beta}\bigl(c\|x-y\|_2\bigr), x, y \in \mathbb{R}^d, \beta > \frac{d}{2} \colon c \in (0,\infty)\right\},$$
where $\mathfrak{K}_v$ is the modified Bessel function of the third kind of order v;
4. Inverse multiquadrics: $\mathcal{K} = \{(1+\|\frac{x-y}{c}\|_2^2)^{-\beta}, x, y \in \mathbb{R}^d, \beta > 0 \colon c \in (0,\infty)\}$;
5. Splines:
$$\mathcal{K} = \left\{\prod_{j=1}^{d}\left(1 - \frac{|x_j - y_j|}{c_j}\right)\mathbb{1}_{\{|x_j-y_j|\le c_j\}}, x, y \in \mathbb{R}^d \colon c_j \in (0,\infty), \forall j = 1,\dots,d\right\};$$
6. Radial basis functions:
$$\mathcal{K} = \left\{\int_{(0,\infty)} \mathrm{e}^{-\sigma\|x-y\|_2^2}\,\mathrm{d}\Lambda(\sigma), x, y \in \mathbb{R}^d \colon \Lambda \in M_+^1\bigl((0,\infty)\bigr)\right\}.$$

In all these examples, it is easy to check that every $k \in \mathcal{K}$ is bounded and characteristic (as $\operatorname{supp}(\Upsilon) = \mathbb{R}^d$ or in turn satisfies (3.9)) and, therefore, $\|\cdot\|_{\mathcal{F}_H}$ is a metric on $M_+^1(\mathbb{R}^d)$.

Topology induced by $\|\cdot\|_{\mathcal{F}_H}$: Sriperumbudur *et al.* [34] showed that for any bounded kernel k,

$$\mathfrak{D}_k(\mathbb{P},\mathbb{Q}) \le \sup_{x\in\mathcal{X}}\sqrt{k(x,x)}\,\mathrm{TV}(\mathbb{P},\mathbb{Q}),$$

where TV is the total variation distance. This means there can be two distinct $\mathbb{P}$ and $\mathbb{Q}$ which need not be distinguished by $\mathfrak{D}_k$ but are distinguished in total variation, that is, $\mathfrak{D}_k$ induces a topology that is weaker (or coarser) than the strong topology on $M_+^1(\mathcal{X})$. Therefore, it is of interest to understand the topology induced by $\|\cdot\|_{\mathcal{F}_H}$. The following result shows that under additional conditions on $\mathcal{K}$, $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak-topology on $M_+^1(\mathcal{X})$. A special case of this result is already proved in Sriperumbudur *et al.* ([34], Theorem 23), for $\mathfrak{D}_k$ when $\mathcal{X}$ is compact.

Theorem 3.2. *Let $\mathcal{X}$ be a Polish space that is locally compact Hausdorff. Suppose* $(\mathbb{P}_{(n)})_{n\in\mathbb{N}} \subset M_+^1(\mathcal{X})$ *and* $\mathbb{P} \in M_+^1(\mathcal{X})$.

(a) *If there exists a $k \in \mathcal{K}$ such that $k(\cdot,x) \in C_0(\mathcal{X})$ for all $x \in \mathcal{X}$ and*

$$\int\int k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) > 0 \qquad \forall \mu \in M_b(\mathcal{X}) \setminus \{0\}. \tag{3.10}$$

Then

$$\|\mathbb{P}_{(n)} - \mathbb{P}\|_{\mathcal{F}_H} \to 0 \quad \Longrightarrow \quad \mathbb{P}_{(n)} \rightsquigarrow \mathbb{P} \qquad \text{as } n \to \infty.$$

(b) *If* $\mathcal{K}$ *is uniformly bounded, that is,* $\sup_{k\in\mathcal{K},x\in\mathcal{X}} k(x,x) < \infty$ *and satisfies the following property* (P):

$$\forall x \in \mathcal{X}, \forall \epsilon > 0, \exists \text{ open } U_{x,\epsilon} \subset \mathcal{X} \text{ such that } \big\|k(\cdot,x) - k(\cdot,y)\big\|_{\mathcal{H}_k} < \epsilon, \forall k \in \mathcal{K}, \forall y \in U_{x,\epsilon}.$$

Then

$$\mathbb{P}_{(n)} \rightsquigarrow \mathbb{P} \quad \Longrightarrow \quad \|\mathbb{P}_{(n)} - \mathbb{P}\|_{\mathcal{F}_H} \to 0 \qquad \text{as } n \to \infty.$$

Proof. (a) Define $\mathcal{H}_*$ to be the RKHS associated with the reproducing kernel k_*. Suppose $k_* \in \mathcal{K}$ satisfies $k_*(\cdot,x) \in C_0(\mathcal{X}), \forall x \in \mathcal{X}$. By Steinwart and Christmann [35], Lemma 4.28 (see Sriperumbudur, Fukumizu and Lanckriet [32], Theorem 5), it follows that the inclusion $\mathrm{id}: \mathcal{H}_* \to C_0(\mathcal{X})$ is well-defined and continuous. In addition, since $k_* \in \mathcal{K}$ satisfies (3.10), it follows from Sriperumbudur, Fukumizu and Lanckriet ([33], Proposition 4), that $\mathcal{H}_*$ is dense in $C_0(\mathcal{X})$ w.r.t. the uniform norm. We would like to mention that this denseness result simply follows from the Hahn–Banach theorem (Rudin [28], Theorem 3.5 and the remark following Theorem 3.5), which says that $\mathcal{H}_*$ is dense in $C_0(\mathcal{X})$ if and only if

$$\mathcal{H}_*^{\perp} := \left\{\mu \in M_b(\mathcal{X}) \colon \int f \,\mathrm{d}\mu = 0, \forall f \in \mathcal{H}_*\right\} = \{0\}.$$

It is easy to check that $\mathcal{H}_*^{\perp} = \{0\}$ if and only if

$$\mu \mapsto \int k(\cdot,x)\,\mathrm{d}\mu(x), \qquad \mu \in M_b(\mathcal{X})$$

is injective, which is then equivalent to (3.10). Since $\mathcal{H}_*$ is dense in $C_0(\mathcal{X})$ in the uniform norm, for any $f \in C_0(\mathcal{X})$ and every $\epsilon > 0$, there exists a $g \in \mathcal{H}_*$ such that $\|f - g\|_\infty \le \epsilon$. Therefore,

$$\begin{aligned}
|\mathbb{P}_{(n)} f - \mathbb{P} f| &= \big|\mathbb{P}_{(n)}(f-g) + \mathbb{P}(g-f) + (\mathbb{P}_{(n)} g - \mathbb{P} g)\big| \\
&\le \mathbb{P}_{(n)}|f-g| + \mathbb{P}|f-g| + |\mathbb{P}_{(n)} g - \mathbb{P} g| \\
&\le 2\epsilon + |\mathbb{P}_{(n)} g - \mathbb{P} g| \\
&\le 2\epsilon + \|g\|_{\mathcal{H}_*} \mathfrak{D}_{k_*}(\mathbb{P}_{(n)}, \mathbb{P}) \le 2\epsilon + \|g\|_{\mathcal{H}_*} \|\mathbb{P}_{(n)} - \mathbb{P}\|_{\mathcal{F}_H}.
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, $\|g\|_{\mathcal{H}_*} < \infty$ and $\|\mathbb{P}_{(n)} - \mathbb{P}\|_{\mathcal{F}_H} \to 0$ as $n \to \infty$, we have $\mathbb{P}_{(n)} f \to \mathbb{P} f$ for all $f \in C_0(\mathcal{X})$ as $n \to \infty$, which means $\mathbb{P}_{(n)}$ converges to $\mathbb{P}$ vaguely. Since vague convergence and weak convergence are equivalent on the set of Radon probability measures (Berg, Christensen and Ressel [4], page 51), which is same as $M_+^1(\mathcal{X})$ since $\mathcal{X}$ is Polish, the result follows.

(b) Since $\mathcal{K}$ is uniformly bounded, it is easy to see that

$$\sup_{f\in\mathcal{F}_H, x\in\mathcal{X}} \big|f(x)\big| = \sup_{k\in\mathcal{K},x\in\mathcal{X}} \sup_{\|f\|_{\mathcal{H}_k \le 1}} \big|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\big| = \sup_{k\in\mathcal{K},x\in\mathcal{X}} \sqrt{k(x,x)} < \infty,$$

which means $\mathcal{F}_H$ is uniformly bounded. Now, for a given $x \in \mathcal{X}$ and $\epsilon > 0$, pick some $y \in U_{x,\epsilon}$ such that $\|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}_k} < \epsilon$ for all $k \in \mathcal{K}$. This means, for any $f \in \mathcal{F}_H$,

$$\left|f(x) - f(y)\right| \le \|f\|_{\mathcal{H}_k} \left\|k(\cdot, x) - k(\cdot, y)\right\|_{\mathcal{H}_k} < \epsilon,$$

which implies $\mathcal{F}_H$ is equicontinuous on $\mathcal{X}$. The result therefore follows from Dudley [11], Corollary 11.3.4, which shows that if $\mathbb{P}_{(n)} \rightsquigarrow \mathbb{P}$ then $\mathbb{P}_{(n)}$ converges to $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$ as $n \to \infty$. □

Comparing (3.9) and (3.10), it is clear that one requires a stronger condition for weak convergence than for $\|\cdot\|_{\mathcal{F}_H}$ being just a metric. However, these conditions can be shown to be equivalent for bounded continuous translation invariant kernels on $\mathbb{R}^d$, that is, kernels of the type in (3.1). This is because if k satisfies (3.1), then

$$\int\int k(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = \|\widehat{\mu}\|^2_{L^2(\mathbb{R}^d,\Upsilon)}$$

and, therefore, (3.10) holds if and only if $\mathrm{supp}(\Upsilon) = \mathbb{R}^d$, which is indeed the characterization for k being characteristic. Here, $\widehat{\mu}$ represents the Fourier transform of μ defined as $\widehat{\mu}(\omega) = \int \mathrm{e}^{\sqrt{-1}\omega^T x}\,\mathrm{d}\mu(x)$, $\omega \in \mathbb{R}^d$. Therefore, for $\mathcal{K}$ in Example 1, convergence in $\|\cdot\|_{\mathcal{F}_H}$ implies weak convergence on $M^1_+(\mathbb{R}^d)$. However, for the converse to hold, $\mathcal{K}$ has to satisfy (P) in Theorem 3.2, which is captured in the following example.

Example 2. The following families of kernels satisfy the conditions (3.10), (P) and the uniform boundedness condition of Theorem 3.2 so that $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology on $M^1_+(\mathbb{R}^d)$.

1. $\mathcal{K} = \{\mathrm{e}^{-\sigma\|x-y\|_2^2}, x, y \in \mathbb{R}^d\colon \sigma \in (0, a], a < \infty\}$;
2. $\mathcal{K} = \{\mathrm{e}^{-\sigma\|x-y\|_1}, x, y \in \mathbb{R}^d\colon \sigma \in (0, a], a < \infty\}$;
3. $\mathcal{K} = \{\frac{2(c/2)^{\beta-d/2}}{\Gamma(\beta-d/2)}\|x - y\|_2^{\beta-d/2}\mathfrak{K}_{d/2-\beta}(c\|x - y\|_2), x, y \in \mathbb{R}^d, \beta > \frac{d}{2}\colon c \in (0, a], a < \infty\}$;
4. $\mathcal{K} = \{(1 + \|\frac{x-y}{c}\|_2^2)^{-\beta}, x, y \in \mathbb{R}^d, \beta > 0\colon c \in [a, \infty), a > 0\}$;
5. $\mathcal{K} = \{\prod_{i=1}^d(1 + \frac{|x_i-y_i|^2}{c_i^2})^{-1}, x, y \in \mathbb{R}^d\colon c_i \in [a_i, \infty), a_i > 0, \forall i = 1, \dots, d\}$;
6. $\mathcal{K}$ in Theorem 4.2(b) and (c).

Rate of convergence of $\mathbb{P}_n$ *to* $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$: The following result, which is proved in Section 6.2, presents an exponential inequality for the tail probability of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H}$, which in combination with Theorem 3.2 provides a convergence rate for the weak convergence of $\mathbb{P}_n$ to $\mathbb{P}$.

Theorem 3.3. *Let $X_1, \dots, X_n$ be random samples drawn i.i.d. from $\mathbb{P}$ defined on a measurable space $\mathcal{X}$. Assume there exists $\nu > 0$ such that $\sup_{k\in\mathcal{K},x\in\mathcal{X}} k(x, x) \le \nu$. Then for every $\tau > 0$, with probability at least $1 - 2\mathrm{e}^{-\tau}$ over the choice of $(X_i)_{i=1}^n \sim \mathbb{P}^n$,*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \le 4\sqrt{2}\sqrt{\inf_{\alpha>0}\left\{\alpha + \frac{4\mathrm{e}}{n}\int_\alpha^{2\nu} \log 2\mathcal{N}(\mathcal{K}, \rho, \epsilon)\,\mathrm{d}\epsilon\right\}} + \frac{3\sqrt{2\nu}(\sqrt{2} + \sqrt{\tau})}{\sqrt{n}}, \tag{3.11}$$

where for any $k_1, k_2 \in \mathcal{K}$,

$$\rho(k_1, k_2) := \sqrt{\frac{2}{n^2}\sum_{i<j}^{n}\bigl(k_1(X_i, X_j) - k_2(X_i, X_j)\bigr)^2}. \tag{3.12}$$

In particular, if there exists finite positive constants A and β (that are not dependent on n) such that

$$\log \mathcal{N}(\mathcal{K}, \rho, \epsilon) \le A\left(\frac{2\nu}{\epsilon}\right)^{\beta}, \qquad 0 < \epsilon < 2\nu \tag{3.13}$$

then there exists constants $(D_i)_{i=1}^5$ (dependent only on A, β, ν, τ and not on n) such that

$$\mathbb{P}^n\bigl(\bigl\{(X_1, \dots, X_n) \in \mathcal{X}^n \colon \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} > \lambda(A, \beta, \nu, \tau)\bigr\}\bigr) \le 2\mathrm{e}^{-\tau}, \tag{3.14}$$

where

$$\lambda(A, \beta, \nu, \tau) \le \begin{cases} \dfrac{D_1}{\sqrt{n}}, & 0 < \beta < 1, \\ D_2\sqrt{\dfrac{\log n}{n}} + \dfrac{D_3}{\sqrt{n}}, & \beta = 1, \\ \dfrac{D_4}{n^{1/2\beta}} + \dfrac{D_5}{\sqrt{n}}, & \beta > 1. \end{cases} \tag{3.15}$$

Remark 3.1. (i) If $\mathcal{K}$ is a VC-subgraph, by van der Vaart and Wellner ([38], Theorem 2.6.7), there exists finite constants B and α (that are not dependent on n) such that $\mathcal{N}(\mathcal{K}, \rho, \epsilon) \le B(2\nu/\epsilon)^{\alpha}, 0 < \epsilon < 2\nu$, which implies there exists A and $0 < \beta < 1$ such that (3.13) holds. By Theorem 3.3, this implies $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\mathbb{P}}(n^{-1/2})$, and hence by the Borell–Cantelli lemma, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$. Therefore, it is clear that if $\mathcal{K}$ is a uniformly bounded VC-subgraph, then

$$\mathbb{P}_n \rightsquigarrow \mathbb{P} \qquad \text{a.s. as } n \to \infty \tag{Q}$$

with a rate of convergence of $n^{-1/2}$. Ying and Campbell ([40], Lemma 2) – also see Proposition 6.1 – showed that the Gaussian kernel family in Example 1 is a VC-subgraph (in fact, using the proof idea in Lemma 2 of Ying and Campbell [40] it can be easily shown that Laplacian and inverse multiquadric families are also VC-subgraphs) and, therefore, these kernel classes ensure (3.1) with a convergence rate of $n^{-1/2}$. Instead of directly showing the radial basis function (RBF) class in Example 1 to be a VC-subgraph, Ying and Campbell ([40], see the proof of Corollary 1), bounded the expected suprema of the Rademacher chaos process of degree 2, that is,

$$U_n\bigl(\mathcal{K}; (X_i)_{i=1}^n\bigr) := \mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left|\sum_{i<j}^{n} \varepsilon_i \varepsilon_j k(X_i, X_j)\right| \tag{3.16}$$

indexed by the RBF class, $\mathcal{K}$, by that of the Gaussian class (see (6.13) in the proof of Theorem 4.2(a)) and since the Gaussian class is a VC-subgraph, we obtain $U_n(\mathcal{K}; (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$ for the RBF class. We also show in Theorem 4.2(d) that $U_n(\mathcal{K}; (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$ for the Matérn kernel family in Example 1. Using these bounds in (6.4) and following through the proof of Theorem 3.3 yields that (3.1) holds with a convergence rate of $n^{-1/2}$. Note that this rate of convergence is faster than the rate of $n^{-1/d}$, $d \geq 3$ that is obtained with $\|\cdot\|_{\mathcal{F}_{\mathrm{BL}}}$ (Sriperumbudur *et al.* [31], Corollary 3.5). Here, $(\varepsilon_i)_{i=1}^n$ denote i.i.d. Rademacher random variables.

(ii) We would like to mention that Theorem 3.3 is a variation on Theorem 7 in Sriperumbudur *et al.* [30] where $U_n(\mathcal{K}; (X_i)_{i=1}^n)$ is bounded by the entropy integral in de la Peña and Giné ([7], Corollary 5.1.8), with the lower limit of the integral being zero unlike in Theorem 3.3. This generalization (see Mendelson [25], Srebro, Sridharan and Tewari [29] for a similar result to bound the expected suprema of empirical processes) allows to handle the polynomial growth of entropy number for $\beta \geq 1$ compared to Sriperumbudur *et al.* ([30], Theorem 7). Also, compared to Sriperumbudur *et al.* ([30], Theorem 7), we provide explicit constants in Theorem 3.3.

4. Main results

In this section, we present our main results of demonstrating the optimality of the kernel density estimator in $\|\cdot\|_{\mathcal{F}_H}$ through an exponential concentration inequality in Section 4.1 and a uniform central limit theorem in Section 4.2.

4.1. An exponential concentration inequality for $\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H}$

In this section, we present an exponential inequality for the weak convergence of kernel density estimator on $\mathbb{R}^d$ using which we show the optimality of the kernel density estimator in both strong and weak topologies. This is carried out using the ideas in Section 3, in particular through bounding the tail probability of $\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H}$, where $\mathbb{P}_n * K_h$ is the kernel density estimator. Since $\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} \leq \|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} + \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H}$, the result follows from Theorem 3.3 and bounding the tail probability of $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H}$, again through an application of McDiarmid's inequality, which is captured in Theorem 4.1.

Theorem 4.1. *Let $\mathbb{P}$ have a density $p \in W_1^s(\mathcal{X})$, $s \in \mathbb{N}$ with $(X_i)_{i=1}^n$ being samples drawn i.i.d. from $\mathbb{P}$ defined on an open subset $\mathcal{X}$ of $\mathbb{R}^d$. Assume $\mathcal{K}$ satisfies the following*:

(i) *Every $k \in \mathcal{K}$ is translation invariant, that is, $k(x, y) = \psi(x - y)$, $x, y \in \mathcal{X}$, where ψ is a positive definite function on $\mathcal{X}$*;

(ii) *For every $k \in \mathcal{K}$, $\partial^{\alpha,\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all multi-indexes $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$, $m \in \mathbb{N}$, where $\partial^{\alpha,\alpha} := \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} \partial_{1+d}^{\alpha_1} \cdots \partial_{2d}^{\alpha_d}$*;

(iii) $\exists \nu > 0$ *such that* $\sup_{k \in \mathcal{K}, x \in \mathcal{X}} k(x, x) \leq \nu < \infty$;

(iv) *For* $|\alpha| = m \wedge r$, $\exists \nu_\alpha > 0$ *such that* $\sup_{k' \in \mathcal{K}_\alpha, x \in \mathcal{X}} k'(x, x) \leq \nu_\alpha < \infty$ *where* $\mathcal{K}_\alpha := \{\partial^{\alpha,\alpha} k \colon k \in \mathcal{K}\}$, *where* $1 \le r \le m+s$, $r \in \mathbb{N}$ *is the order of the smoothing kernel* K. *Then for every* $\tau > 0$, *with probability at least* $1-2\mathrm{e}^{-\tau}$ *over the choice of* $(X_i)_{i=1}^n$, *there exists finite constants* $(A_i)_{i=1}^2$ *and* $(B_i)_{i=1}^2$ *(dependent only on* $m, r, s, p, K, \tau, \nu, \nu_\alpha$ *and not on* n*) such that*

$$\|K_h * \mathbb{P}_n - \mathbb{P}_n\|_{\mathcal{F}_H} \le 4\sqrt{2}h^{m\wedge r} \sum_{|\alpha|=m\wedge r} \Theta(\alpha)\sqrt{\mathcal{T}(\mathcal{K}_\alpha, \rho_\alpha, \nu_\alpha)} + \frac{A_1 h^{m\wedge r}}{\sqrt{n}} + A_2 h^r \tag{4.1}$$

and

$$\begin{aligned}\|K_h * \mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \le{}& 4\sqrt{2}h^{m\wedge r} \sum_{|\alpha|=m\wedge r} \Theta(\alpha)\sqrt{\mathcal{T}(\mathcal{K}_\alpha, \rho_\alpha, \nu_\alpha)} + 4\sqrt{2}\sqrt{\mathcal{T}(\mathcal{K}, \rho, \nu)} \\ &+ \frac{B_1 h^{m\wedge r}}{\sqrt{n}} + \frac{B_2}{\sqrt{n}} + A_2 h^r,\end{aligned} \tag{4.2}$$

where $m \wedge r := \min(m, r)$,

$$\mathcal{T}(\mathcal{K}_\alpha, \rho_\alpha, \nu_\alpha) := \inf_{\delta>0}\left\{\delta + \frac{4\mathrm{e}}{n}\int_\delta^{2\nu_\alpha} \log 2\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon)\,\mathrm{d}\epsilon\right\},$$

$$\Theta(\alpha) = \int \frac{\prod_{i=1}^d |t_i|^{\alpha_i}}{\prod_{i=1}^d \alpha_i!}\left|K(t)\right|\mathrm{d}t,$$

