On the Optimal Estimation of Probability Measures in Weak and Strong Topologies

Bharath K. Sriperumbudur
Statistical Laboratory, Center for Mathematical Sciences
Wilberforce Road, Cambridge CB3 0WB, U.K.
bs493@statslab.cam.ac.uk

July 2, 2014

Abstract

Given random samples drawn i.i.d. from a probability measure $P$ (defined on say, $\mathbb{R}^d$), it is well-known that the empirical estimator is an optimal estimator of $P$ 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 $\| \cdot \|_F$ for certain choices of $F$ such that $\| \cdot \|_F$ metrizes the weak topology, where $\| P \|_F := \sup \{ \int f dP : f \in F \}$. In this paper, we investigate this problem of optimal estimation in weak and strong topologies by choosing $F$ to be a unit ball in a reproducing kernel Hilbert space (say $F_H$ 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 $\| \cdot \|_{F_H}$ metrizes the weak topology and the kernel density estimator (with $L^1$ optimal bandwidth) estimates $P$ at dimension independent optimal rate of $n^{-1/2}$ in $\| \cdot \|_{F_H}$ along with providing a uniform central limit theorem for the kernel density estimator.

MSC 2000 subject classification: Primary: 62G07; Secondary: 60F05

Keywords and phrases: adaptive estimation, bounded Lipschitz metric, exponential inequality, kernel density estimator, 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 $P$ defined on a separable metric space $\mathcal{X}$ with $P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ being the empirical measure induced by them. It is well known that $P_n$ is a consistent estimator of $P$ in weak sense as $n \to \infty$, i.e., for any bounded continuous real-valued function $f$ on $\mathcal{X}$, $\int f dP_n \xrightarrow{a.s.} \int f dP$ as $n \to \infty$, written as $P_n \Rightarrow P$. In fact, if nothing is known about $P$, then $P_n$ is probably the most appropriate estimator to use as it is asymptotically efficient and minimax in the sense of [37, Theorem 25.21, Equation (25.22); also see Example 25.24]. In addition, for any Donsker class of functions, $\mathcal{F}$, $\| P_n - P \|_{\mathcal{F}} = O_P(n^{-1/2})$, where

$$\| P_n - P \|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \int f dP_n - \int f dP \right|,$$

i.e., $P_n - P$ is asymptotically of the order of $n^{-1/2}$ uniformly in $\mathcal{F}$ and the processes $f \mapsto \sqrt{n} \int f d(P_n - P)$, $f \in \mathcal{F}$ converge in law to a Gaussian process in $\ell^\infty(\mathcal{F})$, called the $P$-Brownian bridge indexed by $\mathcal{F}$.
where $\ell^\infty(\mathcal{F})$ denotes the Banach space of bounded real-valued 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 ($1 \leq r \leq \infty$) in the sense of achieving the minimax rate over all estimators for densities in certain classes $[8, 22, 37]$. 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, i.e., $\| \cdot \|_F$ and $L^r$. While $\mathbb{P}_n$ is not adequate to estimate $p$, the question arises as to whether $\mathbb{P}_n^*$ defined as $\mathbb{P}_n^*(A) := \int_A p_n(x) \, dx$ for every Borel set $A \subset \mathbb{R}^d$, estimates $\mathbb{P}$ as good as $\mathbb{P}_n$ in the sense that $\| \mathbb{P}_n^* - \mathbb{P} \|_F = O_p(n^{-1/2})$, i.e.,

$$\sup_{f \in \mathcal{F}} \left| \int f(x)p_n(x) \, dx - \int f(x)p(x) \, dx \right| = O_p(n^{-1/2}),$$

and whether the processes $f \mapsto \sqrt{n} \int f(x)(p_n - p)(x) \, dx$, $f \in \mathcal{F}$ 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) holds has been addressed for the kernel density estimator $[41, 36, 16]$ and wavelet density estimator $[18]$ 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 $[27]$ and $[16]$ Section 4.2). For a $\mathbb{P}$-Donsker class $\mathcal{F}$, it is easy to show that (1) and the corresponding uniform central limit theorem (UCLT) hold if $\| \mathbb{P}_n^* - \mathbb{P} \|_F = o_p(n^{-1/2})$, i.e.,

$$\sup_{f \in \mathcal{F}} \left| \int f(x)p_n(x) \, dx - \int f(x) \, d\mathbb{P}_n(x) \right| = o_p(n^{-1/2}).$$

Several recent works $[6, 26, 16, 17, 18, 19]$ have shown that many popular density estimators on $\mathcal{X} = \mathbb{R}$, such as maximum likelihood estimator, kernel density estimator and wavelet estimator satisfy $[22]$ if $\mathcal{F}$ is $\mathbb{P}$-Donsker—the Donsker classes that were considered in these works are: functions of bounded variation, $\{1_{(-\infty,t)} : t \in \mathbb{R}\}$, Hölder, Lipschitz and Sobolev classes on $\mathbb{R}$. In other words, these works show that there exists estimators that are within a $\| \cdot \|_F$-ball of size $o_p(n^{-1/2})$ around $\mathbb{P}_n$ such that they estimate $\mathbb{P}$ consistently in $\| \cdot \|_F$ at the rate of $n^{-1/2}$, i.e., 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 $\| \cdot \|_{F_n}$ using a kernel density estimator, $p_n$ (with $L^1$ optimal bandwidth) on $\mathbb{R}^d$ where under certain conditions on $\mathcal{K}$, $\| \cdot \|_{F_n}$ with

$$\mathcal{F}_H := \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \| f \|_{\mathcal{H}_k} \leq 1 : f \in \mathcal{H}_k, k \in \mathcal{K} \right\}$$

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) $[2]$ (also see $[5]$ and $[35]$ 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 $\| \cdot \|_{F_n}$ (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 $[15]$ for only $d = 1$ where $\mathcal{F}$ is chosen to be a unit ball of bounded Lipschitz functions, $\mathcal{F}_{BL}$ on $\mathbb{R}$, defined as

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

with $\| \cdot \|_2$ being the Euclidean norm. In comparison, our work generalizes the result of $[15]$ to any $d$ by working with $\mathcal{F}_H$. 

2
We would like to mention that in principle our choice of $\mathcal{F}_H$ is abstract, but particular choices have crucial computational advantages (see below). 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 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 that $\| \mathbb{P}_n - \mathbb{P} \|_{\mathcal{F}_H} = O_{a.s.}(n^{-1/2})$ by bounding the expected suprema of $U$-processes—specifically, the homogeneous Rademacher chaos process of degree 2—indexed by a uniformly bounded Vapnik-Cervonenkis (VC)-subgraph class $\mathcal{K}$ (see [7] Chapter 5) for details on $U$-processes). Since Theorems 2 and 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 we present our second main result which provides an exponential inequality for the tail probabilities of $\| \mathbb{P}_n - \mathbb{P} \|_{\mathcal{F}_H} = \| \mathbb{P}_n * K_h - \mathbb{P}_n \|_{\mathcal{F}_H}$, where $\mathbb{P}_n * K_h$ is the kernel density estimator with bandwidth $h$, * represents the convolution and $K_h = h^{-d}K(h \cdot)$ 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 $U$-processes (see Theorem 2 and 3) 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 5 (also see Remark 4) that guarantee this behavior for the kernel density estimator. It proved a similar result for $\mathcal{F}_{BL}$ with $d = 1$ which can be generalized to any $d \geq 2$ using [31] Corollary 3.5). However, for $d > 2$, it can only be shown that the kernel density estimator with $L^1$ optimal bandwidth is within $\| \cdot \|_{\mathcal{F}_{BL}}$-ball of size $o(n^{-1/d})$—for $d = 2$, it is $o_p(\sqrt{\log n} / \sqrt{n})$—around $\mathbb{P}_n$, instead of $o_p(n^{-1/2})$ as with $\| \cdot \|_{\mathcal{F}_H}$.

Now given that [1] holds for $\mathcal{F} = \mathcal{F}_H$ (see Theorem 4 for detailed conditions and Theorem 5 for examples), it is of interest to know whether the processes $f \mapsto \sqrt{n} \int f (\mathbb{P}_n * K_h - \mathbb{P})$, $f \in \mathcal{F}_H$ 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 [16] Theorem 3 which ensure this UCLT in $\ell^\infty(\mathcal{F}_H)$ for any general $\mathcal{K}$ that induces $\mathcal{F}_H$, in Theorem 6 we present concrete examples of $\mathcal{K}$ for which $\mathcal{F}_H$ is $\mathbb{P}$-Donsker so that the following UCLTs are obtained: $\sqrt{n} (\mathbb{P}_n - \mathbb{P}) \rightsquigarrow \ell^\infty(\mathcal{F}_H)$ $\mathbb{G}_\mathcal{P}$ and $\sqrt{n} (\mathbb{P}_n * K_h - \mathbb{P}) \rightsquigarrow \ell^\infty(\mathcal{F}_H) \mathbb{G}_\mathcal{P}$, where $\mathbb{G}_\mathcal{P}$ 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 [15] Theorem 1 for $\mathcal{F}_{BL}$ with $d = 1$ under the condition that $\mathbb{P}$ satisfies $\int_{\mathbb{R}} |x|^\gamma \mu(dx) < \infty$ for some $\gamma > 1/2$, which shows that additional conditions are required on $\mathbb{P}$ to obtain a UCLT while working with $\mathcal{F}_{BL}$ 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 $\mathcal{F}_{BL}$), which we discuss in Section 4, 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 $\| \mathbb{P}_n * K_h - \mathbb{P}_n \|_{\mathcal{F}_{BL}}$ as in the former case, the problem reduces to a maximization problem over $\mathbb{R}$ in contrast to an infinite dimensional optimization problem in $\mathcal{F}_{BL}$. 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—e.g., 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 [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 [15] Theorem 1, [17] Theorem 2 and [19] Theorem 3). Along the lines of [15] Theorem 1, in Section 5 we also discuss the optimal adaptive estimation of $\mathbb{P}$ in weak and strong topologies.

Various notations and definitions that are used throughout the paper are collected in Section 2. The missing proofs of the results are provided in Section 4 and supplementary results are collected in an 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 $\|F\| := \sup_{x \in \mathcal{X}} |F(x)|$. $C(\mathcal{X})$ 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, $\| \cdot \|_\infty$, which we alternate denote as $\| \cdot \|_\infty$. $M_1(\mathcal{X})$ denotes the space of all Borel probability measures defined on $\mathcal{X}$ while $M_1(\mathcal{X})$ denotes the space of all finite signed Borel measures on $\mathcal{X}$. $L^r(\mathcal{X},\mu)$ denotes the Banach space of $r$-power ($r \geq 1$) $\mu$-integrable functions where $\mu$ is a Borel measure defined on $\mathcal{X}$. We will write $L^r(\mathcal{X})$ for $L^r(\mathcal{X},\mu)$ if $\mu$ is a Lebesgue measure on $\mathcal{X} \subseteq \mathbb{R}^d$. $W^r_2(\mathbb{R}^d)$ 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 $L^1(\mathbb{R}^d)$. $\mathcal{F}_{BL}$ denotes the unit ball of bounded Lipschitz functions on $\mathcal{X} = \mathbb{R}^d$ as shown in (4). A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ is called a positive definite (pd) kernel if, for all $n \in \mathbb{N}$, $(\alpha_1, \ldots, \alpha_n) \in \mathcal{C}^n$ and all $(x_1, \ldots, x_n) \in \mathcal{X}^n$, we have

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

where $\bar{\alpha}$ 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. $\mathcal{F}_H$ denotes the unit ball of RKHS functions indexed by a cone of positive definite kernels, $\mathcal{K}$ as shown in (4). The convolution $f * g$ of two measurable functions $f$ and $g$ on $\mathbb{R}^d$ is defined as $(f * g)(x) := \int_{\mathbb{R}^d} f(y)g(x-y)\,dy$, provided the integral exists for all $x \in \mathbb{R}^d$. Similarly, the convolution of $\mu \in M_1(\mathbb{R}^d)$ and measurable $f$ is defined as $(f * \mu)(x) := \int_{\mathbb{R}^d} f(x-y)\,d\mu(y)$ if the integral exists for all $x \in \mathbb{R}^d$.

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

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, $\{x_1, \ldots, x_n\}$, if for each of its $2^n$ subsets, there exists $C \in \mathcal{C}$ such that $C \cap \{x_1, \ldots, x_n\}$ yields the subset. The Vapnik-Chervonenkis (VC)-index, $VC(\mathcal{C})$ of the class $\mathcal{C}$ is the smallest $n$ for which no set of size $n$ is shattered by $\mathcal{C}$, i.e.,

$$VC(\mathcal{C}) = \inf \left\{ n : \max_{x_1, \ldots, x_n} |\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}| < 2^n \right\}. $$

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 $\mathcal{F}$, i.e., $\{(x,t): t < f(x)\} : f \in \mathcal{F}\}$ 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 $\{g : \rho(f,g) < \epsilon\}$ of radius $\epsilon$ needed to cover $\mathcal{F}$.

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

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

where $K : \mathbb{R}^d \rightarrow \mathbb{R}$ is the smoothing kernel that satisfies $K(x) = K(-x), \quad x \in \mathbb{R}^d$, $K \in L^1(\mathbb{R}^d)$ and $\int_{\mathbb{R}^d} K(x)\,dx = 1$ with $K_h(x) := h^{-d}K(x/h)$ and $0 < h := h_n \rightarrow 0$ as $n \rightarrow \infty$. $K$ is said to be of order
\( r > 0 \) if
\[
\int_{\mathbb{R}^d} \prod_{i=1}^{d} y_i^{\alpha_i} K(y) \, dy = 0 \text{ for } 0 < |\alpha| \leq r - 1, \quad \text{and} \quad \int_{\mathbb{R}^d} \prod_{i=1}^{d} |y_i|^{|\alpha_i|} |K(y)| \, dy < \infty, \quad \text{for } |\alpha| = r,
\]
where \( y = (y_1, \ldots, y_d) \), \( \alpha = (\alpha_1, \ldots, \alpha_d) \), \( \alpha_i \geq 0, \forall i = 1, \ldots, d \) and \( |\alpha| := \sum_{i=1}^{d} \alpha_i \). We refer the reader to [3, 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 [3]. 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 [3, 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) \( \forall y \in \mathcal{X}, \forall f \in \mathcal{H}_k, f(k(\cdot, y)) \in \mathcal{H}_k = f(y) \).

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 [3, Chapter 1, Theorem 1]. 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. Starting from Definition 1 it can be shown that \( \mathcal{H}_k = \text{span}\{k(\cdot, x) : x \in \mathcal{X}\} \) where the closure is taken w.r.t. the RKHS norm (see [3, Chapter 1, Theorem 3]), which means the kernel function, \( k \) generates the RKHS. Examples of \( k \) include the Gaussian kernel, \( k(x, y) = \exp(-\sigma \|x - y\|^2_2) \), \( x, y \in \mathbb{R}^d, \sigma > 0 \) that induces the following Gaussian RKHS,

\[
\mathcal{H}_k = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \int |\hat{f}(\omega)|^2 e^{\|\omega\|^2_2/4\sigma} \, d\omega < \infty \right\}
\]

and the Matérn kernel, \( k(x, y) = \frac{2^{1-\beta}}{\Gamma(\beta)} \|x - y\|^{\beta-d/2} \mathcal{J}_{d/2-\beta}(\|x - y\|_2), x, y \in \mathbb{R}^d, \beta > d/2 \) that induces the Sobolev space, \( H^\beta_2 \),

\[
\mathcal{H}_k = H^\beta_2 = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \int (1 + \|\omega\|^2_2)^\beta |\hat{f}(\omega)|^2 \, d\omega < \infty \right\}.
\]

Here, \( \hat{f} \) denotes the Fourier transform of \( f \), \( \Gamma \) is the Gamma function and \( \mathcal{J}_\alpha \) is the modified Bessel function of the third kind of order \( \alpha \), where \( \nu \) controls the smoothness of \( k \). Clearly, \( L^2(\mathbb{R}^d) \) is not an RKHS as the Dirac distribution, \( \delta \notin L^2(\mathbb{R}^d) \) though it satisfies the reproducing property—also \( L^2(\mathbb{R}^d) \) does not consist of functions. An important characterization for positive definiteness on \( \mathbb{R}^d \) is given by Bochner’s theorem.
Theorem 1: where \( \nu \) presents an alternate representation to \( \| \cdot \| \) using which we discuss the relation of \( \| \cdot \| \) to other classical distances on probabilities. First, in Proposition 1, we provide an alternate expression for \( \| \cdot \| \) about the topology induced by \( \| \cdot \| \). This representation is particularly useful as it shows that \( \Phi \) can be seen as completely determined by the kernels, \( k \in \mathcal{K} \) and does not depend on the individual functions in the corresponding RKHS.