ρ *is defined as in* (3.12) *and for any* $k_1, k_2 \in \mathcal{K}_\alpha$,

$$\rho_\alpha(k_1, k_2) = \sqrt{\frac{2}{n^2}\sum_{i<j}^n \bigl(k_1(X_i, X_j) - k_2(X_i, X_j)\bigr)^2}.$$

In addition, suppose there exists finite constants C_α, $C_\mathcal{K}$, ω_α *and* $\omega_\mathcal{K}$ *(that are not dependent on* n*) such that*

$$\log\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon) \le C_\alpha\left(\frac{2\nu_\alpha}{\epsilon}\right)^{\omega_\alpha}, \qquad 0 < \epsilon < 2\nu_\alpha, \text{ for } |\alpha| = m\wedge r \tag{4.3}$$

and

$$\log\mathcal{N}(\mathcal{K}, \rho, \epsilon) \le C_\mathcal{K}\left(\frac{2\nu}{\epsilon}\right)^{\omega_\mathcal{K}}, \qquad 0 < \epsilon < 2\nu. \tag{4.4}$$

Define $\omega_\star := \max\{\omega_\alpha\colon |\alpha| = m\wedge r\}$. *If*

$$\sqrt{(\log n)^{\mathbb{1}_{\{\omega_\star=1\}}}}\, n^{((\omega_\star\vee 1)-1)/(2\omega_\star)} h^{m\wedge r} \to 0, \qquad \sqrt{n}h^r \to 0 \text{ as } h\to 0, n\to\infty, \tag{4.5}$$

then

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = \mathrm{o}_{\mathrm{a.s.}}\bigl(n^{-1/2}\bigr) \tag{4.6}$$

and, *therefore*,

$$\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}\Big(\sqrt{(\log n)^{\mathbb{1}_{\{\omega_{\mathcal{K}}=1\}}}}\, n^{-(\omega_{\mathcal{K}}\wedge 1)/(2\omega_{\mathcal{K}})}\Big). \tag{4.7}$$

Remark 4.1. (i) Theorem 4.1 shows that the kernel density estimator with bandwidth, h converging to zero sufficiently fast as given by conditions in (4.5) is within $\|\cdot\|_{\mathcal{F}_H}$-ball of size $\mathrm{o}_{\text{a.s.}}(n^{-1/2})$ around $\mathbb{P}_n$ and behaves like $\mathbb{P}_n$ in the sense that it converges to $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$ at a dimension independent rate of $n^{-1/2}$ (see Theorem 3.3) as long as $\mathcal{K}$ is not too big, which is captured by $\omega_{\mathcal{K}} < 1$ in (4.4). In addition, if $\mathcal{K}$ satisfies the conditions in Theorem 3.2, then the kernel density estimator converges weakly to $\mathbb{P}$ a.s. at the rate of $n^{-1/2}$. Since we are interested in the optimality of $\mathbb{P}_n * K_h$ in both strong and weak topologies, it is interest to understand whether the asymptotic behavior in (4.6) holds for $h^* \simeq n^{-1/(2s+d)}$ where h^* is the optimal bandwidth (of the kernel density estimator) for the estimation of p in L^1 norm. It is easy to verify that if

$$r > s + \frac{d}{2} \quad \text{and} \quad m > \frac{(2s+d)(\omega_\star - 1)}{2\omega_\star} \vee \frac{d}{2} \tag{4.8}$$

then h^* satisfies (4.5) and, therefore, $\|K_{h^*} * \mathbb{P}_n - \mathbb{P}_n\|_{\mathcal{F}_H} = \mathrm{o}_{\text{a.s.}}(n^{-1/2})$ so that $\|K_{h^*} * \mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}(n^{-1/2})$ if $\mathcal{K}$ is not too big. This means for an appropriate choice of $\mathcal{K}$ (i.e., $\omega_{\mathcal{K}} < 1$), the kernel density estimator $K_h * \mathbb{P}_n$ with $h = h^*$ is optimal in both weak (induced by $\|\cdot\|_{\mathcal{F}_H}$) and strong topologies unlike $\mathbb{P}_n$ which is only an optimal estimator of $\mathbb{P}$ in the weak topology. In Theorem 4.2, we present examples of $\mathcal{K}$ for which the kernel density estimator is optimal in both strong and weak topologies (induced by $\|\cdot\|_{\mathcal{F}_H}$). Under the conditions in (4.8), it can be shown that $h^{**} \simeq (n/\log n)^{-1/(2s+d)}$, which is the optimal bandwidth for the estimation of p in sup-norm, also satisfies (4.5) and, therefore, (4.6) and (4.7) hold for $h = h^{**}$.

(ii) The condition on r in (4.8) coincides with the one obtained for $\{\mathbb{1}_{(-\infty,t]}: t \in \mathbb{R}\}$ in Bickel and Ritov [6] and bounded variation and Lipschitz classes with $d = 1$ in Giné and Nickl [15], see Remarks 7 and 8. This condition shows that for the kernel density estimator with bandwidth h^* to be optimal in the weak topology (assuming $\omega_* \le 1$, $\omega_{\mathcal{K}} < 1$ and $\mathcal{K}$ satisfying the conditions in Theorem 3.2), the order of the kernel has to be chosen higher by $\frac{d}{2}$ than the usual (the usual being estimating p using the kernel density estimator in L^1-norm). An interesting aspect of the second condition in (4.8) is that the smoothness of kernels in $\mathcal{K}$ should increase with either d or the size of $\mathcal{K}_\alpha$ for $\mathbb{P}_n * K_h$ with $h = h^*$ or $h = h^{**}$ to lie in $\|\cdot\|_{\mathcal{F}_H}$-ball of size $\mathrm{o}_{\text{a.s.}}(n^{-1/2})$ around $\mathbb{P}_n$. If $\mathcal{K}_\alpha$ is large, that is, $\omega_\star > 1$, then the choice of m depends on the smoothness s of p and therefore s has to be known a priori to pick k appropriately. Also, since the smoothness of kernels in $\mathcal{K}$ should grow with d for (4.6) to hold, it implies that the rate in (4.7) holds under weaker metrics on the space of probabilities. On the other hand, it is interesting to note that as long $\mathcal{K}$ satisfies the conditions in Theorem 3.2, each of these weaker metrics metrize the weak topology.

(iii) If $\mathcal{K}$ is singleton, then it is easy to verify that the first terms in (4.1) and (4.2) are of order $h^{m\wedge r}/\sqrt{n}$ – use the idea in Remark 6.1(ii) for (4.18) – and the second term in (4.2) is of order $n^{-1/2}$ (see (6.5)). Therefore, the claims of Theorem 4.1 hold as if $\omega_\star \le 1$ and $\omega_{\mathcal{K}} \le 1$.

Proof. Note that

$$\|K_h * \mathbb{P}_n - \mathbb{P}_n\|_{\mathcal{F}_H} \le \big\|K_h * (\mathbb{P}_n - \mathbb{P}) - (\mathbb{P}_n - \mathbb{P})\big\|_{\mathcal{F}_H} + \|K_h * \mathbb{P} - \mathbb{P}\|_{\mathcal{F}_H}. \tag{4.9}$$

(a) *Bounding* $\|K_h * (\mathbb{P}_n - \mathbb{P}) - (\mathbb{P}_n - \mathbb{P})\|_{\mathcal{F}_H}$:
By defining $\mathcal{B}_0 := \|K_h * (\mathbb{P}_n - \mathbb{P}) - (\mathbb{P}_n - \mathbb{P})\|_{\mathcal{F}_H}$, we have

$$\begin{aligned}\mathcal{B}_0 &= \sup_{f \in \mathcal{F}_H} \left| \int f(x)\,\mathrm{d}\big(K_h * (\mathbb{P}_n - \mathbb{P})\big)(x) - \int f(x)\,\mathrm{d}(\mathbb{P}_n - \mathbb{P})(x) \right| \\ &= \sup_{f \in \mathcal{F}_H} \left| \int (f * K_h - f)\,\mathrm{d}(\mathbb{P}_n - \mathbb{P}) \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}},\end{aligned}$$

where $\mathcal{G} := \{f * K_h - f \colon f \in \mathcal{F}_H\}$. We now obtain a bound on $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}}$ through an application of McDiarmid's inequality. To this end, consider

$$\|g\|_\infty \le \|f * K_h - f\|_\infty = \sup_{x \in \mathcal{X}} \left| \int \big(f(x + ht) - f(x)\big) K(t)\,\mathrm{d}t \right|. \tag{4.10}$$

Since every $k \in \mathcal{K}$ is m-times differentiable, by Steinwart and Christmann [35], Corollary 4.36, every $f \in \mathcal{F}_H$ is m-times continuously differentiable and for any $k \in \mathcal{K}$, $f \in \mathcal{H}_k$,

$$\big|\partial^\alpha f(x)\big| \le \|f\|_{\mathcal{H}_k} \sqrt{\partial^{\alpha,\alpha} k(x, x)}, \qquad x \in \mathcal{X} \tag{4.11}$$

for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le m$. Therefore, Taylor series expansion of $f(x + th)$ around x gives

$$\begin{aligned} f(x + th) - f(x) = &\sum_{0 < |\alpha| \le (m \wedge r) - 1} h^{|\alpha|} \Lambda_\alpha(t) \partial^\alpha f(x) \\ &+ h^{m \wedge r} \sum_{|\alpha| = m \wedge r} \Lambda_\alpha(t) \partial^\alpha f(x + h D_\theta t), \end{aligned} \tag{4.12}$$

where

$$\Lambda_\alpha(t) := \frac{\prod_{i=1}^d t_i^{\alpha_i}}{\prod_{i=1}^d \alpha_i!},$$

$D_\theta = \operatorname{diag}(\theta_1, \ldots, \theta_d)$ and $0 < \theta_i < 1$ for all $i = 1, \ldots, d$. Using (4.12) in (4.10) along with the regularity of K, we have

$$\begin{aligned} \|g\|_\infty \le \|f * K_h - f\|_\infty &\le h^{m \wedge r} \sup_{x \in \mathcal{X}} \left| \sum_{|\alpha| = m \wedge r} \int \Lambda_\alpha(t) K(t) \partial^\alpha f(x + h D_\theta t)\,\mathrm{d}t \right| \\ &\le h^{m \wedge r} \sup_{x \in \mathcal{X}} \sum_{|\alpha| = m \wedge r} \int \Lambda_\alpha\big(|t|\big) \big|K(t)\big| \big|\partial^\alpha f(x + h D_\theta t)\big|\,\mathrm{d}t, \end{aligned} \tag{4.13}$$

where $|t|_i := |t_i|, \forall i = 1, \dots, d$. Using (4.11) in (4.13), for any $g \in \mathcal{G}$, we get

$$
\begin{aligned}
\|g\|_\infty &\le \|f * K_h - f\|_\infty \\
&\le h^{m\wedge r} \sup_{k\in\mathcal{K},x\in\mathcal{X}} \sum_{|\alpha|=m\wedge r} \int \Lambda_\alpha\big(|t|\big)\big|K(t)\big|\sqrt{\partial^{\alpha,\alpha}k(x+hD_\theta t, x+hD_\theta t)}\,\mathrm{d}t \\
&\overset{\text{(i)}}{\le} h^{m\wedge r} \sum_{|\alpha|=m\wedge r} \sqrt{\sup_{k\in\mathcal{K},x\in\mathcal{X}} \partial^{\alpha,\alpha}k(x,x)} \int \Lambda_\alpha\big(|t|\big)\big|K(t)\big|\,\mathrm{d}t \\
&= L_{m,r}h^{m\wedge r},
\end{aligned}
\tag{4.14}
$$

where

$$L_{m,r} := \sum_{|\alpha|=m\wedge r} \sqrt{\nu_\alpha}\Theta(\alpha) < \infty.$$

Now, let us consider

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{G} \overset{(\star)}{\le} \frac{2}{n}\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i g(X_i)\right| = \frac{2}{n}\mathbb{E}\sup_{f\in\mathcal{F}_H}\left|\sum_{i=1}^n \varepsilon_i (f * K_h - f)(X_i)\right|, \tag{4.15}$$

where we have invoked the symmetrization inequality (van der Vaart and Wellner [38], Lemma 2.3.1) in $(\star)$ with $(\varepsilon_i)_{i=1}^n$ being the Rademacher random variables. By McDiarmid's inequality, for any $\tau > 0$, with probability at least $1 - \mathrm{e}^{-\tau}$,

$$
\begin{aligned}
\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{G} &\le \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{G} + \|g\|_\infty\sqrt{\frac{2\tau}{n}} \\
&\le \frac{2}{n}\mathbb{E}\sup_{f\in\mathcal{F}_H}\left|\sum_{i=1}^n \varepsilon_i (f * K_h - f)(X_i)\right| + L_{m,r}h^{m\wedge r}\sqrt{\frac{2\tau}{n}},
\end{aligned}
\tag{4.16}
$$

where (4.14) and (4.15) are used in (4.16). Define

$$R_n(\mathcal{F}_H) := \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (f * K_h - f)(X_i)\right|,$$

where $\mathbb{E}_\varepsilon$ denotes the expectation w.r.t. $(\varepsilon_i)_{i=1}^n$ conditioned on $(X_i)_{i=1}^n$. Applying McDiarmid's inequality to $R_n(\mathcal{F}_H)$, we have for any $\tau > 0$, with probability at least $1 - \mathrm{e}^{-\tau}$,

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (f * K_h - f)(X_i)\right| \\
&\quad \le R_n(\mathcal{F}_H) + L_{m,r}h^{m\wedge r}\sqrt{\frac{2\tau}{n}}.
\end{aligned}
\tag{4.17}
$$

Bounding $R_n(\mathcal{F}_H)$ yields

$$
\begin{aligned}
&R_n(\mathcal{F}_H)\\
&\quad = \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \int \big(f(X_i+th)-f(X_i)\big)K(t)\,\mathrm{d}t\right|\\
&\quad \overset{(4.12)}{=} \frac{h^{m\wedge r}}{n}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}_H}\left|\sum_{|\alpha|=m\wedge r}\int \Lambda_\alpha(t)K(t)\sum_{j=1}^n \varepsilon_j \partial^\alpha f(X_j+hD_\theta t)\,\mathrm{d}t\right|\\
&\quad = \frac{h^{m\wedge r}}{n}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}_H}\left|\sum_{|\alpha|=m\wedge r}\int \Lambda_\alpha(t)K(t)\left\langle f,\sum_{j=1}^n \varepsilon_j \partial^\alpha k(\cdot,X_j+hD_\theta t)\right\rangle_{\mathcal{H}_k}\mathrm{d}t\right|\\
&\quad \le \frac{h^{m\wedge r}}{n}\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\sum_{|\alpha|=m\wedge r}\int \Lambda_\alpha\big(|t|\big)\big|K(t)\big|\left\|\sum_{j=1}^n \varepsilon_j \partial^\alpha k(\cdot,X_j+hD_\theta t)\right\|_{\mathcal{H}_k}\mathrm{d}t\\
&\quad = \frac{h^{m\wedge r}}{n}\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\sum_{|\alpha|=m\wedge r}\int \Lambda_\alpha\big(|t|\big)\big|K(t)\big|\sqrt{\sum_{i,j=1}^n \varepsilon_i\varepsilon_j \partial^{\alpha,\alpha}k(X_i+hD_\theta t,X_j+hD_\theta t)}\,\mathrm{d}t.
\end{aligned}
$$

Since k is translation invariant, we have

$$
\begin{aligned}
R_n(\mathcal{F}_H) &= \frac{h^{m\wedge r}}{n}\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\sum_{|\alpha|=m\wedge r}\Theta(\alpha)\sqrt{\sum_{i,j=1}^n \varepsilon_i\varepsilon_j\partial^{\alpha,\alpha}k(X_i,X_j)} &(4.18)\\
&\le \frac{\sqrt{2}h^{m\wedge r}}{n}\sum_{|\alpha|=m\wedge r}\Theta(\alpha)\sqrt{\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j\partial^{\alpha,\alpha}k(X_i,X_j)\right|}+\frac{h^{m\wedge r}L_{m,r}}{\sqrt{n}}\\
&= \frac{\sqrt{2}h^{m\wedge r}}{n}\sum_{|\alpha|=m\wedge r}\Theta(\alpha)\sqrt{\mathbb{E}_\varepsilon \sup_{k'\in\mathcal{K}_\alpha}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j k'(X_i,X_j)\right|}+\frac{h^{m\wedge r}L_{m,r}}{\sqrt{n}} &(4.19)\\
&\overset{(\dagger)}{\le} 2\sqrt{2}h^{m\wedge r}\sum_{|\alpha|=m\wedge r}\Theta(\alpha)\sqrt{\mathcal{T}(\mathcal{K}_\alpha,\rho_\alpha,\nu_\alpha)}+\frac{3L_{m,r}h^{m\wedge r}}{\sqrt{n}}, &(4.20)
\end{aligned}
$$

where we used Lemma A.2 in (†) with $\theta=\frac{3}{4}$. Combining (4.16), (4.17) and (4.20), we have that for any $\tau>0$, with probability at least $1-2\mathrm{e}^{-\tau}$,

$$
\mathcal{B}_0 \le 4\sqrt{2}h^{m\wedge r}\sum_{|\alpha|=m\wedge r}\Theta(\alpha)\sqrt{\mathcal{T}(\mathcal{K}_\alpha,\rho_\alpha,\nu_\alpha)}+\frac{A_1h^{m\wedge r}}{\sqrt{n}}, \tag{4.21}
$$

where $A_1:=(6+\sqrt{18\tau})L_{m,r}$.

(b) *Bounding* $\|K_h*\mathbb{P}-\mathbb{P}\|_{\mathcal{F}_H}$:

Defining $\mathcal{B}_1 := \|\mathbb{P} * K_h - \mathbb{P}\|_{\mathcal{F}_H}$, we have

$$\begin{aligned}\mathcal{B}_1 &= \sup_{f\in\mathcal{F}_H}\left|\int f(x)\,\mathrm{d}(\mathbb{P}*K_h)(x) - \int f(x)\,\mathrm{d}\mathbb{P}(x)\right| \\ &= \sup_{f\in\mathcal{F}_H}\left|\frac{1}{h^d}\int\int f(x)K\left(\frac{x-y}{h}\right)\mathrm{d}\mathbb{P}(y)\,\mathrm{d}x - \int f(x)\,\mathrm{d}\mathbb{P}(x)\right| \\ &= \sup_{f\in\mathcal{F}_H}\left|\int\left(\int\big(f(x+th)-f(x)\big)\,\mathrm{d}\mathbb{P}(x)\right)K(t)\,\mathrm{d}t\right| \\ &= \sup_{f\in\mathcal{F}_H}\left|\int\big((\tilde{f}*p)(ht)-(\tilde{f}*p)(0)\big)K(t)\,\mathrm{d}t\right|,\end{aligned} \tag{4.22}$$

where $\tilde{f}(x) = f(-x)$. Since $p \in L^1(\mathbb{R}^d)$ and $\partial^\alpha f$ is bounded for all $|\alpha| \le m$, by Folland [12], Proposition 8.10, we have $\partial^\alpha(f * p) = (\partial^\alpha f) * p$ for $|\alpha| \le m$. In addition, since $\partial^\alpha f$ is continuous for all $|\alpha| \le m$ and $\partial^\beta p \in L^1(\mathbb{R}^d)$ for $|\beta| \le s$, by extension of Giné and Nickl [15], Lemma 5(b), to $\mathbb{R}^d$, we have $\partial^{\alpha+\beta}(f*p) = \partial^\beta((\partial^\alpha f)*p) = (\partial^\alpha f)*(\partial^\beta p)$ for $|\alpha| \le m$, $|\beta| \le s$, which means for all $f \in \mathcal{F}_H$, $f * p$ is $m+s$-differentiable. Therefore, using the Taylor series expansion of $(\tilde{f} * p)(ht)$ around zero (as in (4.12)) along with the regularity of K in (4.22), we have

$$\begin{aligned}\mathcal{B}_1 &= \sup_{f\in\mathcal{F}_H}\left|h^r\sum_{|\alpha|+|\beta|=r}\int\Lambda_{\alpha+\beta}(t)K(t)\big(\partial^\alpha\tilde{f}*\partial^\beta p\big)(hD_\theta t)\,\mathrm{d}t\right| \\ &\le h^r\sup_{f\in\mathcal{F}_H}\sum_{|\alpha|+|\beta|=r}\int\Lambda_{\alpha+\beta}\big(|t|\big)\big|K(t)\big|\big|\big(\partial^\alpha\tilde{f}*\partial^\beta p\big)(hD_\theta t)\big|\,\mathrm{d}t \\ &\le h^r\sup_{f\in\mathcal{F}_H}\sum_{|\alpha|+|\beta|=r}\int\Lambda_{\alpha+\beta}\big(|t|\big)\big|K(t)\big|\big(\big|\partial^\alpha\tilde{f}\big|*\big|\partial^\beta p\big|\big)(hD_\theta t)\,\mathrm{d}t.\end{aligned} \tag{4.23}$$

Since

$$\begin{aligned}\big(\big|\partial^\alpha\tilde{f}\big|*\big|\partial^\beta p\big|\big)(hD_\theta t) &= \int\big|\partial^\alpha f(x-hD_\theta t)\big|\big|\partial^\beta p(x)\big|\,\mathrm{d}x \\ &\overset{(4.11)}{\le} \int\sqrt{\partial^{\alpha,\alpha}k(x-hD_\theta t, x-hD_\theta t)}\big|\partial^\beta p(x)\big|\,\mathrm{d}x \\ &\overset{(\mathrm{i})}{\le} \big\|\partial^\beta p\big\|_{L^1(\mathbb{R}^d)}\sqrt{\sup_{k\in\mathcal{K},x\in\mathcal{X}}\partial^{\alpha,\alpha}k(x,x)},\end{aligned} \tag{4.24}$$

using (4.24) in (4.23), we obtain

$$\mathcal{B}_1 \le A_2 h^r, \tag{4.25}$$

where $A_2 := \sum_{|\alpha|+|\beta|=r} \Theta(\alpha+\beta)\sqrt{\nu_\alpha}\|\partial^\beta p\|_{L^1(\mathbb{R}^d)}$ and $(\alpha+\beta)_i = \alpha_i + \beta_i, \forall i = 1, \dots, d$. Using (4.21) and (4.25) in (4.9), we obtain the result in (4.1). Since

$$\|K_h * \mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \le \|K_h * \mathbb{P}_n - \mathbb{P}_n\|_{\mathcal{F}_H} + \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H}, \tag{4.26}$$

the result in (4.2) follows from Theorem 3.3 and (4.1). Under the entropy number conditions in (4.3), it is easy to check (see (3.15)) that

$$\sum_{|\alpha|=m\wedge r} \Theta(\alpha)\sqrt{\mathcal{T}(\mathcal{K}_\alpha, \rho_\alpha, \nu_\alpha)} = \mathrm{O}\Big(\sqrt{(\log n)^{\mathbb{1}_{\{\omega_\star=1\}}}}\, n^{-(\omega_\star\wedge 1)/(2\omega_\star)}\Big)$$

and, therefore, (4.6) holds if h satisfies (4.5). Using (4.6) and (3.14) in (4.26), the result in (4.7) follows under the assumption that $\mathcal{K}$ satisfies (4.4). □

Remark 4.2. Since every $k \in \mathcal{K}$ is translation invariant, an alternate proof can be provided by using the representation for $\mathfrak{D}_k$ (following (3.9)) in Proposition 3.1: $\|\mathbb{P} - \mathbb{Q}\|_{\mathcal{F}_H} = \sup_\Upsilon \|\phi_\mathbb{P} - \phi_\mathbb{Q}\|_{L^2(\mathbb{R}^d,\Upsilon)}$, where the supremum is taken over all finite nonnegative Borel measures on $\mathbb{R}^d$. In this case, conditions on the derivatives of $k \in \mathcal{K}$ translate into moment requirements for Υ. However, the current proof is more transparent as it clearly shows why the translation invariance of k is needed; see (4.14) and (4.18).

In the following result (proved in Section 6.3), we present some families of $\mathcal{K}$ that ensure the claims of Theorems 3.3 and 4.1.

Theorem 4.2. *Suppose the assumptions on $\mathbb{P}$ and K in Theorem 4.1 hold and let $0 < a < \infty$. Then for the following classes of kernels,*

(a)

$$\mathcal{K} = \{k(x,y) = \psi_\sigma(x-y), x, y \in \mathbb{R}^d \colon \sigma \in \Sigma\},$$

where $\psi_\sigma(x) = \mathrm{e}^{-\sigma\|x\|_2^2}$ *and* $\Sigma := (0, a]$;

(b)

$$\mathcal{K} = \left\{k(x,y) = \int_0^\infty \psi_\sigma(x-y)\,\mathrm{d}\Lambda(\sigma), x, y \in \mathbb{R}^d \colon \Lambda \in \mathcal{M}_A\right\},$$

where

$$\mathcal{M}_A := \left\{\Lambda \in M_+^1\big((0,\infty)\big) \colon \int_0^\infty \sigma^r\,\mathrm{d}\Lambda(\sigma) \le A < \infty\right\}$$

for some fixed $A > 0$;

(c)

$$\mathcal{K} = \left\{k(x,y) = \int_{(0,\infty)^d} \mathrm{e}^{-(x-y)^T\Delta(x-y)}\,\mathrm{d}\Lambda(\Delta), x, y \in \mathbb{R}^d \colon \Lambda \in \mathcal{Q}_A\right\},$$

where $\Delta := \operatorname{diag}(\sigma_1, \dots, \sigma_d)$,

$$\mathcal{Q}_A := \left\{\Lambda \in M_+^1\big((0,\infty)^d\big) \,\Big|\, \Lambda = \bigotimes_{i=1}^{d} \Lambda_i \colon \Lambda_i \in \mathcal{M}_{A_i}, i = 1, \dots, d\right\}$$

and

$$\mathcal{M}_{A_i} := \left\{\Lambda_i \in M_+^1\big((0,\infty)\big) \colon \sup_{j\in\{1,\dots,d\}} \int_0^\infty \sigma^{\alpha_j}\, \mathrm{d}\Lambda_i(\sigma) \le A_i < \infty\right\}$$

for some fixed constant $A := (A_1, \dots, A_d) \in (0,\infty)^d$ *with* $\sum_{i=1}^d \alpha_i = r$ *and* $\alpha_i \ge 0, \forall i = 1, \dots, d$;

(d)

$$\mathcal{K} = \left\{k(x,y) = A\frac{\|x-y\|_2^{\beta-d/2}}{c^{d/2-\beta}}\mathfrak{K}_{d/2-\beta}\big(c\|x-y\|_2\big), x, y \in \mathbb{R}^d, \beta > m + \frac{d}{2} \colon c \in \Sigma\right\},$$

where $A := \frac{2^{d/2+1-\beta}}{\Gamma(\beta-d/2)}$ *and* $m \in \mathbb{N}$,

$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = O_{\text{a.s.}}(n^{-1/2})$, $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = o_{\text{a.s.}}(n^{-1/2})$ *and* $\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} = O_{\text{a.s.}}(n^{-1/2})$ *for any* h *satisfying* $\sqrt{n}h^r \to 0$ *as* $h \to 0$ *and* $n \to \infty$, *which is particularly satisfied by* $h = h^*$ *and* $h = h^{**}$ *if* $r > s + \frac{d}{2}$, *where* h^* *and* h^{**} *are defined in Remark* 4.1(i).

Remark 4.3. (i) The Gaussian RKHS in (3.3) has the property that $\mathcal{H}_\sigma \subset \mathcal{H}_\tau$ if $0 < \sigma < \tau < \infty$, where $\mathcal{H}_\sigma$ is the Gaussian RKHS induced by ψ_σ. This follows since for any $f \in \mathcal{H}_\sigma$,

$$\begin{aligned}\|f\|_{\mathcal{H}_\tau}^2 &:= (4\pi\tau)^{d/2}\int \big|\widehat{f}(\omega)\big|^2 \mathrm{e}^{\|\omega\|_2^2/(4\tau)}\,\mathrm{d}\omega \\ &= (4\pi\tau)^{d/2}\int \big|\widehat{f}(\omega)\big|^2 \mathrm{e}^{\|\omega\|_2^2/(4\sigma)}\mathrm{e}^{\|\omega\|_2^2(1/(4\tau)-1/(4\sigma))}\,\mathrm{d}\omega \le \left(\frac{\tau}{\sigma}\right)^{d/2}\|f\|_{\mathcal{H}_\sigma}^2,\end{aligned} \tag{4.27}$$

which implies for any $k \in \mathcal{K}$, $\mathcal{H}_k \subset \mathcal{H}_a$, where the definition of $\|\cdot\|_{\mathcal{H}_\tau}^2$ for any $\tau > 0$ in the first line of (4.27) is obtained from Wendland [39], Theorem 10.12. From (4.27), it follows that

$$\left\{\|f\|_{\mathcal{H}_\sigma} \le \left(\frac{\sigma}{a}\right)^{d/4} \colon f \in \mathcal{H}_\sigma, \sigma \in (0,a]\right\} \subset \big\{f \in \mathcal{H}_a \colon \|f\|_{\mathcal{H}_a} \le 1\big\} \subset \mathcal{F}_H$$

and

$$\mathcal{F}_H \subset \bigcup_{\sigma\in(0,a]}\left\{f \in \mathcal{H}_a \colon \|f\|_{\mathcal{H}_a} \le \left(\frac{a}{\sigma}\right)^{d/4}\right\} = \mathcal{H}_a, \tag{4.28}$$

where $\mathcal{F}_H := \{\|f\|_{\mathcal{H}_\sigma} \le 1 \colon f \in \mathcal{H}_\sigma, \sigma \in (0,a]\}$.

(ii) The kernel classes in (b) and (c) above are generalizations of the Gaussian family in (a). This can be seen by choosing $\mathcal{M}_A = \{\delta_\sigma \colon \sigma \in \Sigma\}$ in (b) where $A = a^r$ and $\mathcal{M}_{A_i} = \{\delta_\sigma : \sigma \in \Sigma\}$

in (c) with $A_i = a^r$ for all $i = 1, \dots, d$. By choosing

$$\mathcal{M}_A = \left\{ \mathrm{d}\Lambda(\sigma) = \frac{c^{2\beta}}{\Gamma(\beta)} \sigma^{\beta-1} \mathrm{e}^{-\sigma c^2} \, \mathrm{d}\sigma, \beta > 0\colon\ c \in [a, \infty), a > 0 \right\}$$

in (b), the inverse multiquadrics kernel family,

$$\mathcal{K} = \left\{ k(x, y) = \left(1 + \left\| \frac{x-y}{c} \right\|_2^2 \right)^{-\beta}, x, y \in \mathbb{R}^d, \beta > 0\colon\ c \in [a, \infty), a > 0 \right\} \tag{4.29}$$

mentioned in Example 2 is obtained, where $A = a^{-2r} \frac{\Gamma(r+\beta)}{\Gamma(\beta)}$ with Γ being the Gamma function. Similarly, choosing

$$\mathcal{M}_{A_i} = \left\{ \mathrm{d}\Lambda_i(\sigma) = c_i^2 \mathrm{e}^{-\sigma c_i^2} \, \mathrm{d}\sigma : c_i \in [a_i, \infty), a_i > 0 \right\}$$

yields the family

$$\mathcal{K} = \left\{ \prod_{i=1}^{d} \left(1 + \frac{|x_i - y_i|^2}{c_i^2} \right)^{-1}, x, y \in \mathbb{R}^d\colon\ c_i \in [a_i, \infty), a_i > 0, \forall i = 1, \dots, d \right\}$$

in Example 2 where $A_i = \sup_{j \in \{1,\dots,d\}} \alpha_j ! a_i^{-2\alpha_j}$. It is easy to verify that these classes of kernels metrize the weak topology on $M_+^1(\mathbb{R}^d)$.

(iii) Suppose there exists $B > 0$ and $\delta > 0$ such that $\inf_{\Lambda \in \mathcal{M}_A} \int_0^\infty \mathrm{e}^{-\delta\sigma^2} \, \mathrm{d}\Lambda(\sigma) \ge B$ (similarly, there exists $B_i > 0$ and $\delta_i > 0$ such that $\inf_{\Lambda_i \in \mathcal{M}_{A_i}} \int_0^\infty \mathrm{e}^{-\delta_i\sigma^2} \, \mathrm{d}\Lambda_i(\sigma) \ge B_i, i = 1, \dots, d$), where $\mathcal{M}_A$ and $(\mathcal{M}_{A_i})_{i=1}^d$ are defined in (b) and (c) of Theorem 4.2. Then it is easy to show (see Section 6.4 for a proof) that $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology on $M_+^1(\mathbb{R}^d)$, which when combined with the result in Theorem 4.2 yields that for $\mathcal{K}$ in (a)–(c),

$$\mathbb{P}_n \rightsquigarrow \mathbb{P} \quad \text{and} \quad \mathbb{P}_n * K_h \rightsquigarrow \mathbb{P} \qquad \text{a.s.}$$

at the rate of $n^{-1/2}$.

(iv) It is clear from (3.4) that any k in the Matérn family in Example 1 – the family in Theorem 4.2(d) is a special case of this – induces an RKHS which is a Sobolev space,

$$H_c := H_2^{\beta,c} = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)\colon\ \int \left(c^2 + \|\omega\|_2^2\right)^\beta \left|\widehat{f}(\omega)\right|^2 \mathrm{d}\omega < \infty \right\},$$

where $\beta > d/2$ and $c > 0$. Similar to the Gaussian kernel family, it can be shown that $H_c \subset H_\alpha$ for $0 < c < \alpha < \infty$ since for any $f \in H_c$,

$$\|f\|_{H_\alpha}^2 := \frac{2^{1-\beta}}{A\alpha^{2\beta-d}\Gamma(\beta)} \int \left(\alpha^2 + \|\omega\|_2^2\right)^\beta \left|\widehat{f}(\omega)\right|^2 \mathrm{d}\omega \le \left(\frac{\alpha}{c}\right)^d \|f\|_{H_c}^2,$$

where the definition of $\|\cdot\|_{H_\alpha}$ follows from Wendland [39], Theorems 6.13 and 10.12. Therefore, we have

$$\left\{\|f\|_{H_c} \le \left(\frac{c}{a}\right)^{d/2}: f \in H_c, c \in (0,a]\right\} \subset \left\{f \in H_a: \|f\|_{H_a} \le 1\right\} \subset \mathcal{F}_H$$

and

$$\mathcal{F}_H \subset \bigcup_{c\in(0,a]} \left\{f \in H_a: \|f\|_{H_a} \le \left(\frac{a}{c}\right)^{d/2}\right\} = H_a, \tag{4.30}$$

where $\mathcal{F}_H := \{\|f\|_{H_c} \le 1: f \in H_c, c \in (0,a]\}$. Unlike in Example 1, $\mathcal{K}$ in Theorem 4.2(d) requires $\beta > m + \frac{d}{2}$. This is to ensure that every $k \in \mathcal{K}$ is m-times continuously differentiable as required in Theorem 4.1, which is guaranteed by the Sobolev embedding theorem (Folland [12], Theorem 9.17) if $\beta > m + \frac{d}{2}$. Also, since $\mathcal{K}$ metrizes the weak topology (holds for the Gaussian family as well) on $M_+^1(\mathbb{R}^d)$ (see Example 2), we obtain that

$$\mathbb{P}_n \rightsquigarrow \mathbb{P} \quad \text{and} \quad \mathbb{P}_n * K_h \rightsquigarrow \mathbb{P} \qquad \text{a.s.}$$

at the rate of $n^{-1/2}$.

4.2. Uniform central limit theorem

So far, we have presented exponential concentration inequalities for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H}$ and $\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H}$ in Theorems 3.3 and 4.1, respectively, and showed that $\sqrt{n}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}(1)$ and $\sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}(1)$ for families of $\mathcal{K}$ in Theorem 4.2. It is therefore easy to note that if $\mathcal{F}_H$ is $\mathbb{P}$-Donsker, then $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P}$ and so $\sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P}$ (as $\sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = \mathrm{o}_{\text{a.s.}}(1)$) for any h satisfying $\sqrt{n}h^r \to 0$ as $h \to 0$ and $n \to \infty$. Here, $\mathbb{G}_\mathbb{P}$ denotes the $\mathbb{P}$-Brownian bridge indexed by $\mathcal{F}_H$. However, unlike Theorems 3.3 and 4.1 which hold for a general $\mathcal{F}_H$, it is not easy to verify the $\mathbb{P}$-Donsker property of $\mathcal{F}_H$ for any general $\mathcal{K}$. In particular, it is not easy to check whether there exists a pseudometric on $\mathcal{F}_H$ such that $\mathcal{F}_H$ is totally bounded (w.r.t. that pseudometric) and $\mathcal{F}_H$ satisfies the asymptotic equicontinuity condition (see Dudley [10], Theorem 3.7.2) or $\mathcal{F}_H$ satisfies the uniform entropy condition (see van der Vaart and Wellner [38], Theorem 2.5.2) as obtaining estimates on the $L^2(\mathbb{P}_n)$ covering number of $\mathcal{F}_H$ does not appear to be straightforward. On the other hand, in the following result (which is proved in Section 6.5), we show that $\mathcal{F}_H$ is $\mathbb{P}$-Donsker for the classes considered in Theorem 4.2 (with a slight restriction to the parameter space) and, therefore, UCLT in $\ell^\infty(\mathcal{F}_H)$ holds. A similar result holds for any general $\mathcal{K}$ (other than the ones in Theorem 4.2), if $\mathcal{K}$ is singleton consisting of a bounded continuous kernel.

Theorem 4.3. *Suppose the assumptions on $\mathbb{P}$ and K in Theorem 4.1 hold and let $0 < a < b < \infty$. Define $\Sigma := [a, b]$. Then for the following classes of kernels,*

(a)

$$\mathcal{K} = \left\{k(x,y) = \mathrm{e}^{-\sigma\|x-y\|_2^2}, x, y \in \mathbb{R}^d: \sigma \in \Sigma\right\};$$

(b)

$$\mathcal{K}=\left\{k(x,y)=\left(1+\left\|\frac{x-y}{c}\right\|_2^2\right)^{-\beta}, x,y\in\mathbb{R}^d, \beta>0\colon c\in\Sigma\right\};$$

(c)

$$\mathcal{K}=\left\{k(x,y)=A\frac{\|x-y\|_2^{\beta-d/2}}{c^{d/2-\beta}}\mathfrak{K}_{d/2-\beta}\big(c\|x-y\|_2\big), x,y\in\mathbb{R}^d, \beta>m+\frac{d}{2}\colon c\in\Sigma\right\},$$

where $A:=\frac{2^{d/2+1-\beta}}{\Gamma(\beta-d/2)}$ *and* $m\in\mathbb{N}$;

(d) $\mathcal{K}=\{k\}$ *where* k *satisfies the conditions in Theorem* 4.1,

$\sqrt{n}(\mathbb{P}_n-\mathbb{P})\rightsquigarrow_{\ell^\infty(\mathcal{F}_H)}\mathbb{G}_\mathbb{P}$ *and for any* h *satisfying* $\sqrt{n}h^r\to 0$ *as* $h\to 0$ *and* $n\to\infty$, *we have*

$$\sqrt{n}(\mathbb{P}_n*K_h-\mathbb{P})\rightsquigarrow_{\ell^\infty(\mathcal{F}_H)}\mathbb{G}_\mathbb{P},$$

which particularly holds for h^* *and* h^{**} *if* $r>s+\frac{d}{2}$, *where* h^* *and* h^{**} *are defined in Remark* 4.1(i).

Theorem 7 in Giné and Nickl [15] shows the above result for Matérn kernels (i.e., $\mathcal{K}$ in (c) with $c=1$ and $d=1$), but here we generalize it to a wide class of kernels. Theorem 4.3(d) shows that all the kernels (with the parameter fixed a priori, for example, σ in the Gaussian kernel) we have encountered so far – such as in Examples 1 and 2 – satisfy the conditions in Theorem 4.3 and, therefore, yield a UCLT. Note that the kernel classes, $\mathcal{K}$ in Theorem 4.3 are slightly constrained compared to those in Theorem 4.2 and Remark 4.3(ii). This restriction in the kernel class is required as the proof of $\mathcal{F}_H$ being $\mathbb{P}$-Donsker (which in combination with Slutsky's lemma and Theorem 4.2 yields the desired result in Theorem 4.3) critically hinges on the inclusion result shown in (4.28) and (4.30); also see (6.19) for such an inclusion result for $\mathcal{K}$ in Theorem 4.3(b). However, this technique is not feasible for the kernel classes, (b) and (c) in Theorem 4.2 to be shown as $\mathbb{P}$-Donsker, while we reiterate that for any general $\mathcal{K}$, it is usually difficult to check for the Donsker property of $\mathcal{F}_H$.

Combining Theorems 3.2, 4.2 and 4.3, we obtain that the kernel density estimator with bandwidth h^* is an optimal estimator of p in both strong and weak topologies unlike $\mathbb{P}_n$, which estimates $\mathbb{P}$ optimally only in the weak topology. While this optimality result holds in $d=1$ when using $\|\cdot\|_{\mathcal{F}_{\mathrm{BL}}}$ as the loss to measure the optimality of $\mathbb{P}_n*K_h$ in the weak sense, the result does not hold for $d\ge 2$ as discussed before. In addition, for $d=1$, the UCLT for $\sqrt{n}(\mathbb{P}_n-\mathbb{P})$ and $\sqrt{n}(\mathbb{P}_n*K_h-\mathbb{P})$ in $\ell^\infty(\mathcal{F}_{\mathrm{BL}})$ holds only under a certain moment condition on $\mathbb{P}$, that is, $\int|x|^{1+\gamma}\,d\mathbb{P}(x)<\infty$ for some $\gamma>0$ (see Giné and Zinn [20], Theorem 2) while no such condition on $\mathbb{P}$ is required to obtain the UCLT for the above processes in $\ell^\infty(\mathcal{F}_H)$ though both $\|\cdot\|_{\mathcal{F}_{\mathrm{BL}}}$ and $\|\cdot\|_{\mathcal{F}_H}$ metrize the weak topology on $M_+^1(\mathbb{R}^d)$.

5. Discussion

So far we have shown that the kernel density estimator on $\mathbb{R}^d$ with an appropriate choice of bandwidth is an optimal estimator of $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$, that is, in weak topology, similar to $\mathbb{P}_n$. In

Section 5.1, we present a similar result for an alternate metric $\|\cdot\|_{\mathcal{K}_{\mathcal{X}}}$ (defined below) that is topologically equivalent to $\|\cdot\|_{\mathcal{F}_H}$, that is, metrizes the weak topology on $M_+^1(\mathcal{X})$ where $\mathcal{X}$ is a topological space and $\mathcal{K}_{\mathcal{X}} \subset \mathcal{F}_H$, showing that $\mathcal{F}_H$ is not the only class that guarantees the optimality of kernel density estimator in weak and strong topologies. While a result similar to these is shown in $\|\cdot\|_{\mathcal{F}_{\mathrm{BL}}}$ for $d=1$ in Giné and Nickl [16], there is a significant computational advantage associated with $\mathcal{F}_H$ over $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{F}_{\mathrm{BL}}$ in the context of constructing adaptive estimators that are optimal in both strong and weak topologies, which we discuss in Section 5.2.