Alternate representation for \( \| \cdot \| \): The following result—a similar result is proved in 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 \( \| \cdot \| \). This representation is particularly useful as it shows that \( \| \cdot \| \) is completely determined by the kernels, \( k \in \mathcal{K} \) and does not depend on the individual functions in the corresponding RKHS.

Proposition 1. Define \( \mathcal{P}_k := \{ \mathbb{P} \in M_+^1(\mathcal{X}) : \int \sqrt{k(x, x)} \, d\mathbb{P}(x) < \infty, \forall k \in \mathcal{K} \} \) where for all \( k \in \mathcal{K} \), \( k(\cdot, x) : \mathcal{X} \to \mathbb{R} \) is measurable for all \( x \in \mathcal{X} \). Then for any \( \mathbb{P}, \mathbb{Q} \in \mathcal{P}_k \),

\[
\| \mathbb{P} - \mathbb{Q} \|_{\mathcal{F}_k} = \sup_{k \in \mathcal{K}} \mathcal{D}_k(\mathbb{P}, \mathbb{Q})
\]

(9)

where

\[
\mathcal{D}_k(\mathbb{P}, \mathbb{Q}) := \left\| \int k(\cdot, x) \, d\mathbb{P}(x) - \int k(\cdot, x) \, d\mathbb{Q}(x) \right\|_{\mathcal{H}_k}
\]

(10)

\[
= \sqrt{\int \int k(x, y) \, d(\mathbb{P} - \mathbb{Q})(x) \, d(\mathbb{P} - \mathbb{Q})(y)},
\]

(11)

with \( \int k(\cdot, x) \, d\mathbb{P}(x) \) and \( \int k(\cdot, x) \, d\mathbb{Q}(x) \) being defined in Bochner sense [2, Definition 1].

It follows from (9) and (10) that \( \| \mathbb{P} - \mathbb{Q} \|_{\mathcal{F}_k} \) can be interpreted as the supremum distance between the embeddings \( \mathbb{P} \mapsto \int k(\cdot, x) \, d\mathbb{P}(x) \) and \( \mathbb{Q} \mapsto \int k(\cdot, x) \, d\mathbb{Q}(x) \), indexed by \( k \in \mathcal{K} \). Choosing \( k(\cdot, x) \) as

\[
\frac{1}{(2\pi)^d/2} e^{-\sqrt{-1} \pi x_2}, \quad e^{\langle x, x \rangle_2} \quad \text{and} \quad \frac{1}{(4\pi)^d/4} e^{-\|x\|_2^2/4}, \quad x \in \mathbb{R}^d,
\]

(12)

the embedding \( \Phi : M_+^1(\mathcal{X}) \to \mathcal{H}_k \), \( \mathbb{P} \mapsto \int k(\cdot, x) \, d\mathbb{P}(x) \) reduces to the characteristic function, moment generating function (if it exists) and Weierstrass transform of \( \mathbb{P} \) respectively. In this sense, \( \Phi \) 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). In particular, defining
\[
k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{L^2}, \quad x, y \in \mathbb{R}^d
\]
for the choices of \( k(\cdot, x), x \in \mathbb{R}^d \) in \([12]\) yields \( \mathcal{D}_k(\mathbb{P}, \mathbb{Q}) \) as the \( L^2 \) distance between the characteristic functions (resp. moment generating functions and Weierstrass transforms) of \( \mathbb{P} \) and \( \mathbb{Q} \) assuming their densities (w.r.t. the Lebesgue measure) to exist.

**When is \( \| \cdot \|_{\mathcal{F}_K} \) a metric on \( \mathcal{P}_k \)?** While \( \| \cdot \|_{\mathcal{F}_K} \) is a pseudo-metric on \( \mathcal{P}_k \), it is in general not a metric as \( \| \mathbb{P} - \mathbb{Q} \|_{\mathcal{F}_K} = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q} \) as shown by the choice \( \mathcal{K} = \{ k \} \) where \( k(x, y) = (x, y)^2, \, x, y \in \mathbb{R}^d \).

For this choice, it is easy to check that \( \| \mathbb{P} - \mathbb{Q} \|_{\mathcal{F}_K} \) is the Euclidean distance between the means of \( \mathbb{P} \) and \( \mathbb{Q} \) and therefore is not a metric on \( \{ \mathbb{P} \in M^1_+(\mathbb{R}^d) : \int \| x \| \, d\mathbb{P}(x) < \infty \} \) (and hence on \( M^1_+(\mathbb{R}^d) \)). The question of when is \( \mathcal{D}_k \) a metric on \( M^1_+(\mathcal{X}) \) is addressed in \([13, 14, 21, 34]\). By defining any kernel for which \( \mathcal{D}_k \) is a metric as the characteristic kernel, it is easy to see that if any \( k \in \mathcal{K} \) is characteristic, then \( \| \cdot \|_{\mathcal{F}_K} \) is a metric on \( \mathcal{P}_k \). \([34]\) Theorem 7 showed that \( k \) is characteristic if and only if
\[
\int \int k(x, y) \, d\mu(x) \, d\mu(y) > 0, \, \forall \mu \in M_b(\mathcal{X}) \text{ with } \mu(\mathcal{X}) = 0. \tag{13}
\]
Combining this with the Bochner characterization for positive definiteness (see \([7]\), \([34]\) Corollary 4) showed that
\[
\mathcal{D}_k(\mathbb{P}, \mathbb{Q}) = \| \phi_\mathbb{P} - \phi_\mathbb{Q} \|_{L^2(\mathbb{R}^d, \gamma)}, \quad \mathbb{P}, \mathbb{Q} \in M^1_+(\mathbb{R}^d),
\]
using which \( k \) is shown to be characteristic if and only if \( \text{supp}(\Upsilon) = \mathbb{R}^d \). 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 \([13, 14, 21]\), 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 \([33]\) 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 \( \| \cdot \|_{\mathcal{F}_K} \) is a metric on \( M^1_+(\mathbb{R}^d) \).

1. Gaussian: \( \mathcal{K} = \left\{ e^{-\sigma \| x-y \|^2_2}, \, x, y \in \mathbb{R}^d : \sigma \in (0, \infty) \right\} \);
2. Laplacian: \( \mathcal{K} = \left\{ e^{-\sigma \| x-y \|_1}, \, x, y \in \mathbb{R}^d : \sigma \in (0, \infty) \right\} \);
3. Matérn: \( \mathcal{K} = \left\{ \frac{c^{2d-\alpha}}{\Gamma(\beta-\frac{d}{2})2^{\beta-1-\frac{d}{2}}} \| x-y \|_2^{\beta-d/2} \delta_{d/2-\beta}(c \| x-y \|_2), \, x, y \in \mathbb{R}^d, \, \beta > d/2 : c \in (0, \infty) \right\} \);
4. Inverse multiquadrics: \( \mathcal{K} = \left\{ 1 + \| x-y \|_2^{2-\beta}, \, x, y \in \mathbb{R}^d, \, \beta > 0 : c \in (0, \infty) \right\} \);
5. Splines: \( \mathcal{K} = \left\{ \prod_{j=1}^d \left( 1 - \frac{|x_j-y_j|}{c_j} \right) 1_{\{ |x_j-y_j| \leq c_j \}}, \, x, y \in \mathbb{R}^d : c_j \in (0, \infty), \, \forall j = 1, \ldots, d \right\} \);
6. Radial basis functions: \( \mathcal{K} = \left\{ \int_{(0, \infty)} e^{-\sigma \| x-y \|^2_2} \, d\Lambda(\sigma), \, x, y \in \mathbb{R}^d : \Lambda \in M^1_+(((0, \infty)) \right\} \).

In all these examples, it is easy to check that every \( k \in \mathcal{K} \) is bounded and characteristic (as \( \text{supp}(\Upsilon) = \mathbb{R}^d \) or in turn satisfies \([13]\)) and therefore \( \| \cdot \|_{\mathcal{F}_K} \) is a metric on \( M^1_+(\mathbb{R}^d) \).

**Topology induced by \( \| \cdot \|_{\mathcal{F}_K} \):** \([34]\) showed that for any bounded kernel \( k \), \( \mathcal{D}_k(\mathbb{P}, \mathbb{Q}) \leq \sqrt{C}TV(\mathbb{P}, \mathbb{Q}) \), where \( TV \) is the total variation distance and \( C := \sup \left\{ \sqrt{K(x, x)} : x \in \mathcal{X} \right\} \). This means there can be two distinct \( \mathbb{P} \) and \( \mathbb{Q} \) which need not be distinguished by \( \mathcal{D}_k \) but are distinguished in total variation, i.e., \( \mathcal{D}_k \) 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 \( \| \cdot \|_{\mathcal{F}_K} \). The following result shows that under additional conditions on \( \mathcal{K} \), \( \| \cdot \|_{\mathcal{F}_K} \) metrizes the weak-topology on \( M^1_+(\mathcal{X}) \). A special case of this result is already proved in \([34]\) Theorem 23 for \( \mathcal{D}_k \) when \( \mathcal{X} \) is compact.
Theorem 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

$$
\int \int k(x, y) \, d\mu(x) \, d\mu(y) > 0, \ \forall \mu \in M_0(\mathcal{X}) \setminus \{0\}.
$$

Then

$$
\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \to 0 \quad \implies \quad \mathbb{P}_n \rightharpoonup \mathbb{P} \text{ as } n \to \infty.
$$

(b) If $\mathcal{K}$ is uniformly bounded i.e., $\sup_{k \in \mathcal{K}, x \in \mathcal{X}} k(x, x) < \infty$ and satisfies the following property,

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

Then

$$
\mathbb{P}_n \rightharpoonup \mathbb{P} \quad \implies \quad \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} \to 0 \text{ as } n \to \infty.
$$

Proof. (a) Define $\mathcal{H}_* \subset C_0(\mathcal{X})$ 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 [33, Lemma 4.28] (also see [32, Theorem 5]), it follows that the inclusion $\mathcal{H}_* \subset C_0(\mathcal{X})$ is well-defined and continuous. In addition, since $k_* \in \mathcal{K}$ satisfies [14], it follows from [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 [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

$$
\mathcal{H}_*^+ := \left\{ \mu \in M_0(\mathcal{X}) : \int f \, d\mu = 0, \ \forall f \in \mathcal{H}_* \right\} = \{0\}.
$$

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

$$
\mu \mapsto \int k(\cdot, x) \, d\mu(x), \ \mu \in M_0(\mathcal{X})
$$

is injective, which is then equivalent to [14]. 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 $\|f - g\|_{\infty} \leq \epsilon$. Therefore,

$$
\|\mathbb{P}(n)f - \mathbb{P}f\| = \|\mathbb{P}(n)(f - g) + \mathbb{P}(g - f) + \mathbb{P}(n)g - \mathbb{P}g\| \\
\leq \mathbb{P}(n)\|f - g\| + |\mathbb{P}f - \mathbb{P}g| + |\mathbb{P}g - \mathbb{P}f| \\
\leq 2\epsilon + \|f - g\|_{\mathcal{H}_*} \mathbb{D}_b(\mathbb{P}(n), \mathbb{P}) \\
\leq 2\epsilon + \|f - g\|_{\mathcal{H}_*} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H}.
$$

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 [4, p. 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

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

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 $f \in \mathcal{F}_H$,

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

which implies $\mathcal{F}_H$ is equicontinuous on $\mathcal{X}$. The result therefore follows from [11, Corollary 11.3.4] which shows that if $\mathbb{P}(n) \rightharpoonup \mathbb{P}$ then $\mathbb{P}(n)$ converges to $\mathbb{P}$ in $\|\cdot\|_{\mathcal{F}_H}$ as $n \to \infty$.\hfill\qed
Assume there exists $\nu > 1$ least

Example 2. K

Therefore for the tail probability of $\hat{P}$ for the weak convergence of $K$

Comparing (13) and (14), it is clear that one requires a stronger condition for weak convergence than for $\| \cdot \|_{x_H}$ being just a metric. However, these conditions can be shown to be equivalent for bounded continuous translation invariant kernels on $\mathbb{R}^d$, i.e., kernels of the type in (7). This is because if $k$ satisfies (7), then

and therefore (14) holds if and only if $\text{supp}(\Upsilon) = \mathbb{R}^d$, which is indeed the characterization for $k$ being characteristic. Here, $\hat{\mu}$ represents the Fourier transform of $\mu$ defined as $\hat{\mu}(\omega) = \int e^{i\omega^T x} d\mu(x)$, $\omega \in \mathbb{R}^d$. Therefore for $\mathcal{K}$ in Example 1, convergence in $\| \cdot \|_{x_H}$ implies weak convergence on $M_1^d(\mathbb{R}^d)$. However, for the converse to hold, $\mathcal{K}$ has to satisfy (P) in Theorem 2 which is captured in the following example.

Example 2. The following families of kernels satisfy the conditions (14), (P) and the uniform boundedness condition of Theorem 2 so that $\| \cdot \|_{x_H}$ metrizes the weak topology on $M_1^d(\mathbb{R}^d)$.

1. $\mathcal{K} = \left\{ e^{-\sigma \| x-y \|^2_2}, x,y \in \mathbb{R}^d : \sigma \in (0,a], a < \infty \right\}$
2. $\mathcal{K} = \left\{ e^{-\sigma \| x-y \|^2_2}, x,y \in \mathbb{R}^d : \sigma \in (0,a], a < \infty \right\}$
3. $\mathcal{K} = \left\{ \frac{x^{1-d}}{\Gamma(\beta - \frac{d}{2})} \frac{x^{\beta - d/2}}{x^{\frac{d}{2}}}, x,y \in \mathbb{R}^d, \beta > d/2 : c \in (0,a], a < \infty \right\}$
4. $\mathcal{K} = \left\{ (1 + \frac{x-y}{c})^{-\beta}, x,y \in \mathbb{R}^d, \beta > 0 : c \in [a,\infty), a > 0 \right\}$
5. $\mathcal{K} = \left\{ \prod_{i=1}^d (1 + \frac{x-y_i}{a_i})^{-1}, x,y \in \mathbb{R}^d : a_i \in [a_i,\infty), a_i > 0, \forall i = 1,\ldots,d \right\}$
6. $\mathcal{K}$ in Theorem 3

Rate of convergence of $\mathbb{P}_n$ to $\mathbb{P}$ in $\| \cdot \|_{x_H}$: The following result presents an exponential inequality for the tail probability of $\| \mathbb{P}_n - \mathbb{P} \|_{x_H}$, which in combination with Theorem 2 provides a convergence rate for the weak convergence of $\mathbb{P}_n$ to $\mathbb{P}$.

Theorem 3. Let $X_1,\ldots,X_n$ be random samples drawn i.i.d. from $\mathbb{P}$ defined on a measurable space $\mathcal{X}$. Assume there exists $\nu > 0$ such that $\sup_{k \in \mathcal{K}, x \in \mathcal{X}} k(x,x) \leq \nu$. Then for every $\tau > 0$, with probability at least $1 - 2e^{-\tau}$ over the choice of $(X_i)_{i=1}^n \sim \mathbb{P}^n$, $\| \mathbb{P}_n - \mathbb{P} \|_{x_H} \leq 4\sqrt{2} \inf_{\alpha > 0} \left\{ \alpha + \frac{4\epsilon}{\sqrt{n}} \int_0^{2\nu} \log N(\mathcal{K},\rho,\epsilon) d\epsilon \right\} + \frac{3\sqrt{2\nu}(\sqrt{2} + \sqrt{\tau})}{\sqrt{n}}$, (15)

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

$\rho(k_1,k_2) := \sqrt{\frac{2}{n^2} \sum_{i < j} (k_1(X_i,X_j) - k_2(X_i,X_j))^2}$. (16)

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

$\log N(\mathcal{K},\rho,\epsilon) \leq A \left( \frac{2\nu}{\epsilon} \right)^B, 0 < \epsilon < 2\nu$ (17)

then there exists constants $(D_1,D_2)$ (dependent only on $A$, $B$, $\nu$, $\tau$ and not on $n$) such that

$\mathbb{P}_n \left( \left\{ (X_1,\ldots,X_n) \in \mathcal{X}^n : \| \mathbb{P}_n - \mathbb{P} \|_{x_H} > \lambda(A,B,\nu,\tau) \right\} \right) \leq 2e^{-\tau}$, (18)

where

$\lambda(A,B,\nu,\tau) \leq \left\{ \begin{array}{ll} \frac{D_1}{\sqrt{n}}, & 0 < B < 1 \\
\frac{D_2 \log n}{n} + \frac{D_2}{\sqrt{n}}, & \beta = 1 \\
\frac{D_3}{n^\beta} + \frac{D_3}{\sqrt{n}}, & \beta > 1\end{array} \right.$ (19)
Remark 1. (i) If $K$ is a VC-subgraph, by [38, Theorem 2.6.7], there exists finite constants $B$ and $\alpha$ (that are not dependent on $n$) such that $\mathcal{N}(K, \rho, \epsilon) \leq B (2\nu/\epsilon)^{\alpha}$, $0 < \epsilon < 2\nu$, which implies there exists $A$ and $0 < \beta < 1$ such that (17) holds. By Theorem 3, this implies $\|P_n - P\|_{\ell^2} = O_p(n^{-1/2})$ and hence by Borell-Cantelli lemma, $\|P_n - P\|_{\ell^2} \to 0$ as $n \to \infty$. Therefore, it is clear that if $K$ is a uniformly bounded VC-subgraph, then

$$(Q) : \quad P_n \sim P \text{ a.s. as } n \to \infty$$

with a rate of convergence of $n^{-1/2}$. [40, Lemma 2]—also see Proposition [10]—showed that the Gaussian kernel family in Example 1 is a VC-subgraph (in fact, using the proof idea in Lemma 2 of [40] it can be easily shown that Laplacian and inverse multiquadric families are also VC-subgraphs) and therefore these kernel classes ensure $(Q)$ with a convergence rate of $n^{-1/2}$. Instead of directly showing the radial basis function (RBF) class to be a VC-subgraph, [10] see the proof of Corollary 1] bounded the expected suprema of the Rademacher chaos process of degree 2, i.e.,

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

(indexed by the RBF class, $K$, by that of the Gaussian class—see [63] in the proof of Theorem 5a)—and since the Gaussian class is a VC-subgraph, we obtain $U_n(K; (X_i)_{i=1}^n) = O_p(n)$ for the RBF class. We also show in Theorem 4(d) that $U_n(K; (X_i)_{i=1}^n) = O_p(n)$ for the Matérn kernel family. Using these bounds in [39] and following through the proof of Theorem 2 yields that $(Q)$ 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 $\| \cdot \|_{\ell^{2d}}$.