5.1. Optimality in $\|\cdot\|_{\mathcal{K}_{\mathcal{X}}}$

In this section, we consider an alternate metric, $\|\cdot\|_{\mathcal{K}_{\mathcal{X}}}$, which we show in Proposition 5.1 (proved in Section 6.6) to be topologically equivalent to $\|\cdot\|_{\mathcal{F}_H}$ if $\mathcal{K}$ is uniformly bounded, where

$$\mathcal{K}_{\mathcal{X}} := \{k(\cdot, x)\colon\ k \in \mathcal{K}, x \in \mathcal{X}\}$$

and $\mathcal{X}$ is a topological space. Note that $\mathcal{K}_{\mathcal{X}} \subset \mathcal{F}_H$ if $k(x,x) \le 1, \forall x \in \mathcal{X}, k \in \mathcal{K}$, which means a reduced subset of $\mathcal{F}_H$ is sufficient to metrize the weak topology on $M_+^1(\mathcal{X})$.

Proposition 5.1. *Suppose* $\nu := \sup_{k\in\mathcal{K}, x\in\mathcal{X}} k(x,x) < \infty$. *Then for any* $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$

$$\nu^{-1/2}\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_{\mathcal{X}}} \le \|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H} \le \sqrt{2\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_{\mathcal{X}}}}, \tag{5.1}$$

where

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_{\mathcal{X}}} = \sup_{k\in\mathcal{K}} \left\| \int k(\cdot, x)\,\mathrm{d}\mathbb{P}(x) - \int k(\cdot, x)\,\mathrm{d}\mathbb{Q}(x) \right\|_\infty.$$

In addition if $\mathcal{K}$ *satisfies the assumptions in Theorem* 3.2, *then for any sequence* $(\mathbb{P}_{(n)})_{n\in\mathbb{N}} \subset M_+^1(\mathcal{X})$ *and* $\mathbb{P} \in M_+^1(\mathcal{X})$,

$$\|\mathbb{P}_{(n)}-\mathbb{P}\|_{\mathcal{K}_{\mathcal{X}}} \to 0 \iff \|\mathbb{P}_{(n)}-\mathbb{P}\|_{\mathcal{F}_H} \to 0 \iff \mathbb{P}_{(n)} \rightsquigarrow \mathbb{P} \quad \text{as } n \to \infty. \tag{5.2}$$

From (5.1), it simply follows that

$$\sqrt{n}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{K}_{\mathcal{X}}} = \mathrm{O}_{\mathrm{a.s.}}(1), \qquad \sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{K}_{\mathcal{X}}} = \mathrm{o}_{\mathrm{a.s.}}(1)$$

and

$$\sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{K}_{\mathcal{X}}} = \mathrm{O}_{\mathrm{a.s.}}(1)$$

for any $\mathcal{K}$ in Theorem 4.1 with $\omega_* < 1$ and $\omega_{\mathcal{K}} < 1$ (and, therefore, for any $\mathcal{K}$ in Theorem 4.2) with h satisfying $\sqrt{n}h^r \to 0$ as $h \to 0$ and $n \to \infty$. Therefore, if $\mathcal{K}_{\mathcal{X}}$ is $\mathbb{P}$-Donsker, then for any h satisfying these conditions, we obtain

$$\sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{K}_{\mathcal{X}})} \mathbb{G}_{\mathbb{P}}.$$

The following result (proved in Section 6.7) shows that $\mathcal{K}_{\mathcal{X}}$ is a universal Donsker class (i.e., $\mathbb{P}$-Donsker for all probability measures $\mathbb{P}$ on $\mathbb{R}^d$) for $\mathcal{K}$ considered in Theorem 4.2 and therefore we obtain UCLT for $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $\sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P})$ in $\ell^\infty(\mathcal{K}_{\mathcal{X}})$.

Theorem 5.2. *Suppose the assumptions on $\mathbb{P}$ and K in Theorem 4.1 hold. Define $\mathcal{K}_{\mathcal{X}} := \{k(\cdot, x)\colon k \in \mathcal{K}, x \in \mathcal{X}\}$. Then for $\mathcal{K}$ in Theorem 4.2, $\mathcal{K}_{\mathcal{X}}$ is a universal Donsker class and*

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{K}_{\mathcal{X}})} \mathbb{G}_{\mathbb{P}} \quad \text{and} \quad \sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{K}_{\mathcal{X}})} \mathbb{G}_{\mathbb{P}},$$

for h satisfying $\sqrt{n}h^r \to 0$ as $h \to 0$ and $n \to \infty$, which particularly holds for h^* and h^{**} if $r > s + \frac{d}{2}$, where h^* and h^{**} are defined in Remark 4.1(i).

Combining Theorem 3.2 and Proposition 5.1, along with Theorems 4.2 and 5.2, it is clear that the kernel density estimator with bandwidth h^* is an optimal estimator of p in both strong and weak topologies (induced by $\|\cdot\|_{\mathcal{K}_{\mathcal{X}}}$). While this result matches with the one obtained for $\|\cdot\|_{\mathcal{F}_H}$, by comparing Theorems 4.3 and 5.2, we note that the convergence in $\ell^\infty(\mathcal{K}_{\mathcal{X}})$ does not require the restriction in the parameter space as imposed in kernel classes for convergence in $\ell^\infty(\mathcal{F}_H)$ in Theorem 4.3. However, we show in the following section that $\|\cdot\|_{\mathcal{F}_H}$ is computationally easy to deal with than $\|\cdot\|_{\mathcal{K}_{\mathcal{X}}}$.

5.2. Adaptive estimation and computation

Let us return to the fact that there exists estimators that are $o_{\mathbb{P}}(n^{-1/2})$ from $\mathbb{P}_n$ in $\|\cdot\|_{\mathcal{F}}$ (for suitable choice of $\mathcal{F}$) and behave statistically similar to $\mathbb{P}_n$. While we showed this fact through Theorems 4.1 and 4.2 for the kernel density estimator with $\mathcal{F} = \mathcal{F}_H$ (and Proposition 5.1 for $\mathcal{F} = \mathcal{K}_{\mathcal{X}}$), Giné and Nickl [15,16,18] showed the same result with $\mathcal{F}$ being functions of bounded variation, $\{\mathbb{1}_{(-\infty,t]}\colon t \in \mathbb{R}\}$, Hölder, Lipschitz and Sobolev classes on $\mathbb{R}$. Similar result is shown for wavelet density estimators and spline projection estimators in $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]}\colon t \in \mathbb{R}\}$ (Giné and Nickl [17,19]) and maximum likelihood estimators in $\mathcal{F}_{\text{BL}}$ (Nickl [26]). While $\mathbb{P}_n$ is simple and elegant to use in practice, these other estimators that are $o_{\mathbb{P}}(n^{-1/2})$ from $\mathbb{P}_n$ have been shown to improve upon it in the following aspect: without any assumption on $\mathbb{P}$, it is possible to construct adaptive estimators that estimate $\mathbb{P}$ efficiently in $\mathcal{F}$ and at the same time estimate the density of $\mathbb{P}$ (if it exists without a priori assuming its existence) at the best possible convergence rate in some relevant loss over prescribed class of densities. Concretely, Giné and Nickl [18,19] proved the above behavior for kernel density estimator, wavelet density estimator and spline projection estimator on $\mathbb{R}$ for $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]}\colon t \in \mathbb{R}\}$ and sup-norm loss over the Hölder balls. By choosing $\mathcal{F} = \mathcal{F}_{\text{BL}}$ (with $d = 1$), Giné and Nickl [16] showed that the kernel density estimator adaptively estimates $\mathbb{P}$ in weak topology and at the same time estimates the density of $\mathbb{P}$ in strong topology at the best possible convergence rate over Sobolev balls.

The construction of these adaptive estimators involves applying Lepski's method (Lepski, Mammen and Spokoiny [23]) to kernel density estimators (in fact to any of the other estimators we discussed above) that are within a $\|\cdot\|_{\mathcal{F}}$-ball of size smaller than $n^{-1/2}$ around $\mathbb{P}_n$ and then using the exponential inequality of the type in Theorem 4.1 to control the probability of the event that $\sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}}$ is "too large" (see Giné and Nickl [16], Theorem 1, [18], Theorem 2, and [19], Theorem 3, for the optimality of the adaptive estimator in both $\|\cdot\|_{\mathcal{F}}$ and some relevant loss over prescribed class of densities). Using Theorem 4.1, it is quite straightforward in principle to construct an adaptive estimator that is optimal in both strong and weak topologies along the lines of Giné and Nickl ([16], Theorem 1), by incorporating two minor changes in the proof of Theorem 1 in [16]: the first change is to apply Theorem 4.1 in the place of Lemma 1 and extend Lemma 2 in Giné and Nickl [16] from $\mathbb{R}$ to $\mathbb{R}^d$. Informally, the procedure involves computing the bandwidth $\tilde{h}_n$ as

$$\begin{aligned}\tilde{h}_n = \max\Biggl\{ & h \in \mathcal{H}\colon \bigl\|\mathbb{P}_n * (K_h - K_g)\bigr\|_{L^1} \le \sqrt{\frac{A}{ng^d}}, \forall g < h, g \in \mathcal{H} \\ & \text{and } \|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}} \le \frac{n^{-1/2}}{\log n}\Biggr\},\end{aligned} \tag{5.3}$$

where $\mathcal{H} := \{h_k = \rho^{-k}\colon k \in \mathbb{N} \cup \{0\}, \rho^{-k} > (\log n)^2/n\}$ and $\rho > 1$ is arbitrary. Here, A depends on some moment conditions on $\mathbb{P} \in \mathcal{P}(\gamma, L)$, specifically through γ and L, where

$$\mathcal{P}(\gamma, L) = \left\{\mathbb{P} \in M_+^1(\mathbb{R}^d)\colon \int \bigl(1 + \|x\|_2^2\bigr)^{\gamma}\, \mathrm{d}\mathbb{P}(x) \le L\right\}$$

for some $L < \infty$ and $\gamma > \frac{d}{2}$. Along the lines of Theorem 1 in Giné and Nickl [16], the following result can be obtained (we state here without a proof) that shows the kernel density estimator with a purely data-driven bandwidth, $\tilde{h}_n$ to be optimal in both strong and weak topologies.

Theorem 5.3. *Let $(X_i)_{i=1}^n$ be random samples drawn i.i.d. from a probability measure $\mathbb{P} \in \mathcal{P}(\gamma, L)$ for some $L < \infty$ and $\gamma > \frac{d}{2}$. Suppose K is of order r satisfying $r > T + \frac{d}{2}$, $T \in \mathbb{N} \cup \{0\}$ such that $\int_{\mathbb{R}^d}(1 + \|x\|_2^2)^{\gamma} K^2(x)\,\mathrm{d}x < \infty$ where $p \in W_1^s(\mathbb{R}^d)$ for some $0 < s \le T$. If $\mathcal{F}_H$ is $\mathbb{P}$-Donsker (satisfied by $\mathcal{K}$ in Theorem 4.3), then*

$$\|\mathbb{P}_n * K_{\tilde{h}_n} - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\mathbb{P}}\bigl(n^{-1/2}\bigr) \quad \text{and} \quad \sqrt{n}(\mathbb{P}_n * K_{\tilde{h}_n} - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_{\mathbb{P}}.$$

Similarly, for $\mathcal{K}$ in Theorem 4.2, we have

$$\|\mathbb{P}_n * K_{\tilde{h}_n} - \mathbb{P}\|_{\mathcal{K}_{\mathcal{X}}} = \mathrm{O}_{\mathbb{P}}\bigl(n^{-1/2}\bigr) \quad \text{and} \quad \sqrt{n}(\mathbb{P}_n * K_{\tilde{h}_n} - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{K}_{\mathcal{X}})} \mathbb{G}_{\mathbb{P}}.$$

In addition, for any $0 < s \le T$,

$$\|\mathbb{P}_n * K_{\tilde{h}_n} - p\|_{L^1} = \mathrm{O}_{\mathbb{P}}\bigl(n^{-s/(2s+d)}\bigr).$$

We now discuss some computational aspects of the estimator in (5.4), which requires computing $\|\mathbb{P}_n * K_h - \mathbb{P}_n * K_g\|_{L^1}$ and $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}}$. While computing $\|\mathbb{P}_n * K_h - \mathbb{P}_n * K_g\|_{L^1}$ is usually not straightforward, the computation of $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}}$ can be simple depending on the choice of $\mathcal{F}$. In the following, we show that $\mathcal{F} = \mathcal{F}_H$ yields a simple maximization problem over a subset of $(0, \infty)$ depending on the choice of $\mathcal{K}$ and K, in contrast to an infinite dimensional optimization problem that would arise if $\mathcal{F} = \mathcal{F}_{\mathrm{BL}}$ and optimization over $\mathbb{R}^d \times (0, \infty)$ if $\mathcal{F} = \mathcal{K}_{\mathcal{X}}$ therefore demonstrating the computational advantage of working with $\mathcal{F}_H$ over $\mathcal{F}_{\mathrm{BL}}$ and $\mathcal{K}_{\mathcal{X}}$.

Consider $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H}$, which from (3.6) and (3.7) yields

$$\begin{aligned}
&\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} \\
&\quad = \sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \int K_h(X_i - x) k(\cdot, x)\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^n k(\cdot, X_i) \right\|_{\mathcal{H}_k} \\
&\quad = \frac{1}{n}\sup_{k\in\mathcal{K}} \sqrt{\sum_{i,j=1}^n \mathcal{A}(X_i, X_j) + k(X_i, X_j) - 2\int K_h(x) k(X_i - x, X_j)\,\mathrm{d}x},
\end{aligned}$$

which in turn reduces to

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = \sqrt{\frac{1}{n^2}\sup_{k\in\mathcal{K}} \sum_{i,j=1}^n (K_h * K_h * \psi + \psi - 2K_h * \psi)(X_i - X_j)} \tag{5.4}$$

when k is translation invariant, that is, $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, where

$$\mathcal{A}(X_i, X_j) := \int\int K_h(x) K_h(y) k(X_i - x, X_j - y)\,\mathrm{d}x\,\mathrm{d}y.$$

While computing (5.4) is not easy in general, in the following we present two examples where (5.4) is easily computable for appropriate choices of $\mathcal{K}$ and K. Let $\mathcal{K}$ be as in Theorem 4.2(a) (i.e., $\psi(x) := \psi_\sigma(x) = \mathrm{e}^{-\sigma\|x\|_2^2}$, $x \in \mathbb{R}^d$, $\sigma \in \Sigma$) and $K = \pi^{-d/2}\psi_1$. Then

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = \frac{1}{n}\sqrt{\sup_{\sigma\in\Sigma} \mathcal{A}_\psi(\sigma)}, \tag{5.5}$$

where

$$\mathcal{A}_\psi(\sigma) := \sum_{i,j=1}^n \left(\frac{\psi_{\sigma/(2\sigma h^2+1)}(X_i - X_j)}{(2\pi)^d (2\sigma h^2 + 1)^{d/2}} - \frac{2\psi_{\sigma/(\sigma h^2+1)}(X_i - X_j)}{(2\pi)^{d/2}(\sigma h^2 + 1)^{d/2}} + \psi_\sigma(X_i - X_j) \right).$$

Also choosing $\mathcal{K}$ to be as in Remark 4.3, that is,

$$\psi(x) := \phi_\alpha(x) = \prod_{i=1}^d \frac{\alpha^2}{\alpha^2 + x_i^2}, \qquad x \in \mathbb{R}^d, \alpha \in [c, \infty), c > 0,$$

which is a special case of $\mathcal{K}$ in Theorem 4.2(c) and $K = \pi^{-d}\phi_1$ in (5.4) yields

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H} = \frac{1}{n}\sqrt{\sup_{\alpha\in[c,\infty)} \mathcal{A}_\phi(\alpha)},$$

where

$$\mathcal{A}_\phi(\alpha) := \sum_{i,j=1}^{n}\left(\frac{\phi_{\alpha+2h}(X_i - X_j)}{\alpha^{-d}2^d(\alpha+2h)^d} - \frac{2\phi_{\alpha+h}(X_i - X_j)}{2^{d/2}\alpha^{-d}(\alpha+h)^d} + \phi_\alpha(X_i - X_j)\right).$$

In both these examples (where $\|\cdot\|_{\mathcal{F}_H}$ metrizes the weak topology on $M_+^1(\mathbb{R}^d)$), it is clear that one can compute $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_H}$ easily by solving a maximization problem over a subset of $(0,\infty)$, which can be carried out using standard gradient ascent methods. For the choice of K in both these examples, it is easy to see that K is of order 2 and therefore Theorem 4.2 holds if $s < 2 - \frac{d}{2}$.

On the other hand, note that

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{F}_{\mathrm{BL}}} = \frac{1}{n}\sup_{f\in\mathcal{F}_{\mathrm{BL}}}\left|\sum_{i=1}^{n}(K_h * f - f)(X_i)\right|$$

is not easily computable in practice. Also for $\mathcal{F} = \mathcal{K}_{\mathcal{X}}$, we have

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{K}_{\mathcal{X}}} = \sup_{k\in\mathcal{K},y\in\mathcal{X}}\left|\frac{1}{n}\sum_{i=1}^{n}\int K_h(X_i - x)k(y,x)\,\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}k(y,X_i)\right|,$$

which reduces to

$$\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{K}_{\mathcal{X}}} = \sup_{k\in\mathcal{K},y\in\mathcal{X}}\left|\frac{1}{n}\sum_{i=1}^{n}(K_h * \psi - \psi)(y - X_i)\right|$$

when $k(x,y) = \psi(x-y)$. For the choice of K and $\mathcal{K}$ as above (i.e., ψ_σ and ϕ_α), it is easy to verify that the computation of $\|\mathbb{P}_n * K_h - \mathbb{P}_n\|_{\mathcal{K}_{\mathcal{X}}}$ involves solving an optimization problem over $\mathbb{R}^d \times (0,\infty)$ which is more involved than solving the one in (5.5) that is obtained by working with $\mathcal{F}_H$.

In addition to the above application of adaptive estimation, there are various statistical applications where the choice of $\mathcal{F}_H$ can be computationally useful (over $\mathcal{F}_{\mathrm{BL}}$ and $\mathcal{K}_{\mathcal{X}}$), the examples of which include the two-sample and independence testing. As an example, in two-sample testing, $\|\mathbb{P}_n * K_h - \mathbb{Q}_m * K_g\|_{\mathcal{F}_H}$ can be used as a statistic to test for $\mathbb{P} = \mathbb{Q}$ vs. $\mathbb{P} \neq \mathbb{Q}$ based on n and m numbers of random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ respectively, assuming these distributions to have densities w.r.t. the Lebesgue measure. Based on the above discussion, it is easy to verify that $\|\mathbb{P}_n * K_h - \mathbb{Q}_m * K_g\|_{\mathcal{F}}$ is simpler to compute when $\mathcal{F} = \mathcal{F}_H$ compared to the other choices of $\mathcal{F}$ such as $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{F}_{\mathrm{BL}}$. Similarly, computationally efficient test statistics can be obtained for nonparametric independence tests through $\|\cdot\|_{\mathcal{F}_H}$.

6. Proofs

In this section, we present the missing proofs of results in Sections 3 and 4.

6.1. Proof of Proposition 3.1

For any $f \in \mathcal{H}_k$ and $\mathbb{P} \in \mathscr{P}_{\mathcal{K}}$, we have

$$\int f(x)\,\mathrm{d}\mathbb{P}(x) = \int \langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\,\mathrm{d}\mathbb{P}(x) = \left\langle f, \int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x)\right\rangle_{\mathcal{H}_k},$$

where the last equality follows from the assumption that k is Bochner-integrable, that is,

$$\int \|k(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mathbb{P}(x) = \int \sqrt{k(x,x)}\,\mathrm{d}\mathbb{P}(x) < \infty.$$

Therefore, for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_{\mathcal{K}}$,

$$\begin{aligned}\sup_{k\in\mathcal{K}}\sup_{\|f\|_{\mathcal{H}_k}\le 1}\left|\int f(x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right| &= \sup_{k\in\mathcal{K}}\sup_{\|f\|_{\mathcal{H}_k}\le 1}\left\langle f, \int k(\cdot,x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right\rangle_{\mathcal{H}_k}\\ &= \sup_{k\in\mathcal{K}}\left\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right\|_{\mathcal{H}_k},\end{aligned}$$

where the inner supremum is attained at $f = \frac{\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)}{\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\|_{\mathcal{H}_k}}$. Because of the Bochner-integrability of k,

$$\left\langle \int k(\cdot,x)\,\mathrm{d}\mathbb{P}(x), \int k(\cdot,y)\,\mathrm{d}\mathbb{Q}(y)\right\rangle_{\mathcal{H}_k} = \int\int k(x,y)\,\mathrm{d}\mathbb{P}(x)\,\mathrm{d}\mathbb{Q}(y)$$

and (3.7) follows.