(ii) We would like to mention that Theorem 3 is a variation on [30, Theorem 7] where $U_n(K; (X_i)_{i=1}^n)$ is bounded by the entropy integral in [7, Corollary 5.1.8] with the lower limit of the integral being zero unlike in Theorem 3. This generalization (see [25, 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 [30, Theorem 7]. Also, compared to [30, Theorem 7], we provide explicit constants in Theorem 3.

4 Main Results

In this section, we present our main results of demonstrating the optimality of the kernel density estimator in $\| \cdot \|_{\ell^2}$ 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 $\|P_n * K_h - P\|_{\ell^2}$

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 $\|P_n * K_h - P\|_{\ell^2}$, where $P_n * K_h$ is the kernel density estimator. Since $\|P_n * K_h - P\|_{\ell^2} \leq \|P_n * K_h - P_n\|_{\ell^2} + \|P_n - P\|_{\ell^2}$, the result follows from Theorem 3 and bounding the tail probability of $\|P_n * K_h - P_n\|_{\ell^2}$, again through an application of McDiarmid’s inequality, which is captured in Theorem 4.

Theorem 4. Let $P$ 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 $P$ defined on an open subset $\mathcal{X}$ of $\mathbb{R}^d$. Assume $K$ satisfies the following:

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

(ii) For every $k \in K$, $\partial^{\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all multi-indexes $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$, $m \in \mathbb{N}$, where $\partial^{\alpha} := \partial_{a_1}^{\alpha_1} \cdots \partial_{a_d}^{\alpha_d} \partial_{1+d}^{\alpha_{1+d}} \cdots \partial_{2d}^{\alpha_{2d}}$;
(iii) \( \exists \nu > 0 \) such that \( \sup_{k \in K, x \in X} k(x, x) \leq \nu < \infty \);
(iv) For \( |\alpha| = m \wedge r \), \( \exists \nu_\alpha > 0 \) such that \( \sup_{k' \in K_\alpha, x \in X} k'(x, x) \leq \nu_\alpha < \infty \) where \( K_\alpha := \{ \partial^{\alpha} k : k \in K \} \),
where \( 1 \leq r \leq m + s \), \( r \in \mathbb{N} \) is the order of the smoothing kernel \( K \). Then for every \( \tau > 0 \), with probability at least \( 1 - 2e^{-\tau} \) 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 \( m, r, s, K, \tau, \nu, \nu_\alpha \) and not on \( n \)) such that
\[
\| K_h \ast P_n - P_n \|_{\mathcal{F}_H} \leq 4 \sqrt{2} h^{m \wedge r} \sum_{|\alpha| = m \wedge r} \Theta(\alpha) \sqrt{T(K_\alpha, \rho_\alpha, \nu_\alpha)} + \frac{A_1 h^{m \wedge r}}{\sqrt{n}} + A_2 h^r, \tag{21}
\]
and
\[
\| K_h \ast P_n - P \|_{\mathcal{F}_H} \leq 4 \sqrt{2} h^{m \wedge r} \sum_{|\alpha| = m \wedge r} \Theta(\alpha) \sqrt{T(K_\alpha, \rho_\alpha, \nu_\alpha)} + 4 \sqrt{2} \sqrt{T(K, \rho, \nu)} + \frac{B_1 h^{m \wedge r}}{\sqrt{n}} + \frac{B_2}{\sqrt{n}} + A_2 h^r, \tag{22}
\]
where \( m \wedge r := \min(m, r) \),
\[
T(K_\alpha, \rho_\alpha, \nu_\alpha) := \inf_{\delta > 0} \left\{ \delta + \frac{4 \epsilon}{n} \int_{\delta}^{2 \nu_\alpha} \log N(K_\alpha, \rho_\alpha, \epsilon) \, d\epsilon \right\}, \quad \Theta(\alpha) = \int \prod_{i=1}^d |\alpha_i|^{\alpha_i} |K(t)| \, dt,
\]
\( \rho \) is defined as in (16) and for any \( k_1, k_2 \in K_\alpha \),
\[
\rho_\alpha(k_1, k_2) = \sqrt{\frac{2}{n^2} \sum_{i<j}^n (k_1(X_i, X_j) - k_2(X_i, X_j))^2}.
\]
In addition, suppose there exists finite constants \( C_\alpha, C_K, \omega_\alpha \) and \( \omega_K \) (that are not dependent on \( n \)) such that
\[
\log N(K_\alpha, \rho_\alpha, \epsilon) \leq C_\alpha \left( \frac{2 \nu_\alpha}{\epsilon} \right)^{\omega_\alpha}, \quad 0 < \epsilon < 2 \nu_\alpha, \quad \text{for } |\alpha| = m \wedge r, \tag{23}
\]
and
\[
\log N(K, \rho, \epsilon) \leq C_K \left( \frac{2 \nu}{\epsilon} \right)^{\omega_K}, \quad 0 < \epsilon < 2 \nu. \tag{24}
\]
Define \( \omega_* := \max\{ \omega_\alpha : |\alpha| = m \wedge r \} \). If
\[
\sqrt{(\log n)^{1/(\omega_* + 1)} n^{(\omega_* + 1)/2}} h^{m \wedge r} \to 0, \quad \sqrt{n} h^r \to 0 \quad \text{as } h \to 0, \quad n \to \infty, \tag{25}
\]
then
\[
\| P_n \ast K_h - P_n \|_{\mathcal{F}_H} = o_{a.s.}(n^{-1/2}) \tag{26}
\]
and therefore
\[
\| P_n \ast K_h - P \|_{\mathcal{F}_H} = O_{a.s.} \left( \sqrt{(\log n)^{1/(\omega_K + 1)} n^{-\omega_K \wedge 1/2}} \right). \tag{27}
\]
Remark 2. (i) Theorem 4 shows that the kernel density estimator with bandwidth, \( h \) converging to zero sufficiently fast as given by conditions in (26), is within \( \| \cdot \|_{\mathcal{F}_H} \)-ball of size \( o_{a.s.}(n^{-1/2}) \) around \( P_n \) and behaves like \( P_n \) in the sense that it converges to \( P \) in \( \| \cdot \|_{\mathcal{F}_H} \) at a dimension independent rate of \( n^{-1/2} \)—see Theorem 3—as long as \( K \) is not too big, which is captured by \( \omega_K < 1 \) in (24). In addition, if \( K \) satisfies the conditions in Theorem 2, then the kernel density estimator converges weakly to \( P \) a.s. at the rate of \( n^{-1/2} \). Since we are interested in the optimality of \( P_n \ast K_h \) in both strong and weak topologies, it is interest to understand whether the asymptotic behavior in (26) 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 \( L^2 \) norm. It is easy to verify that if
\[
r > s + \frac{d}{2} \quad \text{and} \quad m > \frac{(2s+d)(\omega_* - 1)}{2\omega_*} \vee \frac{d}{2}, \tag{28}
\]
then \( h^* \) satisfies (25) and therefore \( \|K_{h^*} \ast P_n - P_n\|_{\mathcal{F}_H} = o_{a.s.}(n^{-1/2}) \) so that \( \|K_{h^*} \ast P_n - P\|_{\mathcal{F}_H} = O_{a.s.}(n^{-1/2}) \) 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 \( K_{h^*} \ast P_n \) with \( h = h^* \) is optimal in both weak (induced by \( \| \cdot \|_{\mathcal{F}_H} \)) and strong topologies unlike \( P_n \) which is only an optimal estimator of \( P \) in the weak topology. In Theorem 4, 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 (26), it can be shown that \( h^{**} \approx (n/\log n)^{-1/(2+d)} \), which is the optimal bandwidth for the estimation of \( p \) in sup-norm, also satisfies (25) and therefore (26) and (27) hold for \( h = h^{**} \).

(ii) The condition on \( r \) in (28) coincides with the one obtained for \( \{1_{(-\infty,t]} : t \in \mathbb{R}\} \) in (6) and bounded variation and Lipschitz classes with \( d = 1 \) in (16) 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 \( \omega_* \leq 1 \), \( \omega_\mathcal{K} < 1 \) and \( \mathcal{K} \) satisfying the conditions in Theorem 2), the order of the kernel has to be chosen higher by \( \frac{d}{2} \) than the usual (the usual being estimating \( K \) in sup-norm). An interesting aspect of the second condition in (28) is that the smoothness of kernels in \( \mathcal{K} \) should increase with either \( d \) or the size of \( \mathcal{K}_\alpha \) for \( P_n \ast K_h \) with \( h = h^* \) or \( h = h^{**} \) to lie in \( \| \cdot \|_{\mathcal{F}_H} \)-ball of size \( o_{a.s.}(n^{-1/2}) \) around \( P_n \). If \( \mathcal{K}_\alpha \) is large, i.e., \( \omega_\alpha > 1 \), 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 \( d \) (assuming \( \omega_* < 1 \)) for (20) to hold, it implies that the rate in (28) 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 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 (21) and (22) are of order \( h^{m+r}/\sqrt{n} \)—use the idea in Remark 3 ii) for (35) —and the second term in (22) is of order \( n^{-1/2} \) (see (60)). Therefore, the claims of Theorem 4 hold as if \( \omega_* \leq 1 \) and \( \omega_\mathcal{K} \leq 1 \).

**Proof.** Note that

\[
\|K_h \ast P_n - P_n\|_{\mathcal{F}_H} \leq \|K_h \ast (P_n - P) - (P_n - P)\|_{\mathcal{F}_H} + \|K_h \ast P - P\|_{\mathcal{F}_H}.
\]

(a) Bounding \( \|K_h \ast (P_n - P) - (P_n - P)\|_{\mathcal{F}_H} \):

Consider

\[
\|K_h \ast (P_n - P) - (P_n - P)\|_{\mathcal{F}_H} = \sup_{f \in \mathcal{F}_H} \left| \int f(x) d(K_h \ast (P_n - P))(x) - \int f(x) d(P_n - P)(x) \right|
\]

\[
= \sup_{f \in \mathcal{F}_H} \left| \int (f \ast K_h - f) d(P_n - P) \right|
\]

\[
= \|P_n - P\|_{\mathcal{G}},
\]

where \( \mathcal{G} := \{f \ast K_h - f : f \in \mathcal{F}_H\} \). We now obtain a bound on \( \|P_n - P\|_{\mathcal{G}} \) through an application of McDiarmid’s inequality. To this end, consider

\[
\|g\|_{\mathcal{G}} \leq \|f \ast K_h - f\|_{\mathcal{G}} = \sup_{x \in \mathcal{X}} \left| \int (f(x + ht) - f(x)) K(t) dt \right|.
\]

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

\[
|\partial^\alpha f(x)| \leq \|f\|_{\mathcal{H}_k} \sqrt{\partial^{\alpha \cdot \alpha} k(x, x), \ x \in \mathcal{X}}
\]

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

\[
f(x + th) - f(x) = \sum_{0 \leq |\alpha| \leq (m \wedge r) - 1} h^{|\alpha|} \Lambda_\alpha(t) \partial^\alpha f(x) + h^{m \wedge r} \sum_{|\alpha| = m \wedge r} \Lambda_\alpha(t) \partial^\alpha f(x + hD_\theta t),
\]

12
where
\[ \Lambda_\alpha(t) := \prod_{i=1}^{d} t_i^{\alpha_i}, \]
\( D_\theta = \text{diag}(\theta_1, \ldots, \theta_d) \) and \( 0 < \theta_i < 1 \) for all \( i = 1, \ldots, d \). Using (32) in (30) along with the regularity of \( K \), we have

\[
\| g \|_\infty \leq \| f * K_h - f \|_\infty \leq \h m^{\wedge r} \sup_{x \in X} \left| \int \Lambda_\alpha(t) K(t) \partial^\alpha f(x + hD_\theta t) \, dt \right|
\]
\[
\leq \h m^{\wedge r} \sup_{x \in X} \left| \int \Lambda_\alpha(|t|) |K(t)| |\partial^\alpha f(x + hD_\theta t)| \, dt \right|, \quad (33)
\]
where \( |t| := |t_i|, \forall i = 1, \ldots, d \). Using (31) in (33), for any \( g \in G \), we get

\[
\| g \|_\infty \leq \| f * K_h - f \|_\infty \leq \h m^{\wedge r} \sup_{k \in K, x \in X} \left| \int \Lambda_\alpha(|t|) |K(t)| \sqrt{\partial^\alpha \alpha k(x + hD_\theta t, x + hD_\theta t)} \, dt \right|
\]
\[
\leq \h m^{\wedge r} \sup_{|t| = m^{\wedge r}} \sqrt{\sup_{k \in K, x \in X} \partial^\alpha \alpha k(x, x)} \left| \int \Lambda_\alpha(|t|) |K(t)| \, dt \right|
\]
\[
= L_{m,r} \h m^{\wedge r}, \quad (34)
\]
where
\[
L_{m,r} := \sum_{|\alpha| = m^{\wedge r}} \sqrt{\nu_\alpha \Theta(\alpha)} < \infty.
\]

Now, let us consider

\[
\mathbb{E}\| \mathbb{P}_n - \mathbb{P} \|_G^* \leq 2 \frac{1}{n} \mathbb{E} \sup_{g \in G} \left| \sum_{i=1}^{n} \varepsilon_i g(X_i) \right| = 2 \frac{1}{n} \mathbb{E} \sup_{f \in F_H} \left| \sum_{i=1}^{n} \varepsilon_i (f * K_h - f)(X_i) \right|, \quad (35)
\]
where we have invoked the symmetrization inequality (see [38, Lemma 2.3.1]) in (\( \ast \)) with \( (\varepsilon_i)^n_{i=1} \) being the Rademacher random variables. By McDiarmid’s inequality, for any \( \tau > 0 \), with probability at least \( 1 - e^{-\tau} \),

\[
\mathbb{E}\| \mathbb{P}_n - \mathbb{P} \|_G \leq \mathbb{E}\| \mathbb{P}_n - \mathbb{P} \|_G + \| g \|_\infty \sqrt{\frac{2\tau}{n}}
\]
\[
\leq 2 \frac{1}{n} \mathbb{E} \sup_{f \in F_H} \left| \sum_{i=1}^{n} \varepsilon_i (f * K_h - f)(X_i) \right| + L_{m,r} \h m^{\wedge r} \sqrt{\frac{2\tau}{n}}, \quad (36)
\]
where (31) and (34) are used in (36). Define

\[
R_n(F_H) := \mathbb{E}_\varepsilon \sup_{f \in F_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f * K_h - f)(X_i) \right|
\]
where \( \mathbb{E}_\varepsilon \) denotes the expectation w.r.t. \( (\varepsilon_i)^n_{i=1} \) conditioned on \( (X_i)^n_{i=1} \). Applying McDiarmid’s inequality to \( R_n(F_H) \), we have for any \( \tau > 0 \), with probability at least \( 1 - e^{-\tau} \),

\[
\mathbb{E} \sup_{f \in F_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f * K_h - f)(X_i) \right| \leq R_n(F_H) + L_{m,r} \h m^{\wedge r} \sqrt{\frac{2\tau}{n}}, \quad (37)
\]