6.2. Proof of Theorem 3.3

Since $\sup_{k\in\mathcal{K},x\in\mathcal{X}} k(x,x) \le \nu$ and $\mathbb{P}_n, \mathbb{P} \in \mathscr{P}_{\mathcal{K}}$, by Proposition 3.1, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \sup_{k\in\mathcal{K}}\left\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}_n-\mathbb{P})(x)\right\|_{\mathcal{H}_k}.$$

It is easy to check that $\sup_{k\in\mathcal{K}}\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}_n-\mathbb{P})(x)\|_{\mathcal{H}_k}$ satisfies the bounded difference property and therefore, by McDiarmid's inequality, for every $\tau > 0$, with probability at least $1-\mathrm{e}^{-\tau}$,

$$\begin{aligned}\sup_{k\in\mathcal{K}}\left\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}_n-\mathbb{P})(x)\right\|_{\mathcal{H}_k} &\le \mathbb{E}\sup_{k\in\mathcal{K}}\left\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}_n-\mathbb{P})(x)\right\|_{\mathcal{H}_k} + \sqrt{\frac{2\nu\tau}{n}}\\ &\overset{(*)}{\le} 2\mathbb{E}\mathbb{E}_\varepsilon\sup_{k\in\mathcal{K}}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\right\|_{\mathcal{H}_k} + \sqrt{\frac{2\nu\tau}{n}},\end{aligned} \tag{6.1}$$

where $(\varepsilon_i)_{i=1}^n$ represent i.i.d. Rademacher variables, $\mathbb{E}_\varepsilon$ represents the expectation w.r.t. $(\varepsilon_i)_{i=1}^n$ conditioned on $(X_i)_{i=1}^n$, and $(*)$ is obtained by symmetrizing $\mathbb{E}\sup_{k\in\mathcal{K}}\mathfrak{D}_k(\mathbb{P}_n,\mathbb{P})$ (see van der Vaart and Wellner [38], Lemma 2.3.1). Since $\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\|_{\mathcal{H}_k}$ satisfies the bounded difference property, another application of McDiarmid's inequality yields that, for every $\tau>0$, with probability at least $1-\mathrm{e}^{-\tau}$,

$$\begin{aligned}\mathbb{E}\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\right\|_{\mathcal{H}_k}\\ \le \mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\right\|_{\mathcal{H}_k} + \sqrt{\frac{2\nu\tau}{n}}\end{aligned} \tag{6.2}$$

and, therefore, combining (6.1) and (6.2) yields that for every $\tau>0$, with probability at least $1-2\mathrm{e}^{-\tau}$,

$$\begin{aligned}\sup_{k\in\mathcal{K}}\left\|\int k(\cdot,x)\,\mathrm{d}(\mathbb{P}_n-\mathbb{P})(x)\right\|_{\mathcal{H}_k}\\ \le 2\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\right\|_{\mathcal{H}_k} + \sqrt{\frac{18\nu\tau}{n}}.\end{aligned} \tag{6.3}$$

Note that

$$\begin{aligned}\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot,X_i)\right\|_{\mathcal{H}_k} &\le \frac{1}{n}\sqrt{\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j k(X_i,X_j)}\\ &\le \frac{\sqrt{2}}{n}\sqrt{U_n\big(\mathcal{K};(X_i)_{i=1}^n\big)}+\frac{\sqrt{\nu}}{\sqrt{n}},\end{aligned} \tag{6.4}$$

where

$$U_n\big(\mathcal{K};(X_i)_{i=1}^n\big) := \mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j k(X_i,X_j)\right|$$

is the expected suprema of the Rademacher chaos process of degree 2, indexed by $\mathcal{K}$. The proof until this point already appeared in Sriperumbudur *et al.* ([30], see the proof of Theorem 7), but we have presented here for completeness.

The result in (3.11) therefore follows by using (6.4) in (6.3) and bounding $U_n(\mathcal{K};(X_i)_{i=1}^n)$ through Lemma A.2 with $\theta=\frac{3}{4}$. Using (3.13) in (3.11) and solving for α yields (3.14) and (3.15).

Remark 6.1. (i) Note that instead of using McDiarmid's inequality in the above proof, one can directly obtain a version of (6.3) by applying Talagrand's inequality through Theorem 2.1 in Bartlett, Bousquet and Mendelson [3], albeit with worse constants and similar dependency on n.

(ii) If $\mathcal{K}$ is singleton, that is, $\mathcal{K}=\{k\}$, then l.h.s. of (6.4) can be bounded as

$$\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot, X_i)\right\|_{\mathcal{H}_k} \le \frac{1}{n}\sqrt{\mathbb{E}_\varepsilon \sum_{i,j=1}^n \varepsilon_i\varepsilon_j k(X_i, X_j)} \le \frac{1}{n}\sqrt{\mathbb{E}_\varepsilon \sum_{i\ne j} \varepsilon_i\varepsilon_j k(X_i, X_j)} + \frac{\sqrt{\nu}}{\sqrt{n}},$$

and, therefore,

$$\mathbb{E}_\varepsilon \sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i k(\cdot, X_i)\right\|_{\mathcal{H}_k} \le \frac{\sqrt{\nu}}{\sqrt{n}}. \tag{6.5}$$

6.3. Proof of Theorem 4.2

The proof involves showing that the kernels in (a)–(d) satisfy the conditions (i)–(iv) in Theorem 4.1, thereby ensuring that (4.1) and (4.2) hold. However, instead of bounding $\mathcal{T}$ through bounds on the covering numbers of $\mathcal{K}$, we directly bound the expected suprema of the Rademacher chaos process indexed by $\mathcal{K}$ and $\mathcal{K}_\alpha$, that is, $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_\alpha, (X_i)_{i=1}^n)$ which are defined in (3.16) – note that the terms involving $\mathcal{T}$ in (4.1) and (4.2) are in fact bounds on $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_\alpha, (X_i)_{i=1}^n)$ – and show that $U_n(\mathcal{K}, (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$ and $U_n(\mathcal{K}_\alpha, (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$. Using these results in (6.4) and (4.19) and following the proofs of Theorems 3.3 and 4.1, we have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = \mathrm{O}_{\text{a.s.}}(n^{-1/2})$,

$$\|K_h * \mathbb{P}_n - \mathbb{P}_n\|_{\mathcal{F}_H} \le \frac{E_1 h^r}{\sqrt{n}} + A_2 h^r$$

and

$$\|K_h * \mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \le \frac{F_1 h^r}{\sqrt{n}} + A_2 h^r + \frac{F_2}{\sqrt{n}},$$

where E_1 and $(F_i)_{i=1}^2$ are constants that do not depend on n (we do not provide the explicit constants here but can be easily worked out by following the proofs of Theorems 3.3 and 4.1). Therefore, the result follows.

In the following, we show that for $\mathcal{K}$ in (a)–(d), (iv) in Theorem 4.1 holds (note that (i)–(iii) in Theorem 4.1 hold trivially because of the choice of $\mathcal{K}$) along with $U_n(\mathcal{K}, (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$ and $U_n(\mathcal{K}_\alpha, (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$. In order to obtain bounds on $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_\alpha, (X_i)_{i=1}^n)$, we need an intermediate result (see Proposition 6.1 below) – also of independent interest – which is based on the notion of *pseudo-dimension* (Anthony and Bartlett [1], Definition 11.1) of a function class $\mathcal{F}$. It has to be noted that the pseudo-dimension of $\mathcal{F}$ matches with the VC-index of a VC-subgraph class, $\mathcal{F}$ (Anthony and Bartlett [1], Chapter 11, page 153).

Definition 2 (Pseudo-dimension). *Let $\mathcal{F}$ be a set of real valued functions on $\mathcal{X}$ and suppose that $S = \{z_1, \dots, z_n\} \subset \mathcal{X}$. Then S is pseudo-shattered by $\mathcal{F}$ if there are real numbers $r_1, \dots, r_n$ such that for any $b \in \{-1, 1\}^n$ there is a function $f_b \in \mathcal{F}$ with $\text{sign}(f_b(z_i) - r_i) = b_i$ for $i = 1, \dots, n$. The pseudo-dimension or VC-index of $\mathcal{F}$, $VC(\mathcal{F})$ is the maximum cardinality of S that is pseudo-shattered by $\mathcal{F}$.*

Proposition 6.1. *Let*

$$\mathcal{F} = \left\{ f_\sigma(x, y) = \sigma^\theta \prod_{i=1}^d \big(\sigma(x_i - y_i)^2\big)^{\delta_i} \mathrm{e}^{-\sigma(x_i - y_i)^2}, x, y \in \mathbb{R}^d \colon \sigma \in (0, \infty) \right\},$$

where $x := (x_1, \dots, x_d) \in \mathbb{R}^d$, $y := (y_1, \dots, y_d) \in \mathbb{R}^d$, $\theta \geq 0$ and $\delta_i > 0$ for any $i \in \{1, \dots, d\}$. Then $VC(\mathcal{F}) \leq 2$. If $\theta = \delta_1 = \cdots = \delta_d = 0$, then $VC(\mathcal{F}) = 1$.

Proof. Suppose $VC(\mathcal{F}) > 2$. Then there exists a set $S = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^d \colon i \in \{1, 2, 3\}\}$ which is pseudo-shattered by $\mathcal{F}$, where $x_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ and $y_i = (y_{i1}, \dots, y_{id}) \in \mathbb{R}^d$. This implies there exists $(r_1, r_2, r_3) \in \mathbb{R}^3$ such that for any $b \in \{-1, 1\}^3$ there is a function $f_\sigma \in \mathcal{F}$ with $\text{sign}(f_\sigma(x_i, y_i) - r_i) = b_i$ for $i = 1, 2, 3$. Without loss of generality, let us assume the points in S satisfy

$$\|x_2 - y_2\|_2 \leq \|x_1 - y_1\|_2 \leq \|x_3 - y_3\|_2. \tag{P$_{213}$}$$

We now consider two cases.

Case 1: $\|x_2 - y_2\|_2 < \|x_1 - y_1\|_2 < \|x_3 - y_3\|_2$: Let $b = (b_1, b_2, b_3) = (-1, 1, 1)$. Then there exists $\sigma_1 \in (0, \infty)$ such that the following hold:

$$f_{\sigma_1}(x_1, y_1) < r_1, \qquad f_{\sigma_1}(x_2, y_2) \geq r_2, \qquad f_{\sigma_1}(x_3, y_3) \geq r_3.$$

Similarly, for $b = (1, -1, -1)$, there exists $\sigma_2 \in (0, \infty)$ such that the following hold:

$$f_{\sigma_2}(x_1, y_1) \geq r_1, \qquad f_{\sigma_2}(x_2, y_2) < r_2, \qquad f_{\sigma_2}(x_3, y_3) < r_3.$$

This implies $f_{\sigma_2}(x_1, y_1) > f_{\sigma_1}(x_1, y_1)$, $f_{\sigma_1}(x_2, y_2) > f_{\sigma_2}(x_2, y_2)$, $f_{\sigma_1}(x_3, y_3) > f_{\sigma_2}(x_3, y_3)$, that is,

$$\sigma_2^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{1i} - y_{1i})^{2\delta_i} \mathrm{e}^{-\sigma_2 \|x_1 - y_1\|_2^2} > \sigma_1^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{1i} - y_{1i})^{2\delta_i} \mathrm{e}^{-\sigma_1 \|x_1 - y_1\|_2^2},$$

$$\sigma_2^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{2i} - y_{2i})^{2\delta_i} \mathrm{e}^{-\sigma_2 \|x_2 - y_2\|_2^2} < \sigma_1^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{2i} - y_{2i})^{2\delta_i} \mathrm{e}^{-\sigma_1 \|x_2 - y_2\|_2^2},$$

$$\sigma_2^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{3i} - y_{3i})^{2\delta_i} \mathrm{e}^{-\sigma_2 \|x_3 - y_3\|_2^2} < \sigma_1^{\theta + \sum_i \delta_i} \prod_{i=1}^d (x_{3i} - y_{3i})^{2\delta_i} \mathrm{e}^{-\sigma_1 \|x_3 - y_3\|_2^2}.$$

It is clear that $x_{ji} - y_{ji} \neq 0$ for any $i \in \{1, \dots, d\}$ and all $j \in \{1, 2, 3\}$ (otherwise leads to a contradiction). This implies

$$\mathrm{e}^{-(\sigma_1-\sigma_2)\|x_2-y_2\|_2^2} > \left(\frac{\sigma_2}{\sigma_1}\right)^{\theta+\sum_i \delta_i} > \mathrm{e}^{-(\sigma_1-\sigma_2)\|x_1-y_1\|_2^2},$$
$$\mathrm{e}^{-(\sigma_1-\sigma_2)\|x_3-y_3\|_2^2} > \left(\frac{\sigma_2}{\sigma_1}\right)^{\theta+\sum_i \delta_i} > \mathrm{e}^{-(\sigma_1-\sigma_2)\|x_1-y_1\|_2^2}$$

and, therefore,

$$(\sigma_1 - \sigma_2)\big(\|x_2 - y_2\|_2^2 - \|x_1 - y_1\|_2^2\big) < 0 \tag{6.6}$$

and

$$(\sigma_1 - \sigma_2)\big(\|x_3 - y_3\|_2^2 - \|x_1 - y_1\|_2^2\big) < 0,$$

which by our assumption $\|x_2 - y_2\|_2 < \|x_1 - y_1\|_2 < \|x_3 - y_3\|_2$ implies $\sigma_1 > \sigma_2$ and $\sigma_1 < \sigma_2$ leading to a contradiction. Therefore, no 3-point set S satisfying $\|x_2 - y_2\|_2 < \|x_1 - y_1\|_2 < \|x_3 - y_3\|_2$ is pseudo-shattered by $\mathcal{F}$.

Case 2: *At least one equality in* (P_{213}) *holds*: Suppose $\|x_2 - y_2\|_2 = \|x_1 - y_1\|_2 < \|x_3 - y_3\|_2$. Then (6.6) yields a contradiction. Similarly, a contradiction arises if $\|x_2 - y_2\|_2 < \|x_1 - y_1\|_2 = \|x_3 - y_3\|_2$ or $\|x_2 - y_2\|_2 = \|x_1 - y_1\|_2 = \|x_3 - y_3\|_2$.

Since every 3-point set S satisfies (P_{213}), from cases 1 and 2, it follows that no 3-point set S is pseudo-shattered by $\mathcal{F}$, which implies $VC(\mathcal{F}) \leq 2$.

If $\theta = \delta_i = 0$ for all $i \in \{1, \dots, d\}$, then $\mathcal{F} = \{f_\sigma(x, y) = \mathrm{e}^{-\sigma\|x-y\|_2^2}\colon \sigma \in (0, \infty)\}$. Using the same technique as above (also see the proof of Lemma 2 in Ying and Campbell [40]), it can be shown that no two-point is shattered by $\mathcal{F}$ and, therefore, $VC(\mathcal{F}) = 1$. □

Proof of Theorem 4.2. (a) Consider $\mathcal{K}_\alpha := \{\partial^{\alpha,\alpha}\psi_\sigma(\cdot - \cdot)\colon \sigma \in \Sigma\}$ for $|\alpha| = r$. It can be shown that

$$\partial^{\alpha,\alpha}\psi_\sigma(x - y) = \prod_{i=1}^{d}(-1)^{\alpha_i}\sigma^{\alpha_i} H_{2\alpha_i}\big(\sqrt{\sigma}(x_i - y_i)\big)\mathrm{e}^{-\sigma(x_i-y_i)^2},$$

where $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ and H_l denotes the Hermite polynomial of degree l. By expanding $H_{2\alpha_i}$ we obtain

$$\begin{aligned}\partial^{\alpha,\alpha}\psi_\sigma(x - y) &= \sigma^r \prod_{i=1}^{d}\sum_{j=0}^{\alpha_i}\eta_{ij}\big(\sigma(x_i - y_i)^2\big)^j \mathrm{e}^{-\sigma(x_i-y_i)^2} \\ &= \sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d}\prod_{i=1}^{d}\eta_{ij_i}\sigma^{\alpha_i+j_i}(x_i - y_i)^{2j_i}\mathrm{e}^{-\sigma(x_i-y_i)^2},\end{aligned} \tag{6.7}$$

where η_{ij_i} are finite constants and $\eta_{i0} > 0$ for all $i = 1, \ldots, d$. Therefore,

$$\begin{aligned}\sup_{\sigma\in\Sigma,x,y\in\mathbb{R}^d} \partial^{\alpha,\alpha}\psi_\sigma(x-y) &\le \sup_{\sigma\in\Sigma} \sigma^r\left(\sum_{j=0}^{\alpha}\prod_{i=1}^{d}|\eta_{ij_i}|j_i^{j_i}\mathrm{e}^{-j_i}\right)\\ &= a^r\left(\sum_{j=0}^{\alpha}\prod_{i=1}^{d}|\eta_{ij_i}|j_i^{j_i}\mathrm{e}^{-j_i}\right) < \infty,\end{aligned} \tag{6.8}$$

which implies (iv) in Theorem 4.1 is satisfied, where $\sum_{j=0}^{\alpha} := \sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d}$. Defining $\mathcal{B}_2 := U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n)$, we have

$$\begin{aligned}\mathcal{B}_2 &:= \mathbb{E}_\varepsilon \sup_{k'\in\mathcal{K}_\alpha}\left|\sum_{i<j}^{n}\varepsilon_i\varepsilon_j k'(X_i,X_j)\right|\\
&= \mathbb{E}_\varepsilon \sup_{\sigma\in\Sigma}\left|\sum_{p<q}^{n}\varepsilon_p\varepsilon_q \partial^{\alpha,\alpha}\psi_\sigma(X_p,X_q)\right|\\
&\overset{(6.7)}{=} \mathbb{E}_\varepsilon \sup_{\sigma\in\Sigma}\left|\sum_{p<q}^{n}\varepsilon_p\varepsilon_q \sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d}\prod_{i=1}^{d}\eta_{ij_i}\sigma^{\alpha_i+j_i}(X_{pi}-X_{qi})^{2j_i}\mathrm{e}^{-\sigma(X_{pi}-X_{qi})^2}\right|\\
&= \mathbb{E}_\varepsilon \sup_{\sigma\in\Sigma}\left|\sum_{j=0}^{\alpha}\left(\prod_{i=1}^{d}\eta_{ij_i}\right)\sum_{p<q}^{n}\varepsilon_p\varepsilon_q\prod_{i=1}^{d}\sigma^{\alpha_i+j_i}(X_{pi}-X_{qi})^{2j_i}\mathrm{e}^{-\sigma(X_{pi}-X_{qi})^2}\right|\\
&\le \sum_{j=0}^{\alpha}\left(\prod_{i=1}^{d}|\eta_{ij_i}|\right)\mathbb{E}_\varepsilon \sup_{\sigma\in\Sigma}\left|\sum_{p<q}^{n}\varepsilon_p\varepsilon_q\prod_{i=1}^{d}\sigma^{\alpha_i+j_i}(X_{pi}-X_{qi})^{2j_i}\mathrm{e}^{-\sigma(X_{pi}-X_{qi})^2}\right|\\
&= \sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d}\left(\prod_{i=1}^{d}|\eta_{ij_i}|\right)\mathbb{E}_\varepsilon \sup_{k_{j_1\cdots j_d}\in\mathcal{K}_\alpha^{j_1\cdots j_d}}\left|\sum_{p<q}^{n}\varepsilon_p\varepsilon_q k_{j_1\cdots j_d}(X_p,X_q)\right|,\end{aligned} \tag{6.9}$$

where

$$\mathcal{K}_\alpha^{j_1\cdots j_d} := \left\{k_{j_1\cdots j_d}(x,y) = \prod_{i=1}^{d}\sigma^{\alpha_i+j_i}(x_i-y_i)^{2j_i}\mathrm{e}^{-\sigma(x_i-y_i)^2}, x,y\in\mathbb{R}^d\colon \sigma\in\Sigma\right\}.$$

Since

$$\sup_{k_{j_1,\ldots,j_d}\in\mathcal{K}_\alpha^{j_1\cdots j_d},x,y\in\mathbb{R}^d} k_{j_1\cdots j_d}(x,y) \le a^r\mathrm{e}^{-\sum_{i=1}^{d}j_i}\prod_{i=1}^{d}j_i^{j_i} := \zeta_{j_1\cdots j_d} < \infty,$$

by Lemma A.2, we have

$$\begin{aligned}
&\mathbb{E}_\varepsilon \sup_{k_{j_1\cdots j_d}\in\mathcal{K}_\alpha^{j_1\cdots j_d}} \left|\sum_{p<q}^n \varepsilon_p\varepsilon_q k_{j_1\cdots j_d}(X_p, X_q)\right| \\
&\quad\le 2n^2\mathcal{T}\left(\mathcal{K}_\alpha^{j_1\cdots j_d}, \rho_{j_1\cdots j_d}, \frac{\zeta_{j_1\cdots j_d}}{2}\right) + \frac{n\zeta_{j_1\cdots j_d}}{\sqrt{2}},
\end{aligned} \tag{6.10}$$

where $\mathcal{T}$ and $\rho_{j_1\cdots j_d}$ (same as ρ_α but defined on $\mathcal{K}_\alpha^{j_1\cdots j_d}$) are defined in the statement of Theorem 4.1. Since every element of $\mathcal{K}_\alpha^{j_1\cdots j_d}$ is nonnegative and bounded above by $\zeta_{j_1\cdots j_d}$, we obtain the diameter of $\mathcal{K}_\alpha^{j_1\cdots j_d}$ to be bounded above by $\zeta_{j_1\cdots j_d}$ and, therefore, we used $\zeta_{j_1\cdots j_d}/2$ as an argument for $\mathcal{T}$ in (6.10). Proposition 6.1 shows that $\mathcal{K}_\alpha^{j_1\cdots j_d}$ is a VC-subgraph with VC-index, $V := VC(\mathcal{K}_\alpha^{j_1\cdots j_d}) \le 2$ for any $0 \le j_i \le \alpha_i$, $i = 1, \ldots, d$, which by Theorem 2.6.7 in van der Vaart and Wellner [38] implies that

$$\begin{aligned}
&\mathcal{N}\big(\mathcal{K}_\alpha^{j_1\cdots j_d}, \rho_{j_1\cdots j_d}, \epsilon\big) \\
&\quad\le C'V(16\mathrm{e})^V\left(\frac{\zeta_{j_1\cdots j_d}}{\epsilon}\right)^{2(V-1)}, \qquad 0<\epsilon<\zeta_{j_1\cdots j_d}
\end{aligned} \tag{6.11}$$

for some universal constant, C' and, therefore,

$$\mathcal{T}\left(\mathcal{K}_\alpha^{j_1\cdots j_d}, \rho_{j_1\cdots j_d}, \frac{\zeta_{j_1\cdots j_d}}{2}\right) \le \frac{C''_{j_1\cdots j_d}}{n}, \tag{6.12}$$

where $C''_{j_1\cdots j_d}$ is a constant that depends on C', V and $\zeta_{j_1\cdots j_d}$. Combining (6.10) and (6.12) in (6.9), we obtain

$$\begin{aligned}
&U_n\big(\mathcal{K}_\alpha; (X_i)_{i=1}^n\big) \\
&\quad\le n\sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d}\left(\prod_{i=1}^d |\eta_{ij_i}|\right)\left(2C''_{j_1\cdots j_d} + \frac{\zeta_{j_1\cdots j_d}}{\sqrt{2}}\right) \\
&\quad= \mathrm{O}_{\mathbb{P}}(n).
\end{aligned}$$

Also, since $\mathcal{K}$ is a VC-subgraph with $VC(\mathcal{K}) = 1$, from (6.11) we obtain $\mathcal{N}(\mathcal{K}, \rho, \epsilon)$ is a constant independent of ϵ. Following the analysis as above, it is easy to show that $U_n(\mathcal{K}; (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$.

(b) Since $\partial^{\alpha,\alpha}\int_0^\infty \psi_\sigma(x-y)\,\mathrm{d}\Lambda(\sigma) = \int_0^\infty \partial^{\alpha,\alpha}\psi_\sigma(x-y)\,\mathrm{d}\Lambda(\sigma)$ holds by Theorem 2.27(b) in Folland [12], define

$$\mathcal{K}_\alpha := \left\{\int_0^\infty \partial^{\alpha,\alpha}\psi_\sigma(x-y)\,\mathrm{d}\Lambda(\sigma), x, y\in\mathbb{R}^d\colon\ \Lambda\in\mathcal{M}_A\right\}.$$

Therefore,

$$\begin{aligned}\sup_{k'\in\mathcal{K}_\alpha,x,y\in\mathbb{R}^d} k'(x,y) &= \sup_{\Lambda\in\mathcal{M}_A,x,y\in\mathbb{R}^d}\int_0^\infty \partial^{\alpha,\alpha}\psi_\sigma(x-y)\,\mathrm{d}\Lambda(\sigma)\\ &\overset{(6.8)}{\le} \left(\sum_{j=0}^{\alpha}\prod_{i=1}^{d}|\eta_{ij_i}|j_i^{j_i}\mathrm{e}^{-j_i}\right)\sup_{\Lambda\in\mathcal{M}_A}\int_0^\infty \sigma^r\,\mathrm{d}\Lambda(\sigma)\\ &= A\left(\sum_{j=0}^{\alpha}\prod_{i=1}^{d}|\eta_{ij_i}|j_i^{j_i}\mathrm{e}^{-j_i}\right)<\infty,\end{aligned}$$

and so $\mathcal{K}$ satisfies (iv) in Theorem 4.1. Now consider

$$\begin{aligned}U_n\big(\mathcal{K};(X_i)_{i=1}^n\big) &= \mathbb{E}_\varepsilon \sup_{\Lambda\in\mathcal{M}_A}\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q\int_0^\infty \psi_\sigma(X_p-X_q)\,\mathrm{d}\Lambda(\sigma)\right|\\ &\le \mathbb{E}_\varepsilon \sup_{\sigma\in(0,\infty)}\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q\psi_\sigma(X_p-X_q)\right|.\end{aligned}\tag{6.13}$$

By Proposition 6.1, since $\{\psi_\sigma(x-y)\colon \sigma\in(0,\infty)\}$ is a VC-subgraph, carrying out the analysis (following (6.11)) in (a), we obtain $U_n(\mathcal{K};(X_i)_{i=1}^n)=\mathrm{O}_{\mathbb{P}}(n)$. Also,

$$\begin{aligned}U_n\big(\mathcal{K}_\alpha;(X_i)_{i=1}^n\big) &:= \mathbb{E}_\varepsilon \sup_{k'\in\mathcal{K}_\alpha}\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q k'(X_p,X_q)\right|\\ &= \mathbb{E}_\varepsilon \sup_{\Lambda\in\mathcal{M}_A}\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q\int_0^\infty \partial^{\alpha,\alpha}\psi_\sigma(X_p-X_q)\,\mathrm{d}\Lambda(\sigma)\right|\\ &\le \mathbb{E}_\varepsilon \sup_{\Lambda\in\mathcal{M}_A}\int_0^\infty\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q\partial^{\alpha,\alpha}\psi_\sigma(X_p-X_q)\right|\mathrm{d}\Lambda(\sigma)\\ &\le A\mathbb{E}_\varepsilon \sup_{\sigma\in(0,\infty)}\left|\sum_{p<q}^n \varepsilon_p\varepsilon_q\sigma^{-r}\partial^{\alpha,\alpha}\psi_\sigma(X_p-X_q)\right| =: AU_n\big(\mathcal{L};(X_i)_{i=1}^n\big),\end{aligned}$$

where

$$\mathcal{L}:=\left\{\sigma^{-r}\partial^{\alpha,\alpha}\psi_\sigma(x-y),x,y\in\mathbb{R}^d\colon \sigma\in(0,\infty)\right\}.$$

Replicating the analysis in (6.9) for $U_n(\mathcal{L};(X_i)_{i=1}^n)$ in conjunction with Proposition 6.1, it is easy to show that $U_n(\mathcal{L};(X_i)_{i=1}^n)=\mathrm{O}_{\mathbb{P}}(n)$ and, therefore, $U_n(\mathcal{K}_\alpha;(X_i)_{i=1}^n)=\mathrm{O}_{\mathbb{P}}(n)$.

(c) It is easy to check that any $k \in \mathcal{K}$ is of the form

$$k(x, y) = \prod_{i=1}^{d} \int_0^\infty e^{-\sigma_i (x_i - y_i)^2} \, d\Lambda_i(\sigma_i).$$

Therefore,

$$\mathcal{K}_\alpha = \left\{ \prod_{i=1}^{d} \int_0^\infty \partial^{\alpha_i, \alpha_i} \psi_{\sigma_i}(x_i - y_i) \, d\Lambda_i(\sigma_i), x, y \in \mathbb{R}^d \colon \Lambda_i \in \mathcal{M}_{A_i}, i = 1, \ldots, d \right\}$$

and

$$\begin{aligned} \sup_{k' \in \mathcal{K}_\alpha, x, y \in \mathbb{R}^d} k'(x, y) &= \prod_{i=1}^{d} \sup_{\Lambda_i \in \mathcal{M}_{A_i}, x_i, y_i \in \mathbb{R}} \int_0^\infty \partial^{\alpha_i, \alpha_i} \psi_{\sigma_i}(x_i - y_i) \, d\Lambda_i(\sigma_i) \\ &= \prod_{i=1}^{d} A_i \sum_{j=0}^{\alpha_i} |\eta_{ij}| j^j e^{-j} < \infty, \end{aligned}$$

which implies $\mathcal{K}$ satisfies (iv) in Theorem 4.1. Now consider

$$\begin{aligned} U_n\big(\mathcal{K}; (X_i)_{i=1}^n\big) &= \mathbb{E}_\varepsilon \sup_{\Lambda \in \mathcal{Q}_A} \left| \sum_{p<q}^{n} \varepsilon_p \varepsilon_q \int e^{-(X_p - X_q)^T \Delta (X_p - X_q)} \, d\Lambda(\Delta) \right| \\ &\le \mathbb{E}_\varepsilon \sup_{\mathrm{diag}(\Delta) \in (0,\infty)^d} \left| \sum_{p<q}^{n} \varepsilon_p \varepsilon_q e^{-(X_p - X_q)^T \Delta (X_p - X_q)} \right| \\ &=: U_n\big(\mathcal{J}; (X_i)_{i=1}^n\big), \end{aligned}$$

where

$$\mathcal{J} := \left\{ \tilde{k}(x, y) = e^{-(x-y)^T \Delta (x-y)} = \prod_{i=1}^{d} e^{-\sigma_i (x_i - y_i)^2}, x, y \in \mathbb{R}^d : \mathrm{diag}(\Delta) \in (0, \infty)^d \right\}.$$

Define

$$\mathcal{J}_i := \big\{ \tilde{k}^i(x, y) = e^{-\sigma_i (x_i - y_i)^2}, x_i, y_i \in \mathbb{R} \colon \sigma_i \in (0, \infty) \big\}.$$

It is easy to check that for any $\tilde{k}_1, \tilde{k}_2 \in \mathcal{J}$, $\rho(\tilde{k}_1, \tilde{k}_2) \le \sqrt{d} \sum_{i=1}^{d} \rho(\tilde{k}_1^i, \tilde{k}_2^i)$, where $\tilde{k}_1^i, \tilde{k}_2^i \in \mathcal{J}_i$ and $\mathcal{N}(\mathcal{J}, \rho, \epsilon) = \prod_{i=1}^{d} \mathcal{N}(\mathcal{J}_i, \rho, d^{-3/2}\epsilon)$. By Proposition 6.1, since $\mathcal{J}_i$ is a VC-subgraph for any $i = 1, \ldots, d$, from the analysis in (a), we obtain $\mathcal{N}(\mathcal{J}_i, \rho, \epsilon) = O(1)$ and, therefore,

$$U_n\big(\mathcal{K}; (X_i)_{i=1}^n\big) \le U_n\big(\mathcal{J}; (X_i)_{i=1}^n\big) = O_{\mathbb{P}}(n).$$

Similarly,

$$\begin{aligned}
&U_n\big(\mathcal{K}_\alpha;(X_i)_{i=1}^n\big)\\
&\quad:=\mathbb{E}_\varepsilon\sup_{k'\in\mathcal{K}_\alpha}\left|\sum_{p<q}^n\varepsilon_p\varepsilon_q k'(X_p,X_q)\right|\\
&\quad=\mathbb{E}_\varepsilon\sup_{\Lambda_i\in\mathcal{M}_{A_i},i\in[d]}\left|\sum_{p<q}^n\varepsilon_p\varepsilon_q\prod_{i=1}^d\int_0^\infty\partial^{\alpha_i,\alpha_i}\psi_{\sigma_i}(X_{pi}-X_{qi})\,\mathrm{d}\Lambda_i(\sigma_i)\right|\\
&\quad=\mathbb{E}_\varepsilon\sup_{\Lambda_i\in\mathcal{M}_{A_i},i\in[d]}\left|\int_0^\infty\cdots\int_0^\infty\sum_{p<q}^n\varepsilon_p\varepsilon_q\prod_{i=1}^d\partial^{\alpha_i,\alpha_i}\psi_{\sigma_i}(X_{pi}-X_{qi})\prod_{i=1}^d\mathrm{d}\Lambda_i(\sigma_i)\right|\\
&\quad\le\left(\prod_{i=1}^d A_i\right)\mathbb{E}_\varepsilon\sup_{\operatorname{diag}(\Delta)\in(0,\infty)^d}\left|\sum_{p<q}^n\varepsilon_p\varepsilon_q\prod_{i=1}^d\sigma_i^{-\alpha_i}\partial^{\alpha_i,\alpha_i}\psi_{\sigma_i}(X_{pi}-X_{qi})\right|\\
&\quad=:\left(\prod_{i=1}^d A_i\right)U_n\big(\mathcal{I};(X_i)_{i=1}^n\big),
\end{aligned}$$

where $[d]:=\{1,\dots,d\}$ and

$$\mathcal{I}:=\left\{\check{k}(x,y)=\prod_{i=1}^d\sigma_i^{-\alpha_i}\partial^{\alpha_i,\alpha_i}\psi_{\sigma_i}(x_i-y_i),\right.$$
$$\left.x,y\in\mathbb{R}^d:\ (\sigma_1,\dots,\sigma_d)\in(0,\infty)^d\right\}.$$

We now proceed as above to obtain a bound on $U_n(\mathcal{I};(X_i)_{i=1}^n)$ through $\mathcal{N}(\mathcal{I},\rho,\epsilon)$ by defining

$$\mathcal{I}_i:=\big\{\check{k}^i(x,y)=\sigma_i^{-\alpha_i}\partial^{\alpha_i,\alpha_i}\psi_{\sigma_i}(x_i-y_i),x_i,y_i\in\mathbb{R}:\ \sigma_i\in(0,\infty)\big\}$$

and noting that for any $\check{k}_1,\check{k}_2\in\mathcal{I}$, we have $\rho(\check{k}_1,\check{k}_2)\le Bd^{3/2}\rho(\check{k}_1^i,\check{k}_2^i)$ where $\check{k}_1^i,\check{k}_2^i\in\mathcal{I}_i$, $B:=\max_{i\in\{1,\dots,d\}}\sum_{j=0}^{\alpha_i}|\eta_{ij}|j^j\mathrm{e}^{-j}$ and $\mathcal{N}(\mathcal{I},\rho,\epsilon)=\prod_{i=1}^d\mathcal{N}(\mathcal{I}_i,\rho,B^{-1}d^{-3/2}\epsilon)$. Proceeding with the covering number analysis in (a), it can be shown that $\mathcal{I}_i$ is a VC-subgraph with $VC(\mathcal{I}_i)\le 2$ for any $i=1,\dots,d$ and, therefore, $\mathcal{N}(\mathcal{I},\rho,\epsilon)=\mathrm{O}(\epsilon^{-2})$, which means

$$\begin{aligned}
U_n\big(\mathcal{K}_\alpha;(X_i)_{i=1}^n\big)&\le\left(\prod_{i=1}^d A_i\right)U_n\big(\mathcal{I};(X_i)_{i=1}^n\big)\\
&=\mathrm{O}_{\mathbb{P}}(n).
\end{aligned}$$

(d) First we derive an alternate form for $k\in\mathcal{K}$ which will be useful to prove the result. To this end, by Theorem 6.13 in Wendland [39], any $k\in\mathcal{K}$ can be written as the Fourier transform of

$\frac{Ac^{2\beta-d}\Gamma(\beta)}{2^{1-\beta}}(c^2+\|\omega\|_2^2)^{-\beta}$, that is, for any $c>0$,

$$\begin{aligned} k(x,y) &= A\frac{\|x-y\|_2^{\beta-d/2}}{c^{d/2-\beta}}\mathfrak{K}_{d/2-\beta}\big(c\|x-y\|_2\big) \\ &= \frac{Ac^{2\beta-d}\Gamma(\beta)}{(2\pi)^{d/2}2^{1-\beta}}\int_{\mathbb{R}^d} \mathrm{e}^{-\sqrt{-1}(x-y)^T\omega}\big(c^2+\|\omega\|_2^2\big)^{-\beta}\,\mathrm{d}\omega. \end{aligned} \tag{6.14}$$

By the Schönberg representation for radial kernels (see (3.2)), it follows from Wendland [39], Theorem 7.15, that

$$\big(c^2+\|\omega\|_2^2\big)^{-\beta} = \frac{1}{\Gamma(\beta)}\int_0^\infty \mathrm{e}^{-t\|\omega\|_2^2}t^{\beta-1}\mathrm{e}^{-c^2t}\,\mathrm{d}t. \tag{6.15}$$

Combining (6.14) and (6.15), we have

$$k(x,y) = \frac{Ac^{2\beta-d}}{(2\pi)^{d/2}2^{1-\beta}}\int_{\mathbb{R}^d} \mathrm{e}^{-\sqrt{-1}(x-y)^T\omega}\int_0^\infty \mathrm{e}^{-t\|\omega\|_2^2}t^{\beta-1}\mathrm{e}^{-c^2t}\,\mathrm{d}t\,\mathrm{d}\omega,$$

which after applying Fubini's theorem yields

$$\begin{aligned} k(x,y) &= \frac{Ac^{2\beta-d}}{(2\pi)^{d/2}2^{1-\beta}}\int_0^\infty\int_{\mathbb{R}^d} \mathrm{e}^{-\sqrt{-1}(x-y)^T\omega}\mathrm{e}^{-t\|\omega\|_2^2}\,\mathrm{d}\omega\,t^{\beta-1}\mathrm{e}^{-c^2t}\,\mathrm{d}t \\ &= \frac{c^{2\beta-d}}{\Gamma(\beta-d/2)}\int_0^\infty \mathrm{e}^{-(\|x-y\|_2^2)/(4t)}t^{\beta-1-d/2}\mathrm{e}^{-c^2t}\,\mathrm{d}t. \end{aligned} \tag{6.16}$$

Note that

$$\sup_{k\in\mathcal{K},x,y\in\mathcal{X}} k(x,y) \le \sup_{c\in(0,a]}\frac{c^{2\beta-d}}{\Gamma(\beta-d/2)}\int_0^\infty t^{\beta-1-d/2}\mathrm{e}^{-c^2t}\,\mathrm{d}t = 1,$$

implying that $\mathcal{K}$ satisfies (iii) in Theorem 4.1. Using (6.16) in $U_n(\mathcal{K};(X_i)_{i=1}^n)$, we have

$$\begin{aligned} &U_n\big(\mathcal{K};(X_i)_{i=1}^n\big) \\ &\quad= \mathbb{E}_\varepsilon\sup_{k\in\mathcal{K}}\left|\sum_{i<j}^n\varepsilon_i\varepsilon_j k(X_i,X_j)\right| \\ &\quad= \mathbb{E}_\varepsilon\sup_{c\in(0,a]}\frac{c^{2\beta-d}}{\Gamma(\beta-d/2)}\left|\sum_{i<j}^n\varepsilon_i\varepsilon_j\int_0^\infty \mathrm{e}^{-(\|X_i-X_j\|_2^2)/(4t)}t^{\beta-1-d/2}\mathrm{e}^{-c^2t}\,\mathrm{d}t\right| \\ &\quad\le \mathbb{E}_\varepsilon\sup_{t\in(0,\infty)}\left|\sum_{i<j}^n\varepsilon_i\varepsilon_j\mathrm{e}^{-(\|X_i-X_j\|_2^2)/(4t)}\right|\sup_{c\in(0,a]}\frac{c^{2\beta-d}}{\Gamma(\beta-d/2)}\left|\int_0^\infty t^{\beta-1-d/2}\mathrm{e}^{-c^2t}\,\mathrm{d}t\right| \\ &\quad= \mathbb{E}_\varepsilon\sup_{\sigma\in(0,\infty)}\left|\sum_{i<j}^n\varepsilon_i\varepsilon_j\mathrm{e}^{-\sigma\|X_i-X_j\|_2^2}\right|, \end{aligned}$$

and, therefore, it follows (see Remark 3.1(i)) that $U_n(\mathcal{K}; (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$. Now for $|\alpha| = m \wedge r$, let us consider

$$k'(x, y) := \partial^{\alpha,\alpha} k(x, y) = \frac{c^{2\beta-d}}{\Gamma(\beta - d/2)} \int_0^\infty \left(\partial^{\alpha,\alpha} \mathrm{e}^{-(\|x-y\|_2^2)/(4t)}\right) t^{\beta-1-d/2} \mathrm{e}^{-c^2 t}\, \mathrm{d}t$$
$$= \frac{c^{2\beta-d}}{\Gamma(\beta - d/2)} \int_0^\infty \left((4t)^{m\wedge r} \partial^{\alpha,\alpha} \mathrm{e}^{-(\|x-y\|_2^2)/(4t)}\right) t^{\beta-1-d/2} (4t)^{-(m\wedge r)} \mathrm{e}^{-c^2 t}\, \mathrm{d}t$$
$$= \frac{c^{2\beta-d}}{\Gamma(\beta - d/2)4^{m\wedge r}} \int_0^\infty \left((4t)^{m\wedge r} \partial^{\alpha,\alpha} \mathrm{e}^{-(\|x-y\|_2^2)/(4t)}\right) t^{\beta-1-d/2-(m\wedge r)} \mathrm{e}^{-c^2 t}\, \mathrm{d}t, \quad (6.17)$$

where the equality in the first line follows from Folland [12], Theorem 2.27(b). The above implies

$$\sup_{k' \in \mathcal{K}_\alpha, x, y \in \mathcal{X}} k'(x, y) \le \sup_{\sigma \in (0,\infty), x, y \in \mathcal{X}} \left|\sigma^{-(m\wedge r)} \partial^{\alpha,\alpha} \mathrm{e}^{-\sigma\|x-y\|_2^2}\right| \frac{\Gamma(\beta - d/2 - m \wedge r) a^{2(m\wedge r)}}{\Gamma(\beta - d/2)4^{m\wedge r}}$$
$$< \infty,$$

therefore satisfying (iv) in Theorem 4.1. Using (6.17) we now obtain a bound on $\mathcal{B}_3 := U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n)$ as follows by defining $B := \Gamma(\beta - \frac{d}{2})4^{m\wedge r}$.

$$\mathcal{B}_3 = \mathbb{E}_\varepsilon \sup_{k' \in \mathcal{K}_\alpha} \left|\sum_{i<j}^n \varepsilon_i \varepsilon_j k'(X_i, X_j)\right|$$
$$= \mathbb{E}_\varepsilon \sup_{c \in (0,a]} \frac{c^{2\beta-d}}{B} \left|\sum_{i<j}^n \varepsilon_i \varepsilon_j \int_0^\infty \left((4t)^{m\wedge r} \partial^{\alpha,\alpha} \mathrm{e}^{-(\|X_i-X_j\|_2^2)/(4t)}\right) t^{\beta-1-d/2-(m\wedge r)} \mathrm{e}^{-c^2 t}\, \mathrm{d}t\right|$$
$$\le \mathbb{E}_\varepsilon \sup_{t \in (0,\infty)} \left|\sum_{i<j}^n \varepsilon_i \varepsilon_j (4t)^{m\wedge r} \partial^{\alpha,\alpha} \mathrm{e}^{-(\|X_i-X_j\|_2^2)/(4t)}\right|$$
$$\times \sup_{c \in (0,a]} \frac{c^{2\beta-d}}{B} \int_0^\infty t^{\beta-1-d/2-(m\wedge r)} \mathrm{e}^{-c^2 t}\, \mathrm{d}t$$
$$\le \frac{\Gamma(\beta - d/2 - m \wedge r) a^{2(m\wedge r)}}{\Gamma(\beta - d/2)4^{m\wedge r}} \mathbb{E}_\varepsilon \sup_{\sigma \in (0,\infty)} \left|\sum_{i<j}^n \varepsilon_i \varepsilon_j \sigma^{-(m\wedge r)} \partial^{\alpha,\alpha} \mathrm{e}^{-\sigma\|X_i-X_j\|_2^2}\right|,$$

and so $U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n) = \mathrm{O}_{\mathbb{P}}(n)$, which follows from the proof of Theorem 4.2(ii). □

Remark 6.2. Note that instead of following the indirect route – showing $\mathcal{K}_\alpha^{j_1\cdots j_d}$ to be a VC-subgraph and then bounding $U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n)$ – of showing the result in Theorem 4.1 for the Gaussian kernel family as presented in (a), one can directly get the result by obtaining a bound on $\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon)$ as presented in Proposition B.1, under the assumption that $\mathcal{X} = (a_0, b_0)^d$ for some $-\infty < a_0 < b_0 < \infty$. The advantage with the analysis in (a) is that the result holds for $\mathcal{X} = \mathbb{R}^d$ rather than a bounded subset of $\mathbb{R}^d$. Also the proof technique in (a) is useful and interesting as it avoids the difficult problem of bounding the covering numbers of $\mathcal{K}$ and $\mathcal{K}_\alpha$ for kernel classes in (b)–(d) while allowing to handle these classes easily through (a).