13
Bounding $R_n(F_H)$ yields

$$R_n(F_H) = \mathbb{E}_f \sup_{f \in F_H} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \int (f(X_i + th) - f(X_i))K(t) \, dt \right|$$

which yields

$$R_n(F_H) = \frac{h^{m^\wedge r}}{n} \mathbb{E}_x \sup_{f \in F_H} \left| \sum_{|\alpha| = m^\wedge r} \Lambda_\alpha(t)K(t) \sum_{j=1}^n \varepsilon_j \partial^\alpha f(X_j + hD_\theta t) \, dt \right|$$

Since $k$ is translation invariant, we have

$$R_n(F_H) = \frac{h^{m^\wedge r}}{n} \mathbb{E}_x \sup_{k \in K} \sum_{|\alpha| = m^\wedge r} \Theta(\alpha) \left( \sum_{i,j=1}^n \varepsilon_i \varepsilon_j \partial^\alpha k(X_i, X_j) \right) \leq \frac{\sqrt{2}h^{m^\wedge r}}{n} \sum_{|\alpha| = m^\wedge r} \Theta(\alpha) \mathbb{E}_x \sup_{k' \in K} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j \partial^\alpha k(X_i, X_j) + \frac{h^{m^\wedge r}L_{m^\wedge r}}{\sqrt{n}}$$

Combining (36), (37) and (40), we have that for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\left\| K_h \ast (\mathbb{P}_n - \mathbb{P}) - (\mathbb{P}_n - \mathbb{P}) \right\|_{F_H} \leq 4\sqrt{2}h^{m^\wedge r} \sum_{|\alpha| = m^\wedge r} \Theta(\alpha) \sqrt{\mathcal{T}(K_\alpha, p_\alpha, \nu_\alpha)} + \frac{A_1 h^{m^\wedge r}}{\sqrt{n}}.$$
where \( \tilde{f}(x) = f(-x) \). Since \( p \in L^1(\mathbb{R}^d) \) and \( \partial^\alpha f \) is bounded for all \(|\alpha| \leq m\), by \cite[Proposition 8.10]{12}, we have \( \partial^\alpha(f*p) = (\partial^\alpha f)*p \) for \(|\alpha| \leq m\). In addition, since \( \partial^\alpha f \) is continuous for all \(|\alpha| \leq m\) and \( \partial^\beta p \in L^1(\mathbb{R}^d) \) for \(|\beta| \leq s\), by extension of \cite[Lemma 5(b)]{10} to \( \mathbb{R}^d \), we have \( \partial^{\alpha+\beta}(f*p) = \partial^\alpha((\partial^\beta f)*p) = (\partial^\beta f)* (\partial^\beta p) \) for \(|\alpha| \leq m\), \(|\beta| \leq s\), which means for all \( f \in \mathcal{F}_H, f*p \) is \( m+s \)-differentiable. Therefore, using the Taylor series expansion of \((f*p)(ht)\) around zero (as in \cite[23]{12}), along with the regularity of \( K \) in \cite[42]{12}, we have

\[
\|P*K_h - P\|_{\mathcal{F}_H} = \sup_{f \in \mathcal{F}_H} h^r \sum_{|\alpha|+|\beta|=r} \int \Lambda_{\alpha+\beta}(t)K(t)(\partial^\alpha \tilde{f} + \partial^\beta p)(hD_0 t) \, dt \\
\leq h^r \sup_{f \in \mathcal{F}_H} \sum_{|\alpha|+|\beta|=r} \int \Lambda_{\alpha+\beta}(|t|)|K(t)||\partial^\alpha \tilde{f} + \partial^\beta p|(hD_0 t) \, dt \\
\leq h^r \sup_{f \in \mathcal{F}_H} \sum_{|\alpha|+|\beta|=r} \int \Lambda_{\alpha+\beta}(|t|)|K(t)|(\partial^\alpha \tilde{f} + |\partial^\beta p|)(hD_0 t) \, dt.
\]

Since

\[
(\partial^\alpha \tilde{f} + |\partial^\beta p|)(hD_0 t) = \int |\partial^\alpha f(x - hD_0 t)| |\partial^\beta p(x)| \, dx \\
\leq \int \sqrt{\partial^\alpha,\alpha hD_0 t, x - hD_0 t} |\partial^\beta p(x)| \, dx \\
\leq \|\partial^\beta p\|_{L^1(\mathbb{R}^d)} \sup_{k \in K, x \in \mathcal{X}} \partial^\alpha,\alpha k(x, x),
\]

using \cite[11]{14} in \cite[15]{14}, we obtain

\[
\|P*K_h - P\|_{\mathcal{F}_H} \leq A_2 h^r,
\]

where \( A_2 := \sum_{|\alpha|+|\beta|=r} \Theta(\alpha + \beta) \sqrt{\partial^\beta p} \|\partial^\beta p\|_{L^1(\mathbb{R}^d)} \) and \((\alpha + \beta)_i = \alpha_i + \beta_i, \forall i = 1, \ldots, d\). Using \cite[11]{14} and \cite[16]{14} in \cite[22]{14}, we obtain the result in \cite[21]{14}. Since

\[
\|K_h * P_n - P\|_{\mathcal{F}_H} \leq \|K_h * P_n - P_{n+1}\|_{\mathcal{F}_H} + \|P_n - P\|_{\mathcal{F}_H},
\]

the result in \cite[22]{14} follows from Theorem \cite[4]{14} and \cite[21]{14}. Under the entropy number conditions in \cite[24]{14}, it is easy to check—see \cite[10]{14}—that

\[
\sum_{|\alpha|=m^r} \Theta(\alpha) \sqrt{T(K_\alpha, \rho, \nu, \sigma)} = O \left( \sqrt{\log n} \frac{k(\omega_\alpha)}{n^{\frac{\omega_\alpha}{m^r}}} \right)
\]

and therefore \cite[25]{14} holds if \( h \) satisfies \cite[25]{14}. Using \cite[26]{14} and \cite[18]{14} in \cite[10]{14}, the result in \cite[27]{14} follows under the assumption that \( K \) satisfies \cite[24]{14}.

\textbf{Remark 3.} Since every \( k \in \mathcal{K} \) is translation invariant, an alternate proof can be provided by using the representation for \( \mathcal{D}_k \) (following \cite[13]{14}) in Proposition \cite[11]{14} \( \|P - Q\|_{\mathcal{F}_H} = \sup_\Upsilon \|\phi_P - \phi_Q\|_{L^2(\mathbb{R}^d, \Upsilon)} \) where the supremum is taken over all finite non-negative Borel measures on \( \mathbb{R}^d \). In this case, conditions on the derivatives of \( k \in \mathcal{K} \) translate into moment requirements for \( \Upsilon \). However, the current proof is more transparent as it clearly shows why the translation invariance of \( k \) is needed—see the inequalities following \cite[38]{14} and before \cite[38]{14}.

In the following result, we present some families of \( \mathcal{K} \) that ensure the claims of Theorems \cite[3]{14} and \cite[4]{14}.

\textbf{Theorem 5.} Suppose the assumptions on \( P \) and \( K \) in Theorem \cite[4]{14} hold and let \( 0 < a < \infty \). Then for the following classes of kernels,

(a)

\[
\mathcal{K} = \left\{ k(x, y) = \psi_\sigma(x - y), x, y \in \mathbb{R}^d : \sigma \in \Sigma \right\}
\]

where \( \psi_\sigma(x) = e^{-\sigma \|x\|^2} \) and \( \Sigma := (0, a] \);
(b) \[ K = \left\{ k(x, y) = \int_{0}^{\infty} \psi_{\sigma}(x - y) d\Lambda(\sigma), \; x, y \in \mathbb{R}^d : \; \Lambda \in \mathcal{M}_A \right\} \]

where \[ \mathcal{M}_A := \left\{ \Lambda \in M_+((0, \infty)) : \int_{0}^{\infty} \sigma^\beta d\Lambda(\sigma) \leq A < \infty \right\} \]

for some fixed \( A > 0 \);

(c) \[ K = \left\{ k(x, y) = \int_{(0, \infty)^d} e^{-(x-y)^T \Delta(x-y)} d\Lambda(\Delta), \; x, y \in \mathbb{R}^d : \; \Lambda \in \mathcal{Q}_A \right\} \]

where \( \Delta := \text{diag}(\sigma_1, \ldots, \sigma_d) \),

\[ \mathcal{Q}_A := \left\{ \Lambda \in M_+((0, \infty)^d) \mid \Lambda = \bigotimes_{i=1}^{d} \Lambda_i : \Lambda_i \in \mathcal{M}_{A_i}, \; i = 1, \ldots, d \right\} \]

and

\[ \mathcal{M}_{A_i} := \left\{ \Lambda_i \in M_+((0, \infty)) : \sup_{j \in \{1, \ldots, d\}} \int_{0}^{\infty} \sigma^{\alpha_j} d\Lambda_i(\sigma) \leq A_i < \infty \right\} \]

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

(d) \[ K = \left\{ k(x, y) = \sum_{i=1}^{d} \left( \frac{c}{\beta} \right)^{\frac{\beta - 1}{2}} \left( \frac{\beta}{2} \right)^{\frac{1}{2}} \right\} x - y \right\}_2, \; x, y \in \mathbb{R}^d, \; \beta > m + \frac{d}{2}, \; m \in \mathbb{N} : \; c \in (0, a] \right\}, \]

where \( A := \frac{2^{\frac{d}{2}} + 1 - \beta}{\Gamma(\beta - \frac{d}{2})} \).

\[ \|P_n - P\|_{\mathcal{F}_H} = O_{a,n} (n^{-1/2}), \; \|P_n \ast K_h - P_n\|_{\mathcal{F}_H} = O_{a,n} (n^{-1/2}) \text{ and } \|P_n \ast K_h - P\|_{\mathcal{F}_H} = O_{a,n} (n^{-1/2}) \] for any \( h \) satisfying \( \sqrt{\pi h} \to 0 \) as \( h \to 0 \) and \( n \to \infty \), which is particularly satisfied by \( h = h^* \) and \( h = h^{**} \) if \( r > s + \frac{d}{2} \).

Remark 4. (i) The Gaussian RKHS in [13] 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 \( \psi_{\sigma} \). This follows since for any \( f \in \mathcal{H}_{\sigma} \),

\[ \|f\|_{\mathcal{H}_{\sigma}}^2 := \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{-|\sigma|/4} e^{-|\omega|/4} d\omega = \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{-|\sigma|/4} e^{-|\omega|/4} e^{-\omega_0^2/4} d\omega \leq \left( \frac{T}{\sigma} \right)^{\frac{d}{2}} \|f\|_{\mathcal{H}_{\sigma}}^2, \] (47)

which implies for any \( k \in K, \mathcal{H}_k \subset \mathcal{H}_{\sigma} \), where the definition of \( \| \cdot \|_{\mathcal{H}_{\sigma}}^2 \) for any \( \tau > 0 \) in the first line of [17] is obtained from [39, Theorem 10.12]. From [17], it follows that

\[ \left\{ f \in \mathcal{H}_{\sigma} : \|f\|_{\mathcal{H}_{\sigma}} \leq \left( \frac{\sigma}{a} \right)^{d/4} \right\} \subset \left\{ f \in \mathcal{H}_a : \|f\|_{\mathcal{H}_a} \leq 1 \right\} \subset \mathcal{F}_H \]

and

\[ \mathcal{F}_H \subset \bigcup_{\sigma \in (0, a]} \left\{ f \in \mathcal{H}_a : \|f\|_{\mathcal{H}_a} \leq \left( \frac{a}{\sigma} \right)^{d/4} \right\} = \mathcal{H}_a, \] (48)

where \( \mathcal{F}_H := \{ \|f\|_{\mathcal{H}_a} \leq 1 : f \in \mathcal{H}_a, \sigma \in (0, a] \} \).

(ii) The kernel classes in (b) and (c) above are generalizations of the Gaussian family in (a). This can be seen by choosing \( \mathcal{M}_A = \{ \delta_{\sigma} : \sigma \in \Sigma \} \) 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 \( i = 1, \ldots, d \). By choosing

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

16
in (b), the inverse multiquadrics kernel family,

\[ K = \left\{ k(x, y) = \left( 1 + \frac{\|x - y\|}{c} \right)^{-\beta} \right\} \]

where the definition of \( K \) is obtained, where \( A = a^{-\frac{2\beta}{\Gamma(r + \beta)}} \) with \( \Gamma \) being the Gamma function. Similarly, choosing

\[ M_{A_i} = \{ dA_i(\sigma) = c_i^2 e^{-\sigma c_i^2} d\sigma : c_i \in [a_i, \infty), a_i > 0 \} \]

yields the family

\[ K = \left\{ \prod_{i=1}^{d} \left( 1 + \frac{|x_i - y_i|^2}{c_i^2} \right)^{-1} \right\} \]

in Example 2 where \( A_i = \sup_{x \in \{1, \ldots, d\}} \alpha_j a_i^{-2\alpha_j} \). 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 \( \inf_{\alpha \in M_A} \int_0^\infty e^{-\delta \sigma^2} dA(\sigma) \geq B \) (similarly, \( B_i > 0 \) and \( \delta_i > 0 \) such that \( \inf_{\alpha \in M_{A_i}} \int_0^\infty e^{-\delta_i \sigma^2} dA_i(\sigma) \geq B_i, i = 1, \ldots, d \)), where \( M_A \) and \( (M_{A_i})_{i=1}^d \) are defined in (b) and (c) of Theorem 3. 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 3 yields that for \( K \) in (a)–(c),

\[ P_n \sim P \quad \text{and} \quad P_n * K \sim P \quad \text{a.s.} \]

at the rate of \( n^{-1/2} \).

(iv) It is clear from (49) that any \( k \) in the Matérn family in Example 1—the family in Theorem 4(d)—is a special case of this—induces an RKHS which is a Sobolev space,

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

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 \( f \in H_c \),

\[ \|f\|_{H_c}^2 := \frac{2^{1-\beta}}{A\alpha^{2\beta-d\beta}} \int (\alpha^2 + \|\omega\|_2^{2\beta}) |\hat{f}(\omega)|^2 d\omega \leq \left( \frac{\alpha}{c} \right)^d \|f\|_{H_c}^2, \]

where the definition of \( \| \cdot \|_{H_c} \) follows from [39] Theorems 6.13 and 10.12. Therefore, we have

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

and

\[ \mathcal{F}_H \subset \bigcup_{c \in (0, a]} \left\{ f \in H_\alpha : \|f\|_{H_\alpha} \leq \left( \frac{\alpha}{c} \right)^{d/2} \right\} = H_a, \quad (50) \]

where \( \mathcal{F}_H := \{ \|f\|_{H_a} \leq 1 : f \in H_c, c \in (0, a] \} \). Unlike in Example 1, \( K \) in Theorem 4(d) requires \( \beta > m + \frac{d}{2} \). This is to ensure that every \( k \in K \) is \( m \)-times continuously differentiable as required in Theorem 4, which is guaranteed by the Sobolev embedding theorem [12] Theorem 9.17 if \( \beta > m + \frac{d}{2} \). Also, since \( K \) metrizes the weak topology (holds for the Gaussian family as well) on \( M_+^1(\mathbb{R}^d) \)—see Example 2—we obtain that

\[ P_n \sim P \quad \text{and} \quad P_n * K_h \sim P \quad \text{a.s.} \]

at the rate of \( n^{-1/2} \).

\[ \square \]
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 and 4 respectively and showed that \( \sqrt{n}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_H} = O_{a.s.}(1) \) and \( \sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} = O_{a.s.}(1) \) for families of \( K \) in Theorem 5. It is therefore easy to note that if \( \mathcal{F}_H \) is \( \mathbb{P} \)-Donsker, then \( \sqrt{n}\left(\mathbb{P}_n - \mathbb{P}\right) \Rightarrow_{L^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P} \) and so \( \sqrt{n}\left(\mathbb{P}_n * K_h - \mathbb{P}\right) \Rightarrow_{L^\infty(\mathcal{F}_H)} \mathbb{G}_\mathbb{P} \) (as \( \sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}\|_{\mathcal{F}_H} = o_{a.s.}(1) \)) for any \( h \) satisfying \( \sqrt{nh^r} \to 0 \) 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 and 4 which hold for a general \( H \) and \( \beta \) parameter fixed a priori, e.g., \( \sigma \) in the Gaussian kernel, \( \mathcal{F}_H \) in Theorem 6 are slightly constrained compared to those in Theorem 5 and Remark 4(a, b) satisfy the conditions in Theorem 6 and therefore yield a UCLT. Note that the kernel classes, \( \mathcal{K} \), in Theorem 6 are singleton consisting of a bounded continuous kernel. Therefore, we have

an optimal estimator of \( p \) in both strong and weak topologies unlike \( \mathbb{P}_n \), which estimates \( \mathbb{P} \) optimally.