6.4. Proof of the claim in Remark 4.3(iii)

We show that $\mathcal{K}$ in (a)–(c) satisfy the conditions in Theorem 3.2 and, therefore, metrize the weak topology on $M_+^1(\mathbb{R}^d)$. Note that the families in (a)–(c) are uniformly bounded and every $k \in \mathcal{K}$ is such that $k(\cdot, x) \in C_0(\mathbb{R}^d)$ for all $x \in \mathbb{R}^d$. It therefore remains to check (3.10) and (P) in Theorem 3.2. By Proposition 5 in Sriperumbudur, Fukumizu and Lanckriet [33] (see (17) in its proof), it is clear that (3.10) is satisfied for $\mathcal{K}$ in (a) and (b). For (c),

$$
\begin{aligned}
B &:= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\prod_{j=1}^d\int_0^\infty \mathrm{e}^{-\sigma(x_j-y_j)^2}\,\mathrm{d}\Lambda_j(\sigma)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\prod_{j=1}^d\int_0^\infty \frac{1}{(4\pi\sigma)^{d/2}}\int_{\mathbb{R}} \mathrm{e}^{-\sqrt{-1}\omega_j(x_j-y_j)}\mathrm{e}^{-\omega_j^2/(4\sigma)}\,\mathrm{d}\omega_j\,\mathrm{d}\Lambda_j(\sigma)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{e}^{-\sqrt{-1}\omega^T(x-y)}\prod_{j=1}^d\int_0^\infty \frac{1}{(4\pi\sigma)^{d/2}}\mathrm{e}^{-\omega_j^2/(4\sigma)}\,\mathrm{d}\Lambda_j(\sigma)\,\mathrm{d}\omega\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\big|\widehat{\mu}(\omega)\big|^2\left(\prod_{j=1}^d\int_0^\infty \frac{1}{(4\pi\sigma)^{d/2}}\mathrm{e}^{-\omega_j^2/(4\sigma)}\,\mathrm{d}\Lambda_j(\sigma)\right)\mathrm{d}\omega, \qquad (6.18)
\end{aligned}
$$

where we have invoked Fubini's theorem in the last two lines of (6.18) and $\widehat{\mu}$ denotes the Fourier transform of μ. Since $\operatorname{supp}(\Lambda_j) \neq \{0\}$ for all $j = 1, \dots, d$, the inner integrals in (6.18) are positive for every $\omega_j \in \mathbb{R}$ and so (3.10) holds.

We now show that (P) in Theorem 3.2 is satisfied by $\mathcal{K}$ in (a)–(c). Consider $\mathcal{K}$ in (b). Fix $x \in \mathbb{R}^d$ and $\epsilon > 0$. Define $U_{x,\epsilon} = \{y \in \mathbb{R}^d\colon \|x-y\|_2 < (4\delta\log\frac{2B}{2-\epsilon^2})^{1/4}\}$, where δ and B are as mentioned in the statement of Theorem 4.2. Then for any $k \in \mathcal{K}$ and $y \in U_{x,\epsilon}$,

$$
\begin{aligned}
\big\|k(\cdot,x)-k(\cdot,y)\big\|_{\mathcal{H}_k}^2 &= 2-2\int_0^\infty \mathrm{e}^{-\sigma\|x-y\|_2^2}\,\mathrm{d}\Lambda(\sigma)\\
&\le 2-2\left(\inf_{\Lambda\in\mathcal{M}_A}\int_0^\infty \mathrm{e}^{-\delta\sigma^2}\,\mathrm{d}\Lambda(\sigma)\right)\left(\inf_{\sigma\in(0,\infty)} \mathrm{e}^{-\sigma\|x-y\|_2^2}\mathrm{e}^{\delta\sigma^2}\right)\\
&\le 2-2B\mathrm{e}^{-\|x-y\|_2^4/(4\delta)} < \epsilon^2.
\end{aligned}
$$

For $\mathcal{K}$ in (c), define $U_{x,\epsilon} := \{y \in \mathbb{R}^d\colon \|x-y\|_\infty < (4\min_i \delta_i \log \frac{2\prod_{i=1}^d B_i}{2-\epsilon^2})^{1/4}\}$ for some fixed $x \in \mathbb{R}^d$ and $\epsilon > 0$. Then as above, it is easy to show that for any $k \in \mathcal{K}$ and $y \in U_{x,\epsilon}$,

$$\begin{aligned}\|k(\cdot,x)-k(\cdot,y)\|^2_{\mathcal{H}_k} &= 2-2\prod_{i=1}^d \int_0^\infty \mathrm{e}^{-\sigma(x_i-y_i)^2}\,\mathrm{d}\Lambda_i(\sigma)\\ &\le 2-2\prod_{i=1}^d \Big(\inf_{\Lambda_i\in\mathcal{M}_{A_i}} \int_0^\infty \mathrm{e}^{-\delta_i\sigma^2}\,\mathrm{d}\Lambda_i(\sigma)\Big)\Big(\inf_{\sigma\in(0,\infty)} \mathrm{e}^{-\sigma(x_i-y_i)^2}\mathrm{e}^{\delta_i\sigma^2}\Big)\\ &\le 2-2\prod_{i=1}^d B_i \mathrm{e}^{-(x_i-y_i)^4/(4\delta_i)} \le 2-2\prod_{i=1}^d B_i \mathrm{e}^{-\|x-y\|_\infty^4/(4\min_i\delta_i)} < \epsilon^2,\end{aligned}$$

thereby proving the result.

6.5. Proof of Theorem 4.3

In the following, we prove that the class $\mathcal{F}_H$ induced by the family $\mathcal{K}$ in (a)–(d) are Donsker and, therefore, the result simply follows from Theorem 4.2. To this end, we first prove that $\mathcal{K}$ in (d) is Donsker which will be helpful to prove the claim for the kernel classes in (a)–(c).

(d) Since k is continuous and bounded and $\mathcal{X}$ is separable, by Steinwart and Christmann [35], Lemma 4.33, the RKHS $\mathcal{H}_k$ induced by k is separable and every $f \in \mathcal{H}_k$ is also continuous and bounded. In addition, the inclusion $\mathrm{id}\colon \mathcal{H}_k \to C_b(\mathcal{X})$ is linear and continuous (Steinwart and Christmann [35], Lemma 4.28). Therefore, by Marcus ([24], Theorem 1.1), $\mathcal{F}_H = \{f \in \mathcal{H}_k\colon \|f\|_{\mathcal{H}_k} \le 1\}$ is $\mathbb{P}$-Donsker, that is, $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_{\mathbb{P}}$. Also, $\sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}) = \sqrt{n}(\mathbb{P}_n * K_h - \mathbb{P}_n) + \sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow_{\ell^\infty(\mathcal{F}_H)} \mathbb{G}_{\mathbb{P}}$ by Slutsky's lemma and Theorem 4.1.

(a)–(c) From (4.28), we have

$$\mathcal{F}_H \subset \bigcup_{\sigma\in[a,b]} \Big\{f \in \mathcal{H}_b\colon \|f\|_{\mathcal{H}_b} \le \Big(\frac{b}{\sigma}\Big)^{d/4}\Big\} = \Big\{f \in \mathcal{H}_b\colon \|f\|_{\mathcal{H}_b} \le \Big(\frac{b}{a}\Big)^{d/4}\Big\} =: \mathcal{B}.$$

Using the argument as in (d), it is easy to verify that $\mathcal{H}_b$ is separable and $\mathrm{id}\colon \mathcal{H}_b \mapsto C_b(\mathcal{X})$ is linear and continuous and, therefore, $\mathcal{B}$ is $\mathbb{P}$-Donsker, which implies $\mathcal{F}_H$ is Donsker by van der Vaart and Wellner [38], Theorem 2.10.1. The result therefore follows using Slutsky's lemma and Theorem 4.2. The proof of (c) is similar to that of in (a) but we use (4.30) instead of (4.28). For (b), the result hinges on a relation similar to those in (4.28) and (4.30), which we derive below. Let $\mathcal{K}$ be the kernel family as shown in (4.29). Then for $k \in \mathcal{K}$, let $\mathcal{H}_c$ be the induced RKHS. From Wendland [39], Theorems 6.13 and 10.12, it follows that for any $f \in \mathcal{H}_c$,

$$\|f\|^2_{\mathcal{H}_c} = \frac{\Gamma(\beta)}{2^{1-\beta}} \int |\widehat{f}(\omega)|^2 \frac{c^{-d}}{(c\|\omega\|_2)^{\beta-d/2}\mathfrak{K}_{d/2-\beta}(c\|\omega\|_2)}\,\mathrm{d}\omega.$$

By Wendland [39], Corollary 5.12, since for every $\nu \in \mathbb{R}$, $x \mapsto x^\nu \mathfrak{K}_{-\nu}(x)$ is nonincreasing on $(0,\infty)$, we have that for any $0 < \tau < c < \infty$,

$$\|f\|_{\mathcal{H}_\tau} \leq \left(\frac{c}{\tau}\right)^{d/2} \|f\|_{\mathcal{H}_c}$$

and so $\mathcal{H}_c \subset \mathcal{H}_\tau$. Therefore, we have

$$\mathcal{F}_H \subset \bigcup_{c\in[a,\infty)} \left\{ f \in \mathcal{H}_a \colon \|f\|_{\mathcal{H}_a} \leq \left(\frac{c}{a}\right)^{d/2} \right\} = \mathcal{H}_a.$$

For the choice of $\mathcal{K}$ in Theorem 4.3(b), we have

$$\mathcal{F}_H \subset \bigcup_{c\in[a,b]} \left\{ f \in \mathcal{H}_a \colon \|f\|_{\mathcal{H}_a} \leq \left(\frac{c}{a}\right)^{d/2} \right\} = \left\{ f \in \mathcal{H}_a \colon \|f\|_{\mathcal{H}_a} \leq \left(\frac{b}{a}\right)^{d/2} \right\} \tag{6.19}$$

and the rest follows.

6.6. Proof of Proposition 5.1

By definition,

$$\begin{aligned}
\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_\mathcal{X}} &= \sup_{k\in\mathcal{K},x\in\mathcal{X}} \left| \int k(x,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \right| \\
&= \sup_{k\in\mathcal{K},x\in\mathcal{X}} \left| \int \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}_k}\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \right|.
\end{aligned}$$

Since $\mathcal{K}$ is uniformly bounded, $k(\cdot,x)$ is Bochner-integrable for all $k\in\mathcal{K}$ and $x\in\mathcal{X}$, that is,

$$\int \|k(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mathbb{P}(x) = \int \sqrt{k(x,x)}\,\mathrm{d}\mathbb{P}(x) \leq \sqrt{\nu} \qquad \forall k\in\mathcal{K}, x\in\mathcal{X},$$

and, therefore,

$$\begin{aligned}
\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_\mathcal{X}} &= \sup_{k\in\mathcal{K},x\in\mathcal{X}} \left| \int k(y,x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \right| \\
&= \sup_{k\in\mathcal{K},x\in\mathcal{X}} \left| \left\langle k(\cdot,x), \int k(\cdot,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \right\rangle_{\mathcal{H}_k} \right| \\
&\leq \sup_{k\in\mathcal{K},x\in\mathcal{X}} \|k(\cdot,x)\|_{\mathcal{H}_k} \mathfrak{D}_k(\mathbb{P},\mathbb{Q}) \leq \sqrt{\nu}\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H},
\end{aligned}$$

which proves the lower bound on $\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{F}_H}$ in (5.1). To prove the upper bound, consider

$$\begin{aligned}\|\mathbb{P}-\mathbb{Q}\|^2_{\mathcal{F}_H} &\overset{(3.7)}{=} \sup_{k\in\mathcal{K}} \int\int k(x,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\ &\le \sup_{k\in\mathcal{K}} \int \left|\int k(x,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right| \mathrm{d}|\mathbb{P}-\mathbb{Q}|(x) \\ &\le 2\sup_{k\in\mathcal{K}}\sup_{x\in\mathcal{X}}\left|\int k(x,y)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right| = 2\|\mathbb{P}-\mathbb{Q}\|_{\mathcal{K}_\mathcal{X}},\end{aligned}$$

thereby proving the result in (5.1). Equation (5.2) simply follows from Theorem 3.2 and (5.1).

6.7. Proof of Theorem 5.2

In order to prove Theorem 5.2, we need a lemma (see Lemma 6.2 below) which is based on the notion of *fat-shattering dimension* (see Anthony and Bartlett [1], Definition 11.10), defined as follows.

Definition 3 (Fat-shattering dimension). *Let $\mathcal{F}$ be a set of real-valued functions defined on $\mathcal{X}$. For every $\epsilon > 0$, a set $S = \{z_1,\dots,z_n\} \subset \mathcal{X}$ is said to be ϵ-shattered by $\mathcal{F}$ if there exists real numbers $r_1,\dots,r_n$ such that for each $b \in \{0,1\}^n$ there is a function $f_b \in \mathcal{F}$ with $f_b(z_i) \ge r_i + \epsilon$ if $b_i = 1$ and $f_b(z_i) \le r_i - \epsilon$ if $b_i = 0$, for $1 \le i \le n$. The fat-shattering dimension of $\mathcal{F}$ is defined as*

$$\mathrm{fat}_\epsilon(\mathcal{F}) = \sup\big\{|S| \,|\, S \subset \mathcal{X},\ S \text{ is } \epsilon\text{-shattered by } \mathcal{F}\big\}.$$

Lemma 6.2. *Define*

$$\mathcal{G} := \big\{\mathrm{e}^{-\sigma(\cdot-x)^2} \colon \sigma \in (0,\infty), x \in \mathbb{R}\big\}.$$

Then $\mathrm{fat}_\epsilon(\mathcal{G}) \le 1 + \lfloor\epsilon^{-1}\rfloor$. In addition, there exists a universal constant c' such that for every empirical measure $\mathbb{P}_n$, and every $0 < \epsilon \le 1$,

$$\log\mathcal{N}\big(\mathcal{G}, L^2(\mathbb{P}_n), \epsilon\big) \le c'\left(1+\frac{8}{\epsilon}\right)\log^2\left(\frac{2}{\epsilon}+\frac{16}{\epsilon^2}\right).$$

Proof. Since $\int_{-\infty}^{\infty}|\frac{\mathrm{d}g}{\mathrm{d}y}|\,\mathrm{d}y = 2 < \infty$ for all $g \in \mathcal{G}$ then $\mathcal{G} \subset BV(\mathbb{R})$ where $BV(\mathbb{R})$ is the space of functions of bounded variation on $\mathbb{R}$. Therefore, by Anthony and Bartlett [1], Theorem 11.12, we obtain $\mathrm{fat}_\epsilon(\mathcal{G}) \le 1 + \lfloor\epsilon^{-1}\rfloor$ and Mendelson [25], Theorem 3.2, ensures that there exists a universal constant c' such that for every empirical measure $\mathbb{P}_n$, and every $\epsilon > 0$,

$$\log\mathcal{N}\big(\mathcal{G}, L^2(\mathbb{P}_n), \epsilon\big) \le c'\mathrm{fat}_{\epsilon/8}(\mathcal{G})\log^2\left(\frac{2\mathrm{fat}_{\epsilon/8}(\mathcal{G})}{\epsilon}\right) \le c'\left(1+\frac{8}{\epsilon}\right)\log^2\left(\frac{2}{\epsilon}+\frac{16}{\epsilon^2}\right),$$

thereby yielding the result. □

Proof of Theorem 5.2. (a) Define

$$\mathcal{F}_i := \left\{ \mathrm{e}^{-\sigma_i(\cdot - x_i)^2} \colon \sigma_i \in (0,\infty), x_i \in \mathbb{R} \right\}, \qquad i = 1,\ldots,d.$$

By Lemma 6.2, it is easy to see that there exists $N_i(\epsilon) := \mathcal{N}(\mathcal{F}_i, L^2(\mathbb{P}_n), \epsilon)$ functions

$$\left\{ \mathrm{e}^{-\sigma_{i,1}(\cdot - x_{i,1})^2}, \ldots, \mathrm{e}^{-\sigma_{i,N_i(\epsilon)}(\cdot - x_{i,N_i(\epsilon)})^2} \right\} \subset \mathcal{F}_i$$

such that for any $\epsilon > 0$ and $f \in \mathcal{F}_i$, there exists $l \in \{1,\ldots,N_i(\epsilon)\}$ such that

$$\left\| f - \mathrm{e}^{-\sigma_{i,l}(\cdot - x_{i,l})^2} \right\|_{L^2(\mathbb{P}_n)} \le \epsilon.$$

Now pick $l_i \in \{1,\ldots,N_i(\epsilon)\}$, $i = 1,\ldots,d$. Then for $k(\cdot,x) = \mathrm{e}^{-\sigma\|\cdot - x\|^2}$, we have

$$\begin{aligned}
\left\| \mathrm{e}^{-\sigma\|\cdot - x\|^2} - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right\|_{L^2(\mathbb{P}_n)} &= \left\| \prod_{i=1}^d \mathrm{e}^{-\sigma(\cdot - x_i)^2} - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right\|_{L^2(\mathbb{P}_n)} \\
&\le \left\| \sum_{i=1}^d \left| \mathrm{e}^{-\sigma(\cdot - x_i)^2} - \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right| \right\|_{L^2(\mathbb{P}_n)} \\
&\le \sum_{i=1}^d \left\| \mathrm{e}^{-\sigma(\cdot - x_i)^2} - \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right\|_{L^2(\mathbb{P}_n)} \le \epsilon d.
\end{aligned}$$

This implies $\mathcal{N}(\mathcal{K}_\mathcal{X}, L^2(\mathbb{P}_n), \epsilon d) = \prod_{i=1}^d N_i(\epsilon)$ and, therefore,

$$\log \mathcal{N}\left(\mathcal{K}_\mathcal{X}, L^2(\mathbb{P}_n), \epsilon d\right) = \sum_{i=1}^d \log N_i(\epsilon),$$

which by Lemma 6.2 yields

$$\sup_n \sup_{\mathbb{P}_n} \log \mathcal{N}\left(\mathcal{K}_\mathcal{X}, L^2(\mathbb{P}_n), \epsilon\right) \le c' d \left(1 + \frac{8d}{\epsilon}\right) \log^2\left(\frac{2d}{\epsilon} + \frac{16d^2}{\epsilon^2}\right), \qquad 0 < \epsilon \le 1.$$

It is easy to verify that $\int_0^\infty \sup_n \sup_{\mathbb{P}_n} \log \mathcal{N}(\mathcal{K}_\mathcal{X}, L^2(\mathbb{P}_n), \epsilon) < \infty$. Therefore, $\mathcal{K}_\mathcal{X}$ is a universal Donsker class and the UCLTs follow.

(b) Following the setting in (a) above, for $k(\cdot,x) = \int_0^\infty \mathrm{e}^{-\sigma\|\cdot - x\|_2^2}\,\mathrm{d}\Lambda(\sigma)$, $\Lambda \in \mathcal{M}_A$, we have

$$k(\cdot,x) - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} = \int_0^\infty \left(\mathrm{e}^{-\sigma\|\cdot - x\|_2^2} - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right) \mathrm{d}\Lambda(\sigma)$$

and so

$$\left\| k(\cdot,x) - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right\|_{L^2(\mathbb{P}_n)} \le \int_0^\infty \left\| \mathrm{e}^{-\sigma\|\cdot - x\|_2^2} - \prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot - x_{i,l_i})^2} \right\|_{L^2(\mathbb{P}_n)} \mathrm{d}\Lambda(\sigma) \le \epsilon d,$$

and the claim as in (a) follows.

(c) The idea is similar to that of in (b) where for $k(\cdot,x)=\prod_{i=1}^d\int_0^\infty \mathrm{e}^{-\sigma(\cdot-x_i)^2}\,d\Lambda_i(\sigma)$, $\Lambda_i\in\mathcal{M}_{A_i}$, we have

$$\begin{aligned}\left\|k(\cdot,x)-\prod_{i=1}^d \mathrm{e}^{-\sigma_{i,l_i}(\cdot-x_{i,l_i})^2}\right\|_{L^2(\mathbb{P}_n)} &\le \left\|\sum_{i=1}^d\left|\int_0^\infty \mathrm{e}^{-\sigma(\cdot-x_i)^2}\,d\Lambda_i(\sigma)-\mathrm{e}^{-\sigma_{i,l_i}(\cdot-x_{i,l_i})^2}\right|\right\|_{L^2(\mathbb{P}_n)}\\ &\le \sum_{i=1}^d\left\|\int_0^\infty \mathrm{e}^{-\sigma(\cdot-x_i)^2}\,d\Lambda_i(\sigma)-\mathrm{e}^{-\sigma_{i,l_i}(\cdot-x_{i,l_i})^2}\right\|_{L^2(\mathbb{P}_n)}\\ &\le \sum_{i=1}^d\int_0^\infty \left\|\mathrm{e}^{-\sigma(\cdot-x_i)^2}-\mathrm{e}^{-\sigma_{i,l_i}(\cdot-x_{i,l_i})^2}\right\|_{L^2(\mathbb{P}_n)}\,d\Lambda_i(\sigma)\\ &\le \epsilon d,\end{aligned}$$

and the claim as in (a) follows.