Theorem 7 in [10] 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(d) shows that all the kernels (with the parameter fixed a priori, e.g., \( \sigma \) in the Gaussian kernel) we have encountered so far—such as in Examples 1 and 2—satisfy the conditions in Theorem 6 and therefore yield a UCLT. Note that the kernel classes, \( \mathcal{K} \) in Theorem 6 are slightly constrained compared to those in Theorem 5 and Remark 4(b). 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 5 yields the desired result in Theorem 6) critically hinges on the inclusion result shown in [38] and [39]—also see [35] for such an inclusion result for \( \mathcal{K} \) in Theorem 6(b). However this technique is not feasible for the kernel classes, (b) and (c) in Theorem 5 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, 4 and 6, 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 $\| \cdot \|_{F_{BL}}$ as the loss to measure the optimality of $P_n * K_h$ in the weak sense, the result does not hold for $d \geq 2$ as discussed before. In addition, for $d = 1$, the UCLT for $\sqrt{n}(P_n - P)$ and $\sqrt{n}(P_n * K_h - P)$ in $\ell^{\infty}(F_{BL})$ holds only under a certain moment condition on $P$, i.e., $\int |x|^{1+\gamma} \, dP(x) < \infty$ for some $\gamma > 0$ (see [20] Theorem 2) while no such condition on $P$ is required to obtain the UCLT for the above processes in $\ell^{\infty}(F_H)$ though both $\| \cdot \|_{F_{BL}}$ and $\| \cdot \|_{F_H}$ 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 $P$ in $\| \cdot \|_{F_H}$, i.e., in weak topology, similar to $P_n$. In Section 5.1, we present a similar result for an alternate metric $\| \cdot \|_{K_X}$ (defined below) that is topologically equivalent to $\| \cdot \|_{F_H}$, i.e., metrizes the weak topology on $M^+_1(\mathcal{X})$ where $\mathcal{X}$ is a topological space and $K_X \subset F_H$, showing that $F_H$ 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 $\| \cdot \|_{F_{BL}}$ for $d = 1$ in [15], there is a significant computational advantage associated with $F_H$ over $K_X$ and $F_{BL}$ 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 $\| \cdot \|_{K_X}$

In this section, we consider an alternate metric, $\| \cdot \|_{K_X}$, which we show in Proposition 7 to be topologically equivalent to $\| \cdot \|_{F_H}$ if $K$ is uniformly bounded, where

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

and $\mathcal{X}$ is a topological space. Note that $K_X \subset F_H$ if $k(x, x) \leq 1, \forall x \in \mathcal{X}, k \in K$, which means a reduced subset of $F_H$ is sufficient to metrize the weak topology on $M^+_1(\mathcal{X})$.

Proposition 7. Suppose $\nu := \sup_{k \in K, x \in \mathcal{X}} k(x, x) < \infty$. Then for any $P, Q \in M^+_1(\mathcal{X})$

$$\nu^{-1/2} \|P - Q\|_{K_X} \leq \|P - Q\|_{F_H} \leq 2^{1/2} \|P - Q\|_{K_X},$$

where

$$\|P - Q\|_{K_X} = \sup_{k \in K} \left\| \int k(\cdot, x)\, dP(x) - \int k(\cdot, x)\, dQ(x) \right\|_\infty.$$

In addition if $K$ satisfies the assumptions in Theorem 3 then for any sequence $(P_n)_{n \in \mathbb{N}} \subset M^+_1(\mathcal{X})$ and $P \in M^+_1(\mathcal{X})$,

$$\|P_n - P\|_{K_X} \to 0 \iff \|P_n - P\|_{F_H} \to 0 \iff (P_n) \rightsquigarrow P \text{ as } n \to \infty.$$ (52)

From (51), it simply follows that

$$\sqrt{n} \|P_n - P\|_{K_X} = O_{a.s.}(1), \sqrt{n} \|P_n * K_h - P_n\|_{K_X} = O_{a.s.}(1) \text{ and } \sqrt{n} \|P_n * K_h - P\|_{K_X} = O_{a.s.}(1)$$

for any $K$ in Theorem 3 with $\omega_s < 1$ and $\omega_K < 1$ (and therefore for any $K$ in Theorem 5 with $h$ satisfying $\sqrt{n}h^r \to 0$ as $h \to 0$ and $n \to \infty$). Therefore, if $K_X$ is $P$-Donsker, then for any $h$ satisfying these conditions, we obtain

$$\sqrt{n} (P_n * K_h - P) \rightsquigarrow \ell^\infty(K_X) \text{ in } \mathbb{P}.$$

The following result shows that $K_X$ is a universal Donsker class (i.e., $P$-Donsker for all probability measures $P$ on $\mathbb{R}^d$) for $K$ considered in Theorem 5 and therefore we obtain UCLT for $\sqrt{n}(P_n - P)$ and $\sqrt{n}(P_n * K_h - P)$ in $\ell^\infty(K_X)$.

Theorem 8. Suppose the assumptions on $P$ and $K$ in Theorem 5 hold. Define $K_X := \{k(\cdot, x) : k \in K, x \in \mathcal{X}\}$. Then for $K$ in Theorem $5$ $K_X$ is a universal Donsker class and

$$\sqrt{n}(P_n - P) \rightsquigarrow \ell^\infty(K_X) \text{ in } \mathbb{P} \quad \text{and} \quad \sqrt{n}(P_n * K_h - P) \rightsquigarrow \ell^\infty(K_X) \text{ in } \mathbb{P},$$

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 $r > s + \frac{d}{2}$. 

19
Combining Theorems 2 and Proposition 7 along with Theorems 5 and 8, it is clear that the kernel density estimator with bandwidth \( h^* \) is an optimal estimator of \( \rho \) in both strong and weak topologies (induced by \( \| \cdot \|_{K, \gamma} \)). While this result matches with the one obtained for \( \| \cdot \|_{K, \gamma} \), by comparing Theorems 6 and 8 we note that the convergence in \( \ell^\infty(K, \gamma) \) does not require the restriction in the parameter space as imposed in kernel classes for convergence in \( \ell^\infty(F_H) \) in Theorem 6. However, we show in the following section that \( \| \cdot \|_{F_H} \) is computationally easy to deal with than \( \| \cdot \|_{K, \gamma} \).

5.2 Adaptive Estimation & Computation

Let us return to the fact that there exists estimators that are \( Op(n^{-1/2}) \) from \( \mathbb{P}_n \) in \( \| \cdot \|_F \) (for suitable choice of \( F \)) and behave statistically similar to \( \mathbb{P}_n \). While we showed this fact through Theorems 3 and 5 for the kernel density estimator with \( F = F_H \) (and Proposition 1 for \( F = K_r \)), \( 16 \) \( 17 \) showed the same result in \( F \) being functions of bounded variation, \( \{1_{(-\infty,t]} : t \in \mathbb{R} \} \), Hölder, Lipschitz and Sobolev classes on \( \mathbb{R} \). Similar result is shown for wavelet density estimators and spline projection estimators in

\[
F = \{1_{(-\infty,t]} : t \in \mathbb{R} \} \quad \text{[18, 19]} \quad \text{and maximum likelihood estimators in} \quad \mathcal{F}_{BL} \quad \text{[20]}.
\]

While \( \mathbb{P}_n \) is simple and elegant to use in practice, these other estimators that are \( Op(n^{-1/2}) \) 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 \( F \) 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, \( 17 \) \( 19 \) proved the above behavior for kernel density estimator, wavelet density estimator and spline projection estimator on \( \mathbb{R} \) for \( F = \{1_{(-\infty,t]} : t \in \mathbb{R} \} \) and sup-norm loss over the Hölder balls. By choosing \( F = \mathcal{F}_{BL} \) (with \( d = 1 \)), \( 15 \) 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 \( 28 \) to kernel density estimators (in fact to any of the other estimators we discussed above) that are within a \( \| \cdot \|_F \)-ball of size smaller than \( n^{-1/2} \) around \( \mathbb{P}_n \) and then using the exponential inequality of the type in Theorem 3 to control the probability of the event that \( \sqrt{n}\|\mathbb{P}_n * K_h - \mathbb{P}_n \|_F \) is “too large” (see \( 15 \) Theorem 1, \( 17 \) Theorem 2 and \( 19 \) Theorem 3) for the optimality of the adaptive estimator in both \( \| \cdot \|_F \) and some relevant loss over prescribed class of densities). Using Theorem 3 it is quite straightforward in principle to construct an adaptive estimator that is optimal in both strong and weak topologies along the lines of \( 15 \) Theorem 1 by incorporating two minor changes in the proof of Theorem 1 in \( 15 \): the first change is to apply Theorem 3 in the place of \( 15 \) Lemma 1 and extend \( 15 \) Lemma 2 from \( \mathbb{R} \) to \( \mathbb{R}^d \). Informally, the procedure involves computing the bandwidth \( h_n \) as

\[
\hat{h}_n = \max \left\{ h \in \mathcal{H} : \|\mathbb{P}_n * (K_h - K_g)\|_{L^1} \leq \sqrt{\frac{A}{ng^d}}, \forall g < h, g \in \mathcal{H} \text{ and } \|\mathbb{P}_n * K_h - \mathbb{P}_n\|_F \leq \frac{n^{-rac{1}{2}}}{\log n} \right\},
\]

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

\[
\mathcal{P}(\gamma, L) = \left\{ \mathbb{P} \in M_1^1(\mathbb{R}^d) : \int (1 + \|x\|_2^2)^\gamma d\mathbb{P}(x) \leq L \right\}
\]

for some \( L < \infty \) and \( \gamma > \frac{d}{2} \). Along the lines of Theorem 1 in \( 15 \), the following result can be obtained (we state here without a proof) that shows the kernel density estimator with a purely data-driven bandwidth, \( \hat{h}_n \), to be optimal in both strong and weak topologies.

**Theorem 9.** 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 \( r > T + \frac{r}{2}, T \in \mathbb{N} \cup \{0\} \) such that \( \int_{\mathbb{R}^d}(1 + \|x\|_2^2)^r K^2(x) dx < \infty \) where \( p \in W^r_1(\mathbb{R}^d) \) for some \( 0 < s \leq T \). If \( F_H \) is \( \mathbb{P} \)-Donsker (satisfied by \( K \) in Theorem 6), then

\[
\|\mathbb{P}_n * K_{\hat{h}_n} - \mathbb{P}\|_{F_H} = Op(n^{-1/2}) \quad \text{and} \quad \sqrt{n}(\mathbb{P}_n * K_{\hat{h}_n} - \mathbb{P}) \overset{\ell^\infty(F_H)}{\rightarrow} \mathcal{G}_\mathbb{P}.
\]
Similarly, for $K$ in Theorem 5, we have
\[ \|P_n \ast K_h - P\|_{K_x} = O_P(n^{-1/2}) \quad \text{and} \quad \sqrt{n}(P_n \ast K_h - P) \rightarrow_{L^1} G_P. \]
In addition, for any $0 < s \leq T$,
\[ \|P_n \ast K_h - P\|_{L^s} = O_P \left( n^{-\frac{s}{s+d}} \right). \]

We now discuss some computational aspects of the estimator in (53), which requires computing $\|P_n \ast K_h - P\|_{L^1}$ and $\|P_n \ast K_h - P\|_F$. While computing $\|P_n \ast K_h - P\|_{L^1}$ is usually not straightforward, the computation of $\|P_n \ast K_h - P\|_F$ can be simple depending on the choice of $F$. In the following, we show that $F = F_H$ yields a simple maximization problem over a subset of $(0, \infty)$ depending on the choice of $K$, in contrast to an infinite dimensional optimization problem that would arise if $F = F_{BL}$ and optimization over $\mathbb{R}^d \times (0, \infty)$ if $F = K_X$, therefore demonstrating the computational advantage of working with $F_H$ over $F_{BL}$ and $K_X$.

Consider $\|P_n \ast K_h - P\|_{F_H}$, which from (10) and (11) yields
\[
\|P_n \ast K_h - P\|_{F_H} = \sup_{K \in X} \left\| \frac{1}{n} \sum_{i=1}^n \int K_h(x_i - x)k(\cdot, x) dx - \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) \right\|_{H_B}
\]
\[
= \frac{1}{n} \sup_{K \in K} \left\{ \sum_{i,j=1}^n A(X_i, X_j) + k(X_i, X_j) - 2 \int K_h(x)k(x, x_i, x_j) dx \right\},
\]
which in turn reduces to
\[
\|P_n \ast K_h - P\|_{F_H} = \frac{1}{n^2} \sup_{K \in K} \sum_{i,j=1}^n (K_h \ast K_h \ast \psi + \psi - 2K_h \ast \psi)(X_i - X_j)
\]
when $k$ is translation invariant, i.e., $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, where
\[
A(X_i, X_j) := \int \int K_h(x)K_h(y)k(x - x_i, x_j - y) dx dy.
\]
While computing (54) is not easy in general, in the following we present two examples where (54) is easily computable for appropriate choices of $K$ and $K$. Let $K$ be as in Theorem 5(a) (i.e., $\psi(x) = e^{-\sigma \|x\|^2}, x \in \mathbb{R}^d, \sigma \in \Sigma$) and $K = \pi^{-d/2} \psi_1$. Then
\[
\|P_n \ast K_h - P\|_{F_H} = \frac{1}{n} \sup_{\sigma \in \Sigma} \sum_{i,j=1}^n \left( \frac{\psi_{2\sigma h i+1}(X_i - X_j)}{(2\pi)^d(2\sigma h^2 + 1)^{d/2}} - \frac{\psi_{\sigma h i+1}(X_i - X_j)}{(2\pi)^d(\sigma h^2 + 1)^{d/2}} + \psi(X_i - X_j) \right). \quad (55)
\]
Also choosing $K$ to be as in Remark 4 i.e.,
\[
\psi(x) := \phi_\alpha(x) = \prod_{i=1}^d \frac{\alpha^2}{\alpha^2 + x_i^2}, x \in \mathbb{R}^d, \alpha \in [c, \infty), c > 0,
\]
which is a special case of $K$ in Theorem 5(c) and $K = \pi^{-d} \phi_1$ in (54) yields
\[
\|P_n \ast K_h - P\|_{F_H} = \frac{1}{n} \sup_{\alpha \in (c, \infty)} \sum_{i,j=1}^n \left( \frac{\phi_{\alpha + 2h}(X_i - X_j)}{\alpha^{-d} 2^d (\alpha + 2h)^d} - \frac{2\phi_{\alpha + h}(X_i - X_j)}{2^d \alpha^{-d}(\alpha + h)^d} + \phi(X_i - X_j) \right). \quad (55)
\]
In both these examples (where $\| \cdot \|_{F_H}$ metrizes the weak topology on $M_1^d(\mathbb{R}^d)$), it is clear that one can compute $\|P_n \ast K_h - P\|_{F_H}$ 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 5 holds if \( s < 2 - \frac{d}{2} \).

On the other hand, it is easy to see that
\[
\|\mathbb{P} \ast K_h - \mathbb{P} \|_{BL} = \frac{1}{n} \sup_{f \in \mathcal{F}_{BL}} \left| \sum_{i=1}^{n} (K_h \ast f)(X_i) \right|
\]
is not easily computable in practice. Also for \( \mathcal{F} = \mathcal{K}_\mathcal{X} \), we have
\[
\|\mathbb{P} \ast K_h - \mathbb{P} \|_{\mathcal{K}_\mathcal{X}} = \sup_{k \in \mathcal{K}, \mathcal{Y} \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)k(y, x) dx - \frac{1}{n} \sum_{i=1}^{n} k(y, X_i) \right|
\]
which reduces to
\[
\|\mathbb{P} \ast K_h - \mathbb{P} \|_{\mathcal{K}_\mathcal{X}} = \sup_{k \in \mathcal{K}, \mathcal{Y} \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} (K_h \ast \psi - \psi)(y - X_i) \right|
\]
when \( k(x, y) = \psi(x - y) \). For the choice of \( K \) and \( \mathcal{K} \) as above (i.e., \( \psi, \sigma \) and \( \phi, \alpha \)), it is easy to verify that the computation of \( \|\mathbb{P} \ast K_h - \mathbb{P} \|_{\mathcal{K}_\mathcal{X}} \) involves solving an optimization problem over \( \mathbb{R}^d \times (0, \infty) \) which is more involved than solving the one in (55) that is obtained by working with \( \mathcal{F}_{BL} \).

In addition to the above application of adaptive estimation, there are various statistical applications where the choice of \( \mathcal{F}_{BL} \) can be computationally useful (over \( \mathcal{F}_{BL} \) and \( \mathcal{K}_\mathcal{X} \)), the examples of which include the two-sample and independence testing. As an example, in two-sample testing, \( \|\mathbb{P} \ast K_h - \mathbb{P} \|_{\mathcal{K}_\mathcal{X}} \) 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.

6 Proofs

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

6.1 Proof of Proposition 1

For any \( f \in \mathcal{H}_k \) and \( \mathbb{P} \in \mathcal{P}_\mathcal{K} \), we have
\[
\int f(x) d\mathbb{P}(x) = \int \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} d\mathbb{P}(x) = \left\langle f, \int k(\cdot, x) d\mathbb{P}(x) \right\rangle_{\mathcal{H}_k},
\]
where the last equality follows from the assumption that \( k \) is Bochner-integrable, i.e.,
\[
\int \|k(\cdot, x)\|_{\mathcal{H}_k} d\mathbb{P}(x) = \int \sqrt{k(x, x)} d\mathbb{P}(x) < \infty.
\]
Therefore, for any \( \mathbb{P}, \mathbb{Q} \in \mathcal{P}_\mathcal{K} \),
\[
\sup_{k \in \mathcal{K}} \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int f(x) d(\mathbb{P} - \mathbb{Q})(x) \right| = \sup_{k \in \mathcal{K}} \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \int k(\cdot, x) d(\mathbb{P} - \mathbb{Q})(x) \right\rangle_{\mathcal{H}_k}
\]
\[
= \sup_{k \in \mathcal{K}} \left\| \int k(\cdot, x) d(\mathbb{P} - \mathbb{Q})(x) \right\|_{\mathcal{H}_k},
\]
where the inner supremum is attained at \( f = \frac{\int k(\cdot, x) d(\mathbb{P} - \mathbb{Q})(x)}{\int \|k(\cdot, x) d(\mathbb{P} - \mathbb{Q})(x)\|_{\mathcal{H}_k}} \). Because of the Bochner-integrability of \( k \),
\[
\left\langle \int k(\cdot, x) d\mathbb{P}(x), \int k(\cdot, y) d\mathbb{Q}(y) \right\rangle_{\mathcal{H}_k} = \int \int k(x, y) d\mathbb{P}(x) d\mathbb{Q}(y)
\]
and (11) follows.
6.2 Proof of Theorem 3

Since \( \sup_{k \in \mathcal{K}, x \in X} k(x, x) \leq \nu \) and \( \mathbb{P}_n, \mathbb{P} \in \mathcal{P}_K \), by Proposition 1, we have

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

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

\[
\sup_{k \in \mathcal{K}} \left\| \int k(\cdot, x) \, d(\mathbb{P}_n - \mathbb{P})(x) \right\|_{\mathcal{H}_k} \leq \mathbb{E} \sup_{k \in \mathcal{K}} \left\| \int k(\cdot, x) \, d(\mathbb{P}_n - \mathbb{P})(x) \right\|_{\mathcal{H}_k} + \sqrt{\frac{2 \nu \tau}{n}}. \tag{56}
\]

where \( \{ \varepsilon_i \}_{i=1}^n \) represent i.i.d. Rademacher variables, \( \mathbb{E}_x \) represents the expectation w.r.t. \( (\varepsilon_i)_{i=1}^n \) conditioned on \( (X_i)_{i=1}^n \), and \((*)\) is obtained by symmetrizing \( \mathbb{E} \sup_{k \in \mathcal{K}} \| \int k(\cdot, x) \, d(\mathbb{P}_n - \mathbb{P})(x) \|_{\mathcal{H}_k} \). Since \( \mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} \) satisfies the bounded difference property, another application of McDiarmid’s inequality yields that, for every \( \tau > 0 \), with probability at least \( 1 - e^{-\tau} \),

\[
\mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} \leq \mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} + \sqrt{\frac{2 \nu \tau}{n}}. \tag{57}
\]

and therefore combining (56) and (57) yields that for every \( \tau > 0 \), with probability at least \( 1 - 2e^{-\tau} \),

\[
\sup_{k \in \mathcal{K}} \left\| \int k(\cdot, x) \, d(\mathbb{P}_n - \mathbb{P})(x) \right\|_{\mathcal{H}_k} \leq 2 \mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} + \sqrt{\frac{18 \nu \tau}{n}}. \tag{58}
\]

Note that

\[
\mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} \leq \frac{1}{n} \sqrt{\mathbb{E} \sup_{k \in \mathcal{K}} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j k(X_i, X_j)} \leq \frac{\sqrt{2}}{n} \sqrt{U_n(\mathcal{K}; (X_i)_{i=1}^n)} + \frac{\sqrt{\nu}}{\sqrt{n}}. \tag{59}
\]

where

\[
U_n(\mathcal{K}; (X_i)_{i=1}^n) := \mathbb{E} \sup_{k \in \mathcal{K}} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j k(X_i, X_j)
\]

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

The result in (57) therefore follows by using (59) in (58) and bounding \( U_n(\mathcal{K}; (X_i)_{i=1}^n) \) through Lemma A.2 with \( \theta = \frac{2}{3} \). Using (17) in (15) and solving for \( \alpha \) yields (13) and (19). \( \square \)

Remark 5. (i) Note that instead of using McDiarmid’s inequality in the above proof, one can directly obtain a version of (58) by applying Talagrand’s inequality through Theorem 2.1 in [3], albeit with worse constants and similar dependency on \( n \).

(ii) If \( \mathcal{K} \) is singleton, i.e., \( \mathcal{K} = \{ k \} \), then l.h.s. of (59) can be bounded as

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

and therefore

\[
\mathbb{E} \sup_{k \in \mathcal{K}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} \leq \frac{\sqrt{\nu}}{\sqrt{n}}. \tag{60}
\]
6.3 Proof of Theorem 5

The proof involves showing that the kernels in (a)–(c) satisfy the conditions (i)–(iv) in Theorem 4 thereby ensuring that (21) and (22) hold. However, instead of bounding $\mathcal{T}$ through bounds on the covering numbers of $\mathcal{K}$ (see Remark 4 for a discussion about obtaining bounds on the covering numbers of $\mathcal{K}$), we directly bound the expected suprema of the Rademacher chaos process indexed by $\mathcal{K}$ and $\mathcal{K}_n$, i.e., $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_n, (X_i)_{i=1}^n)$ which are defined in (20)—note that the terms involving $\mathcal{T}$ in (21) and (22) are in fact bounds on $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_n, (X_i)_{i=1}^n)$—and show that $U_n(\mathcal{K}, (X_i)_{i=1}^n) = O_P(n)$ and $U_n(\mathcal{K}_n, (X_i)_{i=1}^n) = O_P(n)$. Using these results in (59) and (60) and following the proofs of Theorems 8 and 4 we have $\|P_n - P\|_{\mathcal{F}_H} = O_{a.s.}(n^{-1/2})$,

$$\|K_h * P_n - P_n\|_{\mathcal{F}_H} \leq \frac{E_1 h^r}{\sqrt{n}} + A_2 h^r$$

and

$$\|K_h * P_n - P\|_{\mathcal{F}_H} \leq \frac{F_1 h^r}{\sqrt{n}} + A_2 h^r + \frac{F_2}{\sqrt{n}}$$

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 8 and 4). Therefore the result follows.

In the following, we show that for $\mathcal{K}$ in (a)–(c), (iv) in Theorem 4 holds (note that (i)–(iii) in Theorem 4 hold trivially because of the choice of $\mathcal{K}$) along with $U_n(\mathcal{K}, (X_i)_{i=1}^n) = O_P(n)$ and $U_n(\mathcal{K}_n, (X_i)_{i=1}^n) = O_P(n)$. In order to obtain bounds on $U_n(\mathcal{K}, (X_i)_{i=1}^n)$ and $U_n(\mathcal{K}_n, (X_i)_{i=1}^n)$, we need an intermediate result (see Proposition 10 below)—also of independent interest—, which is based on the notion of pseudo-dimension (see 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, $\mathcal{F}$ [11 Chapter 11, p. 153].

Definition 2 (Pseudo-dimension). Let $\mathcal{F}$ be a set of real valued functions on $\mathcal{X}$ and suppose that $S = \{z_1, \ldots, z_n\} \subset \mathcal{X}$. Then $\mathcal{F}$ is pseudo-shattered by $\mathcal{F}$ if there are real numbers $r_1, \ldots, r_n$ 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 $i = 1, \ldots, n$. 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 10. Let

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

where $\theta \geq 0$ and $\delta_i > 0$ for any $i \in \{1, \ldots, d\}$. Then $VC(\mathcal{F}) \leq 2$. If $\theta = \delta_1 = \cdots = \delta_d = 0$, then $VC(\mathcal{F}) = 1$.

Proof. Suppose $VC(\mathcal{F}) \geq 3$. It means there exists a set $S = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^d : i \in \{1, 2, 3\}\}$ with property

$$(P_{21a}) : \|x_2 - y_2\|_2 < \|x_1 - y_1\|_2 < \|x_3 - y_3\|_2$$

which is pseudo-shattered by $\mathcal{F}$. This means, there exists $r_1, r_2$ and $r_3$ such that for any $b \in \{-1, 1\}^3$ there is a function $f_b \in \mathcal{F}$ with $\text{sign}(f_b(x_i, y_i) - r_i) = b_i$ for $i = 1, 2, 3$.

Consider $b = (b_1, b_2, b_3) = (-1, 1, 1)$. Then there exists $\sigma_1 \in (0, \infty)$ such that the following hold:

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

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

$$f_{\sigma_2}(x_1, y_1) \geq r_1, \ f_{\sigma_2}(x_2, y_2) < r_2, \ f_{\sigma_2}(x_3, y_3) < r_3.$$
This implies $f_{\sigma_1}(x_1, y_1) > f_{\sigma_1}(x_1, y_1)$, $f_{\sigma_1}(x_2, y_2) > f_{\sigma_2}(x_2, y_2)$, $f_{\sigma_1}(x_3, y_3) > f_{\sigma_2}(x_3, y_3)$, i.e.,

$$\sigma_2^{\theta + \sum \delta_i} \prod_{i=1}^d (x_{1i} - y_{1i})^{2\delta_i} e^{-\sigma_2 \|x_{1i} - y_{1i}\|^2} > \sigma_1^{\theta + \sum \delta_i} \prod_{i=1}^d (x_{1i} - y_{1i})^{2\delta_i} e^{-\sigma_1 \|x_{1i} - y_{1i}\|^2}$$

$$\sigma_2^{\theta + \sum \delta_i} \prod_{i=1}^d (x_{2i} - y_{2i})^{2\delta_i} e^{-\sigma_2 \|x_{2i} - y_{2i}\|^2} < \sigma_1^{\theta + \sum \delta_i} \prod_{i=1}^d (x_{2i} - y_{2i})^{2\delta_i} e^{-\sigma_1 \|x_{2i} - y_{2i}\|^2}$$

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

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

$$e^{-(\sigma_1 - \sigma_2) \|x_{2i} - y_{2i}\|^2} > \left( \frac{\sigma_2}{\sigma_1} \right)^{\theta + \sum \delta_i} e^{-(\sigma_1 - \sigma_2) \|x_{1i} - y_{1i}\|^2}$$

$$e^{-(\sigma_1 - \sigma_2) \|x_{3i} - y_{3i}\|^2} > \left( \frac{\sigma_2}{\sigma_1} \right)^{\theta + \sum \delta_i} e^{-(\sigma_1 - \sigma_2) \|x_{1i} - y_{1i}\|^2}$$

and therefore

$$(\sigma_1 - \sigma_2) (\|x_{2i} - y_{2i}\|^2 - \|x_{1i} - y_{1i}\|^2) < 0 \quad \text{and} \quad (\sigma_1 - \sigma_2) (\|x_{3i} - y_{3i}\|^2 - \|x_{1i} - y_{1i}\|^2) < 0,$$

(61) which by (P_{213}) implies $\sigma_1 > \sigma_2$ and $\sigma_1 < \sigma_2$ leading to a contradiction. Similarly, it can be shown that any 3-point set $S$ with property $(P_{ijk})$, $i \neq j \neq k \in \{1, 2, 3\}$ cannot be pseudo-shattered by $F$. Note that replacing (P_{213}) by

$$(E_{213}) : \quad \|x_{2i} - y_{2i}\| = \|x_{1i} - y_{1i}\| < \|x_{3i} - y_{3i}\|$$

also leads to a contradiction (by (61)) and so is the case for any $(E_{ijk})$, $i \neq j \neq k \in \{1, 2, 3\}$. This means that no 3-point set $S$ is shattered by $F$, contradicting the assumption that $VC(F) \geq 3$ and therefore $VC(F) \leq 2$.

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

**Proof of Theorem 2**: (a) Consider $K_{\alpha} := \{\partial^{\alpha, \alpha} \psi_\sigma(\cdot - \cdot) : \sigma \in \Sigma\}$ for $|\alpha| = r$. It can be shown that

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

where $H_l$ denotes the Hermite polynomial of degree $l$. By expanding $H_{2\alpha_i}$, we obtain

$$\partial^{\alpha, \alpha} \psi_\sigma(x - y) = \sigma^r \prod_{i=1}^d \sum_{j=0}^{\alpha_i} \eta_{ij} (\sigma(x_i - y_i)^2)^j e^{-\sigma(x_i - y_i)^2}$$

$$= \sum_{j_1=0}^{\alpha_1} \cdots \sum_{j_d=0}^{\alpha_d} \prod_{i=1}^d \eta_{ij_i} \sigma^{\alpha_i + j_i} (x_i - y_i)^{2j_i} e^{-\sigma(x_i - y_i)^2},$$

(62)

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

$$\sup_{\sigma \in \Sigma} \partial^{\alpha, \alpha} \psi_\sigma(x - y) \leq \sup_{\sigma \in \Sigma} \sigma^r \left( \sum_{j=0}^d \prod_{i=1}^d |\eta_{ij_i}| j_i^r e^{-\sigma_j} \right) = a^r \left( \sum_{j=0}^d \prod_{i=1}^d |\eta_{ij_i}| j_i^r e^{-\sigma_j} \right) < \infty,$$

(63)
which implies (iv) in Theorem 4 is satisfied, where \( \sum_{j=0}^{\alpha_i} = \sum_{j=1}^{\alpha_1} \cdots \sum_{j=d}^{\alpha_d} \). Now consider

\[
U_n(K_{\alpha}; (X_i)_{i=1}^n) := E_x \sup_{k^* \in K_{\alpha}} \left| \sum_{i < j} \varepsilon_i \varepsilon_j k^*(X_i, X_j) \right|
= E_x \sup_{\sigma \in \Sigma} \left( \sum_{p < q} \varepsilon_p \varepsilon_q \sup_{\alpha, \psi} (X_p, X_q) \right)
= E_x \sup_{\sigma \in \Sigma} \left( \sum_{p < q} \varepsilon_p \varepsilon_q \sup_{\alpha, \psi} (X_p, X_q) \right)
\]

By Lemma A.2, we have

\[
U_n(K_{\alpha}; (X_i)_{i=1}^n) := E_x \sup_{k^* \in K_{\alpha}} \left| \sum_{i < j} \varepsilon_i \varepsilon_j k^*(X_i, X_j) \right|
= E_x \sup_{\sigma \in \Sigma} \left( \sum_{p < q} \varepsilon_p \varepsilon_q \sup_{\alpha, \psi} (X_p, X_q) \right)
\]

Proposition 10 shows that \( K \) to be bounded above by \( \rho_i \) and therefore we used \( \alpha_j \) and \( i \) as an argument for \( T \) in [65]. Proposition [11] shows that \( K_{\alpha}^{j_1 \cdots j_d} \) is a VC-subgraph with VC-index, \( V := VC(K_{\alpha}^{j_1 \cdots j_d}) \leq 2 \) for any \( 0 \leq j_1 \leq \alpha_1, i = 1, \ldots, d \), which by [38] Theorem 2.6.7 implies that

\[
\mathcal{N}(K_{\alpha}^{j_1 \cdots j_d}, \rho_{j_1 \cdots j_d}, \varepsilon) \leq C'V(16\varepsilon)V \left( \frac{\zeta_{j_1 \cdots j_d}}{\varepsilon} \right)^{2(V-1)}, 0 < \varepsilon < \zeta_{j_1 \cdots j_d}
\]

for some universal constant, \( C' \), and therefore

\[
T \left( K_{\alpha}^{j_1 \cdots j_d}, \rho_{j_1 \cdots j_d}, \zeta_{j_1 \cdots j_d} \right) \leq \frac{C'_{j_1 \cdots j_d}}{n},
\]

where \( C'_{j_1 \cdots j_d} \) is a constant that depends on \( C' \), \( V \) and \( \zeta_{j_1 \cdots j_d} \). Combining [65] and [67] in [64], we obtain

\[
U_n(K_{\alpha}; (X_i)_{i=1}^n) \leq n \sum_{j_1=0}^{\alpha_1} \cdots \sum_{j_d=0}^{\alpha_d} \left( \prod_{i=1}^{d} |\eta_{j_i}| \right) \left( 2C''_{j_1 \cdots j_d} + \frac{\zeta_{j_1 \cdots j_d}}{\sqrt{2}} \right) = O_p(n).
\]