(d) From (6.16), we have

$$k(x,y)=\frac{(c^2/4)^{\beta-d/2}}{\Gamma(\beta-d/2)}\int_0^\infty \mathrm{e}^{-\sigma\|x-y\|_2^2}\sigma^{d/2-\beta-1}\mathrm{e}^{-c^2/(4\sigma)}\,d\sigma,$$

which is of the form in (b) where $d\Lambda(\sigma)=\frac{(c^2/4)^{\beta-d/2}}{\Gamma(\beta-d/2)}\sigma^{d/2-\beta-1}\mathrm{e}^{-c^2/(4\sigma)}\,d\sigma$ and the result follows from (b). □

Appendix A: Supplementary results

In the following, we present supplementary results that are used in the proofs of Theorems 3.3 and 4.1. Before we present a result to bound $U_n(\mathcal{K};(X_i)_{i=1}^n)$, we need the following lemma. We refer the reader to de la Peña and Giné ([7], Proposition 4.3.1 and equation 5.1.9) for generalized versions of this result. However, here, we provide a bound with explicit constants.

Lemma A.1. *Let $\mathcal{A}$ be a finite subset of $\mathbb{R}^{l(l-1)/2}$ and $(\varepsilon_i)_{i=1}^l$ be independent Rademacher variables. For any $a\in\mathcal{A}$, define $a:=(a_{ij})_{1\le i<j\le n}$. Suppose $\sup_{a\in\mathcal{A}}\|a\|_2\le R<\infty$, then for any $0<\theta<1$,*

$$\mathbb{E}\sup_{a\in\mathcal{A}}\left|\sum_{i<j}^l \varepsilon_i\varepsilon_j a_{ij}\right|\le\frac{\mathrm{e}R}{\theta}\log\frac{|\mathcal{A}|}{1-\theta}\tag{A.1}$$

and, therefore,

$$\mathbb{E}\sup_{a\in\mathcal{A}}\left|\sum_{i<j}^l \varepsilon_i\varepsilon_j a_{ij}\right|<\mathrm{e}R\big(1+\sqrt{\log|\mathcal{A}|}\big)^2.\tag{A.2}$$

Proof. For $\lambda > 0$, consider

$$\mathrm{e}^{\lambda \mathbb{E}\sup_{a\in\mathcal{A}}|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|} \le \mathbb{E}\mathrm{e}^{\lambda \sup_{a\in\mathcal{A}}|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|} = \mathbb{E}\sup_{a\in\mathcal{A}} \mathrm{e}^{\lambda|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|}$$
$$\le \sum_{a\in\mathcal{A}} \mathbb{E}\mathrm{e}^{\lambda|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|} = \sum_{a\in\mathcal{A}} \mathbb{E}\sum_{c=0}^{\infty} \frac{\lambda^c|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|^c}{c!}.$$

By the hypercontractivity of homogeneous Rademacher chaos of degree 2 (de la Peña and Giné [7], Theorem 3.2.2), we have

$$\mathbb{E}\left|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}\right|^c \le (c-1)^c\left(\mathbb{E}\left|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}\right|^2\right)^{c/2} \le (c-1)^c\left(\sum_{i<j}^{l} a_{ij}^2\right)^{c/2}, \qquad c \ge 2$$

and

$$\mathbb{E}\left|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}\right| \le \left(\mathbb{E}\left|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}\right|^2\right)^{1/2} \le \left(\sum_{i<j}^{l} a_{ij}^2\right)^{1/2},$$

which implies

$$\mathrm{e}^{\lambda \mathbb{E}\sup_{a\in\mathcal{A}}|\sum_{i<j}^{l}\varepsilon_i\varepsilon_j a_{ij}|} \le \sum_{a\in\mathcal{A}}\sum_{c=0}^{\infty} \frac{\lambda^c c^c\|a\|_2^c}{c!}.$$

Using $c^c/c! \le \mathrm{e}^c$ and choosing $\lambda = \frac{\theta}{\mathrm{e}R}$ for some $0 < \theta < 1$, we obtain the desired result in (A.1). Using $-\log(1-\theta) < \theta/(1-\theta)$ for $0 < \theta < 1$ in (A.1) and taking infimum over $\theta \in (0, 1)$ (where the infimum is obtained at $\theta = \sqrt{\log|\mathcal{A}|}/(1+\sqrt{\log|\mathcal{A}|})$) yields (A.2). □

The following result is based on the standard chaining argument to obtain a bound on the expected suprema of the Rademacher chaos process of degree 2. While the reader can to refer to de la Peña and Giné [7], Corollary 5.18, for a general result to bound the expected suprema of the Rademacher chaos process of degree m, we present a bound with explicit constants and with the lower limit of the entropy integral away from zero. This allows one to handle classes whose entropy number grows polynomially (for $\beta \ge 1$ in Theorem 3.3) in contrast to the entropy integral bound in de la Peña and Giné [7], Equation 5.1.22, where the integral diverges to infinity. Similar modification to the Dudley entropy integral bound on the expected suprema of empirical processes is carried out in Mendelson [25].

Lemma A.2. *Suppose $\mathcal{G}$ is a class of real-valued functions on $\mathcal{X}\times\mathcal{X}$ and $(\varepsilon_i)_{i=1}^n$ be a independent Rademacher variables. Define $\beta := \sup_{g_1,g_2\in\mathcal{G}}\rho(g_1,g_2)$. Then, for any $(x_i)_{i=1}^n \subset \mathcal{X}$ and $0 < \theta < 1$,*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i<j}^{n}\varepsilon_i\varepsilon_j g(x_i,x_j)\right| \le 2\sqrt{2}n^2\left(\inf_{\alpha>0}\left\{\alpha + \frac{3\mathrm{e}}{\theta}\int_{\alpha}^{\beta}\frac{1}{n}\log\frac{\mathcal{N}(\mathcal{G},\rho,\epsilon)}{\sqrt{1-\theta}}\,\mathrm{d}\epsilon\right\}\right) + \frac{n}{\sqrt{2}}\sup_{g\in\mathcal{G}}\rho(g,0),$$

where for any $g_1, g_2 \in \mathcal{G}$, $\rho(g_1, g_2) = \sqrt{\frac{2}{n^2}\sum_{i<j}^n (g_1(x_i, x_j) - g_2(x_i, x_j))^2}$ *and therefore*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j g(x_i, x_j)\right| < 2\sqrt{2}n^2\left(\inf_{\alpha>0}\left\{\alpha + \frac{3\mathrm{e}}{n}\int_\alpha^\beta \left(1+\sqrt{\log\mathcal{N}(\mathcal{G},\rho,\epsilon)}\right)^2 \mathrm{d}\epsilon\right\}\right) + \frac{n}{\sqrt{2}}\sup_{g\in\mathcal{G}}\rho(g, 0).$$

Proof. Let $\delta_0 := \sup_{g_1,g_2\in\mathcal{G}}\rho(g_1, g_2)$ and for any $l \in \mathbb{N}$, let $\delta_l := 2^{-l}\delta_0$. For each $l \in \mathbb{N}\cup\{0\}$, let $\mathcal{G}_l := \{g_l^1, \dots, g_l^{\mathcal{N}(\mathcal{G},\rho,\delta_l)}\}$ be a ρ-cover of $\mathcal{G}$ at scale δ_l. For any M, any $g \in \mathcal{G}$ can be expressed as

$$g = (g - g_M) + \sum_{l=1}^M (g_l - g_{l-1}) + g_0,$$

where $g_l \in \mathcal{G}_l$ and $\mathcal{G}_0 := \mathcal{G}$. Note that $\rho(g_l, g_{l-1}) \le \rho(g, g_l) + \rho(g, g_{l-1}) \le \delta_l + \delta_{l-1} = 3\delta_l$. Consider

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j g(x_i, x_j)\right| \\
&\quad\le \mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j\big(g(x_i, x_j) - g_M(x_i, x_j)\big)\right| + \mathbb{E}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j g_0(x_i, x_j)\right| \\
&\qquad + \sum_{l=1}^M \mathbb{E}\sup_{\substack{g_l\in\mathcal{G}_l, g_{l-1}\in\mathcal{G}_{l-1}\\ \rho(g_l,g_{l-1})\le 3\delta_l}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j\big(g_l(x_i, x_j) - g_{l-1}(x_i, x_j)\big)\right|.
\end{aligned} \tag{A.3}$$

Note that

$$\begin{aligned}
\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j\big(g(x_i, x_j) - g_M(x_i, x_j)\big)\right| &\le \mathbb{E}\sum_{j=1}^n \varepsilon_j^2 \sup_{g\in\mathcal{G}}\sqrt{\sum_{i<j}^n\big(g(x_i, x_j) - g_M(x_i, x_j)\big)^2} \\
&= \frac{n^2}{\sqrt{2}}\sup_{g\in\mathcal{G}}\rho(g, g_M) \le \frac{n^2\delta_M}{\sqrt{2}},
\end{aligned} \tag{A.4}$$

$$\begin{aligned}
\mathbb{E}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j g_0(x_i, x_j)\right| &\le \left(\mathbb{E}\left|\sum_{i<j}^n \varepsilon_i\varepsilon_j g_0(x_i, x_j)\right|^2\right)^{1/2} \\
&\le \left(\sum_{i<j}^n g_0^2(x_i, x_j)\right)^{1/2} = \frac{n}{\sqrt{2}}\sup_{g\in\mathcal{G}}\rho(g, 0),
\end{aligned} \tag{A.5}$$

and by Lemma A.1,

$$\begin{aligned}
&\mathbb{E} \sup_{\substack{g_l \in \mathcal{G}_l, g_{l-1} \in \mathcal{G}_{l-1} \\ \rho(g_l, g_{l-1}) \le 3\delta_l}} \left| \sum_{i<j}^{n} \varepsilon_i \varepsilon_j \big(g_l(x_i, x_j) - g_{l-1}(x_i, x_j) \big) \right| \\
&\quad \le \frac{3\mathrm{e}\delta_l n}{\theta\sqrt{2}} \log \frac{\mathcal{N}(\mathcal{G}, \rho, \delta_l)\mathcal{N}(\mathcal{G}, \rho, \delta_{l-1})}{1-\theta} \\
&\quad \le \frac{6\mathrm{e}\delta_l n}{\theta\sqrt{2}} \log \frac{\mathcal{N}(\mathcal{G}, \rho, \delta_l)}{\sqrt{1-\theta}}
\end{aligned} \tag{A.6}$$

for any $0 < \theta < 1$. Using (A.4)–(A.6) in (A.3), we have

$$\begin{aligned}
&\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j}^{n} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| \\
&\le \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{6\mathrm{e}n}{\theta\sqrt{2}} \sum_{l=1}^{M} \delta_l \log \frac{\mathcal{N}(\mathcal{G}, \rho, \delta_l)}{\sqrt{1-\theta}} \\
&\le \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{12\mathrm{e}n}{\theta\sqrt{2}} \sum_{l=1}^{M} (\delta_l - \delta_{l+1}) \log \frac{\mathcal{N}(\mathcal{G}, \rho, \delta_l)}{\sqrt{1-\theta}} \\
&\le \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{12\mathrm{e}n}{\theta\sqrt{2}} \int_{\delta_{M+1}}^{\delta_0} \log \frac{\mathcal{N}(\mathcal{G}, \rho, \epsilon)}{\sqrt{1-\theta}} \, \mathrm{d}\epsilon.
\end{aligned} \tag{A.7}$$

For any $\alpha > 0$, pick $M := \sup\{l \colon \delta_l > 2\alpha\}$. This means $\delta_{M+1} \le 2\alpha$ and, therefore, $\delta_M = 2\delta_{M+1} \le 4\alpha$. On the other hand, $\delta_{M+1} > \alpha$ since $\delta_M > 2\alpha$. Using these bounds in (A.7), we obtain

$$\begin{aligned}
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j}^{n} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| &\le 2\sqrt{2} n^2 \alpha + \frac{6\sqrt{2}\mathrm{e}n}{\theta} \int_{\alpha}^{\sup_{g_1, g_2 \in \mathcal{G}} \rho(g_1, g_2)} \log \frac{\mathcal{N}(\mathcal{G}, \rho, \epsilon)}{\sqrt{1-\theta}} \, \mathrm{d}\epsilon \\
&\quad + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0).
\end{aligned} \tag{A.8}$$

Since α is arbitrary, taking infimum over $\alpha > 0$ yields the result. □

Appendix B: Bound on $\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon)$ in Theorem 4.2(a)

The following result presents a bound on $\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon)$ when

$$\mathcal{K} = \big\{ \mathrm{e}^{-\sigma \|x-y\|_2^2}, x, y \in (a_0, b_0)^d, -\infty < a_0 < b_0 < \infty : \sigma \in (0, a] \big\},$$

using which it is easy to check that $\omega_\star < 1$ in Theorem 4.1 and, therefore, the claims shown in Theorem 4.2 follow.

Proposition B.1. *Define* $\Sigma := (0, a]$ *and* $\mathcal{K}_\alpha := \{\partial^{\alpha,\alpha}\psi_\sigma(x-y), x, y \in (a_0, b_0)^d, -\infty < a_0 < b_0 < \infty : \sigma \in \Sigma\}$, *where* $\psi_\sigma(x) = \mathrm{e}^{-\sigma\|x-y\|_2^2}$ *and* $|\alpha| = r$. *Then*

$$\mathcal{N}(\mathcal{K}_\alpha, \rho_\alpha, \epsilon) = \frac{C}{\epsilon},$$

where ρ_α *is defined in Theorem 4.1 and* C *is a constant that depends on* a, a_0, b_0, d *and* r.

Proof. Let $\mathcal{N}(\Sigma, \|\cdot\|_1, \tau)$ be the τ-covering number of Σ and it is easy to verify that

$$N(\tau) := \mathcal{N}\big(\Sigma, \|\cdot\|_1, \tau\big) = \frac{a}{\tau}.$$

Let $\Sigma(\tau) := \{\sigma_1, \dots, \sigma_{N(\tau)}\}$ be the L^1 cover of Σ. Define $\widetilde{\mathcal{K}_\alpha} := \{\partial^{\alpha,\alpha}\psi_\sigma(x-y), x, y \in (a_0, b_0)^d, -\infty < a_0 < b_0 < \infty : \sigma \in \Sigma(\tau)\}$. Using the expression for $\partial^{\alpha,\alpha}\psi_\sigma$ in (6.7), we have

$$\begin{aligned}&\big|\partial^{\alpha,\alpha}\psi_\sigma - \partial^{\alpha,\alpha}\psi_{\sigma_l}\big|(x-y)\\ &\quad\le \sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d} A_{j_1\cdots j_d}\big|\sigma^{r+\sum_{i=1}^d j_i}\mathrm{e}^{-\sigma\|x-y\|_2^2} - \sigma_l^{r+\sum_{i=1}^d j_i}\mathrm{e}^{-\sigma_l\|x-y\|_2^2}\big|,\end{aligned}$$

where $A_{j_1\cdots j_d} := \prod_{i=1}^d |\eta_{ij_i}|(x_i - y_i)^{2j_i} \le (b_0 - a_0)^{2m}\prod_{i=1}^d |\eta_{ij_i}| =: B_{j_1\cdots j_d}$. Note that

$$C := \big|\sigma^{r+\sum_{i=1}^d j_i}\mathrm{e}^{-\sigma\|x-y\|_2^2} - \sigma_l^{r+\sum_{i=1}^d j_i}\mathrm{e}^{-\sigma_l\|x-y\|_2^2}\big|$$

can be bounded as

$$\begin{aligned}C &\le \big|\sigma^{r+\sum_{i=1}^d j_i} - \sigma_l^{r+\sum_{i=1}^d j_i}\big| + a^{r+\sum_{i=1}^d j_i}\big|\mathrm{e}^{-\sigma\|x-y\|_2^2} - \mathrm{e}^{-\sigma_l\|x-y\|_2^2}\big|\\ &\le \left(r + \sum_{i=1}^d j_i - 1\right)a^{r+\sum_{i=1}^d j_i - 1}|\sigma - \sigma_l| + a^{r+\sum_{i=1}^d j_i}|\sigma - \sigma_l|\|x-y\|_2^2\mathrm{e}^{a\|x-y\|_2^2}\\ &\le (2r-1)a^{2r-1}|\sigma - \sigma_l| + d(b_0 - a_0)^2 a^{2r}|\sigma - \sigma_l|\mathrm{e}^{ad(b_0-a_0)^2} \le \mu\tau,\end{aligned}$$

where μ is a constant that depends on a, a_0, b_0, d and r. Therefore,

$$\rho_\alpha\big(\partial^{\alpha,\alpha}\psi_\sigma, \partial^{\alpha,\alpha}\psi_{\sigma_l}\big) \le \big\|\partial^{\alpha,\alpha}\psi_\sigma - \partial^{\alpha,\alpha}\psi_{\sigma_l}\big\|_\infty \le \mu\tau\sum_{j_1=0}^{\alpha_1}\cdots\sum_{j_d=0}^{\alpha_d} B_{j_1\cdots j_d},$$

which yields the result. □

Acknowledgements

The work was carried out while the author was a Research Fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. The author profusely thanks Richard Nickl for many valuable comments and insightful discussions. The author also thanks the associate editor and two anonymous reviewers for their careful review and constructive comments which significantly improved the manuscript.

References

[1] Anthony, M. and Bartlett, P.L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge: Cambridge Univ. Press. MR1741038

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Trans*. *Amer. Math*. *Soc*. **68** 337–404. MR0051437

[3] Bartlett, P.L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *Ann*. *Statist*. **33** 1497–1537. MR2166554

[4] Berg, C., Christensen, J.P.R. and Ressel, P. (1984). *Harmonic Analysis on Semigroups*: *Theory of Positive Definite and Related Functions*. *Graduate Texts in Mathematics* **100**. New York: Springer. MR0747302

[5] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA: Kluwer Academic. MR2239907

[6] Bickel, P.J. and Ritov, Y. (2003). Nonparametric estimators which can be "plugged-in". *Ann*. *Statist*. **31** 1033–1053. MR2001641

[7] de la Peña, V.H. and Giné, E. (1999). *Decoupling*: *From Dependence to Independence*. *Probability and Its Applications* (*New York*). New York: Springer. MR1666908

[8] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation*: *The* L_1 *View*. *Wiley Series in Probability and Mathematical Statistics*: *Tracts on Probability and Statistics*. New York: Wiley. MR0780746

[9] Diestel, J. and Uhl, J.J. Jr. (1977). *Vector Measures*. Providence, RI: Amer. Math. Soc. MR0453964

[10] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge: Cambridge Univ. Press. MR1720712

[11] Dudley, R.M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge: Cambridge Univ. Press. MR1932358

[12] Folland, G.B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*. New York: Wiley.

[13] Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* 20 (J.C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 489–496. Cambridge, MA: MIT Press.

[14] Fukumizu, K., Sriperumbudur, B.K., Gretton, A. and Schölkopf, B. (2009). Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems* 21 (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 473–480. Cambridge, MA: MIT Press.

[15] Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields* **141** 333–387. MR2391158

[16] Giné, E. and Nickl, R. (2008). Adaptation on the space of finite signed measures. *Math*. *Methods Statist*. **17** 113–122. MR2429123

[17] Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann*. *Probab*. **37** 1605–1646. MR2546757

[18] Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596. MR2475673

[19] Giné, E. and Nickl, R. (2010). Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli* **16** 1137–1163. MR2759172

[20] Giné, E. and Zinn, J. (1986). Empirical processes indexed by Lipschitz functions. *Ann. Probab.* **14** 1329–1338. MR0866353

[21] Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B. and Smola, A. (2007). A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems* 19 (B. Schölkopf, J. Platt and T. Hoffman, eds.) 513–520. Cambridge, MA: MIT Press.

[22] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. New York: Springer. MR1618204

[23] Lepski, O.V., Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734

[24] Marcus, D.J. (1985). Relationships between Donsker classes and Sobolev spaces. *Z. Wahrsch. Verw. Gebiete* **69** 323–330. MR0787601

[25] Mendelson, S. (2002). Rademacher averages and phase transitions in Glivenko–Cantelli classes. *IEEE Trans. Inform. Theory* **48** 251–263. MR1872178

[26] Nickl, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* **138** 411–449. MR2299714

[27] Radulović, D. and Wegkamp, M. (2000). Weak convergence of smoothed empirical processes: Beyond Donsker classes. In *High Dimensional Probability, II (Seattle, WA*, 1999). *Progress in Probability* **47** 89–105. Boston, MA: Birkhäuser. MR1857317

[28] Rudin, W. (1991). *Functional Analysis*, 2nd ed. *International Series in Pure and Applied Mathematics*. New York: McGraw-Hill. MR1157815

[29] Srebro, N., Sridharan, K. and Tewari, A. (2010). Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems* 23 (J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel and A. Culotta, eds.) 2199–2207. MIT Press.

[30] Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Lanckriet, G.R.G. and Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems* 22 (Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams and A. Culotta, eds.) 1750–1758. Cambridge, MA: MIT Press.

[31] Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B. and Lanckriet, G.R.G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Stat.* **6** 1550–1599. MR2988458

[32] Sriperumbudur, B.K., Fukumizu, K. and Lanckriet, G. (2011). Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint. In *Advances in Neural Information Processing Systems* 24 (J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira and K.Q. Weinberger, eds.) 1773–1781. Cambridge, MA: MIT Press.

[33] Sriperumbudur, B.K., Fukumizu, K. and Lanckriet, G.R.G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

[34] Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G.R.G. (2010). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11** 1517–1561. MR2645460

[35] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines. Information Science and Statistics*. New York: Springer. MR2450103

[36] van der Vaart, A. (1994). Weak convergence of smoothed empirical processes. *Scand. J. Stat.* **21** 501–504. MR1310093

[37] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

[38] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. New York: Springer. MR1385671

[39] Wendland, H. (2005). *Scattered Data Approximation*. *Cambridge Monographs on Applied and Computational Mathematics* **17**. Cambridge: Cambridge Univ. Press. MR2131724

[40] Ying, Y. and Campbell, C. (2010). Rademacher chaos complexities for learning the kernel problem. *Neural Comput*. **22** 2858–2886. MR2760540

[41] Yukich, J.E. (1992). Weak convergence of smoothed empirical processes. *Scand*. *J*. *Stat*. **19** 271–279. MR1183201

Received July 2014 and revised February 2015