Also, since \( K \) is a VC-subgraph with \( VC(K) = 1 \), from [38] we obtain \( \mathcal{N}(K, \rho, \varepsilon) \) is a constant independent of \( \varepsilon \). Following the analysis as above, it is easy to show that \( U_n(K_{\alpha}; (X_i)_{i=1}^n) = O_p(n) \).
(b) Since $\vartheta^{\alpha, \alpha} \int_0^\infty \psi_\alpha(x - y) \, d\Lambda(\sigma) = \int_0^\infty \vartheta^{\alpha, \alpha} \psi_\alpha(x - y) \, d\Lambda(\sigma)$ holds by [12, Theorem 2.27(b)], define

$$ K_\alpha := \left\{ \int_0^\infty \vartheta^{\alpha, \alpha} \psi_\alpha(x - y) \, d\Lambda(\sigma), \ x, y \in \mathbb{R}^d : \Lambda \in \mathcal{M}_a \right\}. $$

Therefore

$$ \sup_{k' \in K_\alpha, x, y \in \mathbb{R}^d} k'(x, y) = \sup_{\Lambda \in \mathcal{M}_a, x, y \in \mathbb{R}^d} \int_0^\infty \vartheta^{\alpha, \alpha} \psi_\alpha(x - y) \, d\Lambda(\sigma) \leq \left( \sum_{j=0}^\alpha \prod_{i=1}^d |\eta_{ij}| j^i e^{-j} \right) \sup_{\Lambda \in \mathcal{M}_a} \int_0^\infty \sigma^r \, d\Lambda(\sigma) = A \left( \sum_{j=0}^\alpha \prod_{i=1}^d |\eta_{ij}| j^i e^{-j} \right) < \infty, $$

and so $K$ satisfies (iv) in Theorem [11]. Now consider

$$ U_n(K; (X_i)_{i=1}^n) := \mathbb{E}_x \sup_{k' \in K} \left| \sum_{p < q} \varepsilon_p \varepsilon_q k'(X_p, X_q) \right| = \mathbb{E}_x \sup_{\Lambda \in \mathcal{M}_a} \left| \sum_{p < q} \varepsilon_p \varepsilon_q \int_0^\infty \psi_\alpha(X_p - X_q) \, d\Lambda(\sigma) \right| \leq \mathbb{E}_x \sup_{\sigma \in (0, \infty)} \left| \sum_{p < q} \varepsilon_p \varepsilon_q \psi_\alpha(X_p - X_q) \right|. \quad (68) $$

By Proposition [10] since $\{\psi_\alpha(x - y) : \sigma \in (0, \infty)\}$ is a VC-subgraph, carrying out the analysis (following (68)) in (a), we obtain $U_n(K; (X_i)_{i=1}^n) = O_p(n)$. Also,

$$ U_n(K_\alpha; (X_i)_{i=1}^n) := \mathbb{E}_x \sup_{k' \in K_\alpha} \left| \sum_{p < q} \varepsilon_p \varepsilon_q k'(X_p, X_q) \right| = \mathbb{E}_x \sup_{\Lambda \in \mathcal{M}_a} \left| \sum_{p < q} \varepsilon_p \varepsilon_q \int_0^\infty \vartheta^{\alpha, \alpha} \psi_\alpha(X_p - X_q) \, d\Lambda(\sigma) \right| \leq \mathbb{E}_x \sup_{\Lambda \in \mathcal{M}_a} \int_0^\infty \left| \sum_{p < q} \varepsilon_p \varepsilon_q \vartheta^{\alpha, \alpha} \psi_\alpha(X_p - X_q) \right| \, d\Lambda(\sigma) \leq A \mathbb{E}_x \sup_{\sigma \in (0, \infty)} \left| \sum_{p < q} \varepsilon_p \varepsilon_q \sigma^{-r} \vartheta^{\alpha, \alpha} \psi_\alpha(X_p - X_q) \right| =: A U_n(\mathcal{L}; (X_i)_{i=1}^n), $$

where

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

Replicating the analysis in [144] for $U_n(\mathcal{L}; (X_i)_{i=1}^n)$ in conjunction with Proposition [10] it is easy to show that $U_n(\mathcal{L}; (X_i)_{i=1}^n) = O_p(n)$ and therefore $U_n(K_\alpha; (X_i)_{i=1}^n) = O_p(n)$.

(c) It is easy to check that any $k \in K$ 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

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

and

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

27
which implies $\mathcal{K}$ satisfies (iv) in Theorem 4. Now consider

$$U_n(\mathcal{K}; (X_i)_{i=1}^n) := \mathbb{E}_x \sup_{k \in \mathcal{K}} \left| \sum_{p<q} n \epsilon_p \epsilon_q k(X_p, X_q) \right| = \mathbb{E}_x \sup_{\Lambda \in \mathcal{Q}_K} \left| \sum_{p<q} n \epsilon_p \epsilon_q \int e^{-(X_p - X_q)^T \Delta (X_p - X_q)} d\Lambda(\Delta) \right|$$

$$\leq \mathbb{E}_x \sup_{\text{diag}(\Delta) \in (0, \infty)^d} \left| \sum_{p<q} n \epsilon_p \epsilon_q \epsilon^{-(X_p - X_q)^T \Delta (X_p - X_q)} \right| =: U_n(\mathcal{J}; (X_i)_{i=1}^n),$$

where

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

Define

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

It is easy to check that for any $\tilde{k}_1, \tilde{k}_2 \in \mathcal{J}$, $\rho(\tilde{k}_1, \tilde{k}_2) \leq \sqrt{d} \sum_{i=1}^d \rho(\tilde{k}_1^i, \tilde{k}_2^i)$, 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 10 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

$$U_n(\mathcal{K}; (X_i)_{i=1}^n) \leq U_n(\mathcal{J}; (X_i)_{i=1}^n) = O_P(n).$$

Similarly,

$$U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n) := \mathbb{E}_x \sup_{k' \in \mathcal{K}_\alpha} \left| \sum_{p<q} n \epsilon_p \epsilon_q k'(X_p, X_q) \right|$$

$$= \mathbb{E}_x \sup_{\Lambda_i \in \mathcal{M}_{\alpha_i}, \epsilon_i \in [d]} \left| \sum_{p<q} n \epsilon_p \epsilon_q \prod_{i=1}^d \int_0^\infty \partial^{\alpha_i, \alpha_i} \psi_{\sigma_i}(X_{pi} - X_{qi}) d\Lambda_i(\sigma_i) \right|$$

$$= \mathbb{E}_x \sup_{\Lambda_i \in \mathcal{M}_{\alpha_i}, \epsilon_i \in [d]} \left| \int_0^\infty \cdots \int_0^\infty \sum_{p<q} n \epsilon_p \epsilon_q \prod_{i=1}^d \partial^{\alpha_i, \alpha_i} \psi_{\sigma_i}(X_{pi} - X_{qi}) \prod_{i=1}^d d\Lambda_i(\sigma_i) \right|$$

$$\leq \left( \prod_{i=1}^d A_i \right) \mathbb{E}_x \sup_{\text{diag}(\Delta) \in (0, \infty)^d} \sum_{p<q} n \epsilon_p \epsilon_q \prod_{i=1}^d \sigma_i^{-\alpha_i} \partial^{\alpha_i, \alpha_i} \psi_{\sigma_i}(X_{pi} - X_{qi})$$

$$=: \left( \prod_{i=1}^d A_i \right) U_n(\mathcal{I}; (X_i)_{i=1}^n),$$

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

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

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

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

and noting that for any $\tilde{k}_1, \tilde{k}_2 \in \mathcal{I}_i$, we have $\rho(\tilde{k}_1, \tilde{k}_2) \leq B d^{1/2} \rho(\tilde{k}_1^i, \tilde{k}_2^i)$ where $\tilde{k}_1^i, \tilde{k}_2^i \in \mathcal{I}_i$, $B := \max_{i \in \{1, \ldots, d\}} \sum_{j=0}^d |\eta_{ij}| J^2 e^{-j}$ 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 $VC(\mathcal{I}_i) \leq 2$ for any $i = 1, \ldots, d$ and therefore $\mathcal{N}(\mathcal{I}, \rho, \epsilon) = O(\epsilon^{-2})$, which means

$$U_n(\mathcal{K}_\alpha; (X_i)_{i=1}^n) \leq \left( \prod_{i=1}^d A_i \right) U_n(\mathcal{I}; (X_i)_{i=1}^n) = O_P(n).$$

28
(d) First we derive an alternate form for $k \in \mathcal{K}$ which will be useful to prove the result. To this end, by [39 Theorem 6.13], any $k \in \mathcal{K}$ can be written as the Fourier transform of $\frac{Ac^{2d-4}\Gamma(\beta)}{2^d(21-\beta)}(c^2 + \|\omega\|^2)_{2-\beta}^{-\beta}$, i.e., for any $c > 0$,

$$k(x, y) = A \frac{\|x - y\|_{2-\beta}^{-\beta}}{c^2} \int_{\mathbb{R}^d} e^{-\sqrt{\beta}(x-y)^T \omega} e^{-\|\omega\|^2_{2-\beta}} d\omega.$$  

(70)

By the Schönberg representation for radial kernels (see [5]), it follows from [39 Theorem 7.15] that

$$(c^2 + \|\omega\|^2_{2-\beta})^{-\beta} = \frac{1}{\Gamma(\beta)} \int_0^\infty e^{-t\|\omega\|^2_{2-\beta}} t^{\beta-1} e^{-c^2 t} dt.$$  

(71)

Combining (70) and (71), we have

$$k(x, y) = \frac{Ac^{2d-4}\Gamma(\beta)}{2^d(21-\beta)} \int_{\mathbb{R}^d} e^{-\sqrt{\beta}(x-y)^T \omega} \int_0^\infty e^{-t\|\omega\|^2_{2-\beta}} t^{\beta-1} e^{-c^2 t} dt d\omega,$$

which after applying Fubini’s theorem yields

$$k(x, y) = \frac{Ac^{2d-4}\Gamma(\beta)}{2^d(21-\beta)} \int_0^\infty \int_{\mathbb{R}^d} e^{-\sqrt{\beta}(x-y)^T \omega} e^{-t\|\omega\|^2_{2-\beta}} d\omega t^{\beta-1} e^{-c^2 t} dt$$

$$= \frac{c^{2d-4}\Gamma(\beta-d/2)}{\Gamma(\beta-\frac{d}{2})} \int_0^\infty e^{-\frac{1}{4t} \|x-y\|^2_{2-\beta}} t^{\beta-1} e^{-c^2 t} dt.$$  

(72)

Note that

$$\sup_{k \in \mathcal{K}, x, y \in \mathcal{X}} k(x, y) \leq \sup_{c \in (0, a)} \frac{c^{2d-4}\Gamma(\beta-d/2)}{\Gamma(\beta-\frac{d}{2})} \int_0^\infty t^{\beta-1} e^{-\frac{1}{4t} \|x-y\|^2_{2-\beta}} dt = 1,$$

implying that $\mathcal{K}$ satisfies (iii) in Theorem 3. Using (72) in $U_n(K; (X_i)^{n}_{i=1})$, we have

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

$$= E_{\varepsilon} \sup_{c \in (0, a)} \frac{c^{2d-4}\Gamma(\beta-\frac{d}{2})}{\Gamma(\beta-\frac{d}{2})} \left| \sum_{i,j} \varepsilon_i \varepsilon_j \int_0^\infty e^{-\frac{\|X_i-X_j\|^2_{2-\beta}}{4t}} t^{\beta-1/4} e^{-c^2 t} dt \right|$$

$$\leq E_{\varepsilon} \sup_{t \in (0, \infty)} \left| \sum_{i,j} \varepsilon_i \varepsilon_j e^{-\frac{\|X_i-X_j\|^2_{2-\beta}}{4t}} \right| \sup_{c \in (0, a)} \frac{c^{2d-4}\Gamma(\beta-\frac{d}{2})}{\Gamma(\beta-\frac{d}{2})} \left| \int_0^\infty t^{\beta-1/4} e^{-c^2 t} dt \right|$$

$$= E_{\varepsilon} \sup_{\sigma \in (0, \infty)} \left| \sum_{i,j} \varepsilon_i \varepsilon_j e^{-\sigma \|X_i-X_j\|^2_{2-\beta}} \right|,$$

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

$$k'(x, y) := \partial^{\alpha\alpha} k(x, y) = \frac{c^{2d-4}\Gamma(\beta-d/2)}{\Gamma(\beta-\frac{d}{2})} \int_0^\infty \left( \partial^{\alpha\alpha} e^{-\frac{|x-y|^2_{2-\beta}}{4t}} \right) t^{\beta-1/4} e^{-c^2 t} dt$$

$$= \frac{c^{2d-4}\Gamma(\beta-d/2)}{\Gamma(\beta-\frac{d}{2})} \int_0^\infty (At)^{m \wedge r} \partial^{\alpha\alpha} e^{-\frac{|x-y|^2_{2-\beta}}{4t}} t^{\beta-1/4} e^{-c^2 t} dt$$

$$= \frac{c^{2d-4}\Gamma(\beta-d/2)}{\Gamma(\beta-\frac{d}{2})} (At)^{m \wedge r} \partial^{\alpha\alpha} e^{-\frac{|x-y|^2_{2-\beta}}{4t}} t^{\beta-1/4} e^{-c^2 t} dt,$$

(73)

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

$$\sup_{k' \in \mathcal{K}, x, y \in \mathcal{X}} k'(x, y) \leq \sup_{\sigma \in (0, \infty), x, y \in \mathcal{X}} \left| \sigma^{-(m \wedge r)} \partial^{\alpha\alpha} e^{-\sigma \|x-y\|^2_{2-\beta}} \right| \frac{\Gamma(\beta-\frac{d}{2} - m \wedge r)}{\Gamma(\beta-\frac{d}{2})} < \infty,$$
therefore satisfying (iv) in Theorem 4. Using (73) we now obtain a bound on  

\[ U_n(K; (X_i)_{i=1}^n) = \mathbb{E}_x \sup_{K \in K_n} \left| \sum_{i<j}^n \varepsilon_i \varepsilon_j k'(X_i, X_j) \right| \]

\[ = \mathbb{E}_x \sup_{c \in (0, a]} \frac{c^{2\beta - d}}{B} \sum_{i<j}^n \varepsilon_i \varepsilon_j \int_0^\infty \left( 4t \right)^{m \wedge r} \partial^{\alpha, \alpha} e^{-\|X_i - X_j\|^2_{4t}} \left. \right\|_{\beta-1-\frac{d}{2}-(m \wedge r)} e^{-c^2 t} dt \]

\[ \leq \mathbb{E}_x \sup_{c \in (0, a]} \frac{c^{2\beta - d}}{B} \sum_{i<j}^n \varepsilon_i \varepsilon_j \int_0^\infty t^{\beta-1-\frac{d}{2}-(m \wedge r)} e^{-c^2 t} dt \]

\[ \leq \frac{\Gamma(\beta - \frac{d}{2} - m \wedge r) a^{2(m \wedge r)}}{\Gamma(\beta - \frac{d}{2}) 4^{m \wedge r}} \mathbb{E}_x \sup_{\sigma \in (0, \infty)} \left| \sum_{i<j}^n \varepsilon_i \varepsilon_j \sigma^{-(m \wedge r)} \partial^{\alpha, \alpha} e^{-\sigma \|X_i - X_j\|^2} \right| , \]

and therefore it follows from the proof of Theorem 3(ii) that  

\[ U_n(K; (X_i)_{i=1}^n) = O_P(n) . \]

Using the bounds on  

\[ U_n(K; (X_i)_{i=1}^n) \]  

and  

\[ U_n(K; (X_i)_{i=1}^n) \]  

in (73) and (79) respectively and following the proofs of Theorems 3 and 4 yields the desired result.

**Remark 6.** Note that instead of following the indirect route—showing  

\[ K_0 \wedge \ldots \wedge K_d \]  

to be a VC-subgraph and then bounding  

\[ U_n(K; (X_i)_{i=1}^n) \]  

of showing the result in Theorem 4 for the Gaussian kernel family as presented in (a), one can directly get the result by obtaining a bound on  

\[ K \]  

under the assumption that  

\[ X = (a_0, b_0) \]  

for some  

\[ -\infty < a_0 < b_0 < \infty . \]

The advantage with the analysis in (a) is that the result holds for  

\[ X \in \mathbb{R}^d \]  

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  

\[ K \]  

and  

\[ K_\alpha \]  

for kernel classes in (b) and (c) while allowing to handle these classes easily through (a).

### 6.4 Proof of the claim in Remark 4(iii)

We show that  

\[ K \in (a)-(c) \]

satisfy the conditions in Theorem 2 and therefore metrize the weak topology on  

\[ M_2^1(\mathbb{R}^d) \]  

Note that the families in (a)-(c) are uniformly bounded and every  

\[ k'(-, x) \in C_0(\mathbb{R}^d) \]  

for all  

\[ x \in \mathbb{R}^d \].

It therefore remains to check (14) and (P) in Theorem 2.

By Proposition 5 in (82) (see (17) in its proof), it is clear that (14) is satisfied for  

\[ K \]  

in (a) and (b). For (c), define

\[ B := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(x, y) \, d\mu(x) \, d\mu(y) \]

and consider

\[ B = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \prod_{j=1}^d \int_0^\infty e^{-\sigma (x_j - y_j)^2} d\Lambda_j(\sigma) \, d\mu(x) \, d\mu(y) \]

\[ = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \prod_{j=1}^d \int_0^\infty \frac{1}{(4\pi \sigma)^{d/2}} \int_{\mathbb{R}} e^{-\sqrt{1} \omega_j(x_j - y_j)} e^{-\omega_j^2 \sigma \frac{(4\pi \sigma)^{d/2}}{\pi^d}} \, d\omega_j \, d\Lambda_j(\sigma) \, d\mu(x) \, d\mu(y) \]

\[ = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-\sqrt{1} \omega^2 (x - y)} \prod_{j=1}^d \int_0^\infty \frac{1}{(4\pi \sigma)^{d/2}} e^{-\omega_j^2 \sigma \frac{(4\pi \sigma)^{d/2}}{\pi^d}} \, d\Lambda_j(\sigma) \, d\omega \, d\mu(x) \, d\mu(y) \]

\[ = \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2 \left( \prod_{j=1}^d \int_0^\infty \frac{1}{(4\pi \sigma)^{d/2}} e^{-\omega_j^2 \sigma \frac{(4\pi \sigma)^{d/2}}{\pi^d}} \, d\Lambda_j(\sigma) \right) \, d\omega , \]  

(74)
where we have invoked Fubini’s theorem in the last two lines of (49) and denotes the Fourier transform of μ. Since supp(Λ_j) \neq \{0\} for all j = 1, \ldots, d, the inner integrals in (49) are positive for every \omega_j \in \mathbb{R} and so (44) holds.

We now show that (P) in Theorem 2 is satisfied by \mathcal{K} in (a)–(c). Consider \mathcal{K} in (b). Fix x \in \mathbb{R}^d and \epsilon > 0. Define \text{U}_{x,\epsilon} = \{y \in \mathbb{R}^d : \|x - y\| < (4 \delta_1 \log \frac{2n^d}{2\epsilon})^{1/4}\}, where \delta and B are as mentioned in the statement of Theorem 5. Then for any \epsilon > 0. Define \text{U}_{x,\epsilon} := \{y \in \mathbb{R}^d : \|x - y\| < (4 \min_i \delta_i \log \frac{2n^d}{2\epsilon})^{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 \text{U}_{x,\epsilon},

\[ \|k(\cdot, x) - k(\cdot, y)\|^2_{\mathcal{H}_k} = 2 - 2 \prod_{i=1}^d \int_0^\infty e^{-\sigma (x_i - y_i)^2} d\Lambda_i(\sigma) \]

\[ \leq 2 - 2 \prod_{i=1}^d \left( \inf_{\Lambda \in \mathcal{M}_d} \int_0^\infty e^{-\delta_i \sigma^2} d\Lambda_i(\sigma) \right) \left( \inf_{\sigma \in (0, \infty)} e^{-\sigma (x_i - y_i)^2} e^{\delta_i \sigma^2} \right) \]

\[ \leq 2 - 2 \prod_{i=1}^d B_i e^{-\frac{(x_i - y_i)^4}{4 \min_i \delta_i}} \]  

\[ \leq 2 - 2 \prod_{i=1}^d B_i e^{-\frac{\|x - y\|^4}{4 \min_i \delta_i}} < \epsilon^2, \]
	herby proving the result.

### 6.5 Proof of Theorem 6

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 5. 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).

**Proof**

**D** Since k is continuous and bounded and \mathcal{X} is separable, by 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 \text{id} : \mathcal{H}_k \rightarrow C_b(\mathcal{X}) is linear and continuous Lemma 4.28. Therefore, by Theorem 1.1, \mathcal{F}_H = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\} is \mathbb{P}--Donsker, i.e., \sqrt{n} (\mathbb{P}_n * K_h - \mathbb{P}) \overset{\mathcal{L}}{\rightarrow} N(0, \mathbb{I}) \mathbb{G}_\mathbb{P}. Also, \sqrt{n} (\mathbb{P}_n * K_h - \mathbb{P}) \overset{\text{weak}}{\rightarrow} \mathcal{F}_H \mathbb{G}_\mathbb{P} by Slutsky’s lemma and Theorem 4.

**A-C** From Theorem 4, we have

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

Using the argument as in (D), it is easy to verify that \mathcal{H}_b is separable and \text{id} : \mathcal{H}_b \rightarrow C_b(\mathcal{X}) is linear and continuous and therefore \mathcal{B} is \mathbb{P}--Donsker, which implies \mathcal{F}_H is Donsker by Theorem 2.10.1. The result therefore follows using Slutsky’s lemma and Theorem 3. The proof of (C) is similar to that of in (A) but we use (44) instead of (45). For (B), the result hinges on a relation to similar those in (45) and (46), which we derive below. Let \mathcal{K} be the kernel family as shown in (45). Then for k \in \mathcal{K}, let \mathcal{H}_c be the induced RKHS. From Theorems 6.13 and 10.12, it follows that for any f \in \mathcal{H}_c,

\[ \|f\|^2_{\mathcal{H}_c} = \frac{\Gamma(\beta)}{2^{1-\beta}} \int |\hat{f}(\omega)|^2 \frac{e^{-d}}{(c\|\omega\|_2)^{\beta-\frac{d}{2}}} d\mathcal{R}_\mathbb{R}^d - \beta(c\|\omega\|_2) \]

31
By Corollary 5.12, since for every $\nu \in \mathbb{R}, x \mapsto x^\nu \mathcal{R}_x(x)$ is non-increasing on $(0, \infty)$, we have that for any $0 < \tau < c < \infty$,
\[ \|f\|_{\mathcal{H}_\tau} \leq \left( \frac{c}{\tau} \right)^{\frac{1}{2}} \|f\|_{\mathcal{H}_c} \]
and so $\mathcal{H}_c \subset \mathcal{H}_\tau$. Therefore, we have
\[ \mathcal{F}_H \subset \bigcup_{c \in [a, \infty)} \left\{ f \in \mathcal{H}_a : \|f\|_{\mathcal{H}_a} \leq \left( \frac{c}{a} \right)^{d/2} \right\} = \mathcal{H}_a. \]
For the choice of $\mathcal{K}$ in Theorem (b), we have
\[ \mathcal{F}_H \subset \bigcup_{c \in [a, b]} \left\{ f \in \mathcal{H}_a : \|f\|_{\mathcal{H}_a} \leq \left( \frac{c}{a} \right)^{d/2} \right\} = \left\{ f \in \mathcal{H}_a : \|f\|_{\mathcal{H}_a} \leq \left( \frac{b}{a} \right)^{d/2} \right\} \quad (75) \]
and the rest follows.

### 6.6 Proof of Proposition 7

By definition,
\[ \|\mathcal{P} - \mathcal{Q}\|_{\mathcal{K}_x} = \sup_{k \in \mathcal{K}, x \in \mathcal{X}} \left| \int k(x, y) d(\mathcal{P} - \mathcal{Q})(y) \right| = \sup_{k \in \mathcal{K}, x \in \mathcal{X}} \left| \int \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} d(\mathcal{P} - \mathcal{Q})(y) \right|. \]
Since $\mathcal{K}$ is uniformly bounded, $k(\cdot, x)$ is Bochner-integrable for all $k \in \mathcal{K}$ and $x \in \mathcal{X}$, i.e.,
\[ \int \|k(\cdot, x)\|_{\mathcal{H}_k} d(\mathcal{P})(x) = \int \sqrt{k(x, x)} d\mathcal{P}(x) \leq \sqrt{\nu}, \forall k \in \mathcal{K}, x \in \mathcal{X}, \]
and therefore
\[ \|\mathcal{P} - \mathcal{Q}\|_{\mathcal{K}_x} = \sup_{k \in \mathcal{K}, x \in \mathcal{X}} \left| \int k(y, x) d(\mathcal{P} - \mathcal{Q})(y) \right| = \sup_{k \in \mathcal{K}, x \in \mathcal{X}} \left| \langle k(\cdot, x), \int k(\cdot, y) d(\mathcal{P} - \mathcal{Q})(y) \rangle_{\mathcal{H}_k} \right| \leq \sup_{k \in \mathcal{K}, x \in \mathcal{X}} \|k(\cdot, x)\|_{\mathcal{H}_k} \mathcal{D}_k(\mathcal{P}, \mathcal{Q}) \leq \sqrt{\nu} \|\mathcal{P} - \mathcal{Q}\|_{\mathcal{F}_H}, \quad (76) \]
which proves the lower bound on $\|\mathcal{P} - \mathcal{Q}\|_{\mathcal{F}_H}$ in (51). To prove the upper bound, consider
\[ \|\mathcal{P} - \mathcal{Q}\|_{\mathcal{F}_H}^{\downarrow} \leq \sup_{k \in \mathcal{K}} \int \left| \int k(x, y) d(\mathcal{P} - \mathcal{Q})(x) \right| \mathcal{D}_k(\mathcal{P}, \mathcal{Q}) d\mathcal{P}(y) \]
\[ \leq \sup_{k \in \mathcal{K}} \int \left| \int k(x, y) d(\mathcal{P} - \mathcal{Q})(y) \right| d\mathcal{P}(x) \]
\[ \leq 2 \sup_{k \in \mathcal{K}} \sup_{x \in \mathcal{X}} \left| \int k(x, y) d(\mathcal{P} - \mathcal{Q})(y) \right| = 2 \|\mathcal{P} - \mathcal{Q}\|_{\mathcal{K}_x}, \]
thereby proving the result in (51). (52) simply follows from Theorem 2 and (51).

### 6.7 Proof of Theorem 8

In order to prove Theorem 8, we need a lemma (see Lemma 11 below) which is based on the notion of fat-shattering dimension (see Definition 11.10), which we define 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 $S = \{z_1, \ldots, z_n\} \subset \mathcal{X}$ is said to be $\epsilon$-shattered by $\mathcal{F}$ if there exists real numbers $r_1, \ldots, r_n$ such that for each $b \in \{0, 1\}^n$ there is a function $f_b \in \mathcal{F}$ with $f_b(z_i) \geq r_i + \epsilon$ if $b_i = 1$ and $f_b(z_i) \leq r_i - \epsilon$ if $b_i = 0$, for $1 \leq i \leq n$. The fat-shattering dimension of $\mathcal{F}$ is defined as
\[ \text{fat}_\epsilon(\mathcal{F}) = \sup \left\{ |S| : S \subset \mathcal{X}, S \text{ is } \epsilon \text{-shattered by } \mathcal{F} \right\}. \]
Lemma 11. Define
\[ \mathcal{G} := \left\{ e^{-\sigma(-x)^2} : \sigma \in (0, \infty), x \in \mathbb{R} \right\}. \]
Then \( \text{fat}_\varepsilon(\mathcal{G}) \leq 1 + \lfloor \varepsilon^{-1} \rfloor \). In addition, there exists a universal constant \( c' \) such that for every empirical measure \( \mathbb{P}_n \), and every \( 0 < \varepsilon \leq 1 \),
\[ \log \mathcal{N}(\mathcal{G}, L^2(\mathbb{P}_n), \varepsilon) \leq c' \left( 1 + \frac{8}{\varepsilon} \right) \log^2 \left( \frac{2}{\varepsilon} + \frac{16}{\varepsilon^2} \right). \]

Proof. Since \( \int_{-\infty}^{\infty} \frac{d\theta}{\sigma^2} \) \( dy = 2 < \infty \) 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 [1] Theorem 11.12, we obtain \( \text{fat}_\varepsilon(\mathcal{G}) \leq 1 + \lfloor \varepsilon^{-1} \rfloor \) and [25] Theorem 3.2 ensures that there exists a universal constant \( c' \) such that for every empirical measure \( \mathbb{P}_n \), and every \( \varepsilon > 0 \),
\[ \log \mathcal{N}(\mathcal{G}, L^2(\mathbb{P}_n), \varepsilon) \leq c' \text{fat}_\varepsilon(\mathcal{G}) \log^2 \left( \frac{2 \text{fat}_\varepsilon(\mathcal{G})}{\varepsilon} \right) \leq c' \left( 1 + \frac{8}{\varepsilon} \right) \log^2 \left( \frac{2}{\varepsilon} + \frac{16}{\varepsilon^2} \right), \]
thereby yielding the result. \( \square \)

Proof of Theorem 38. (a) Define \( \mathcal{F}_i := \left\{ e^{-\sigma_i(-x)^2} : \sigma_i \in (0, \infty), x_i \in \mathbb{R} \right\}, \ i = 1, \ldots, d \).

By Lemma [11] it is easy to see that there exists \( N_i(\varepsilon) := \mathcal{N}(\mathcal{F}_i, L^2(\mathbb{P}_n), \varepsilon) \) functions
\[ \left\{ e^{-\sigma_i(-x_i)^2}, \ldots, e^{-\sigma_i N_i(\varepsilon)(-x_i N_i(\varepsilon))} \right\} \subset \mathcal{F}_i \]
such that for any \( \varepsilon > 0 \) and \( f \in \mathcal{F}_i \), there exists \( l \in \{1, \ldots, N_i(\varepsilon)\} \) such that
\[ \left\| f - e^{-\sigma_i(-x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \leq \varepsilon. \]

Now pick \( l_i \in \{1, \ldots, N_i(\varepsilon)\}, i = 1, \ldots, d \). Then for \( k(\cdot, x) = e^{-\sigma \|\cdot\|^2}, \) we have
\[
\left\| e^{-\sigma \|\cdot\|^2} - \prod_{i=1}^d e^{-\sigma_i(-x_i)^2} \right\|_{L^2(\mathbb{P}_n)} = \left\| \prod_{i=1}^d e^{-\sigma(-x_i)^2} - \prod_{i=1}^d e^{-\sigma_i(-x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \\
\leq \left\| \sum_{i=1}^d e^{-\sigma(-x_i)^2} - e^{-\sigma_i(-x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \\
\leq \sum_{i=1}^d \left\| e^{-\sigma(-x_i)^2} - e^{-\sigma_i(-x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \leq \varepsilon d.
\]

This implies \( \mathcal{N}(\mathcal{K}_X, L^2(\mathbb{P}_n), \varepsilon d) = \prod_{i=1}^d N_i(\varepsilon) \) and therefore,
\[ \log \mathcal{N}(\mathcal{K}_X, L^2(\mathbb{P}_n), \varepsilon d) = \sum_{i=1}^d \log N_i(\varepsilon), \]
which by Lemma [11] yields
\[ \sup \sup \log \mathcal{N}(\mathcal{K}_X, L^2(\mathbb{P}_n), \varepsilon) \leq c' d \left( 1 + \frac{8d}{\varepsilon} \right) \log^2 \left( \frac{2d}{\varepsilon} + \frac{16d^2}{\varepsilon^2} \right), \ 0 < \varepsilon \leq 1. \]

It is easy to verify that \( \int_{0}^{\infty} \sup_{\mathbb{P}_n} \log \mathcal{N}(\mathcal{K}_X, L^2(\mathbb{P}_n), \varepsilon) < \infty \). Therefore \( \mathcal{K}_X \) is a universal Donsker class and the UCLT’s follow.
(b) Following the setting in (a) above, for \( k(\cdot, x) = \int_0^\infty e^{-\sigma \|x\|^2} d\Lambda(\sigma), \ A \in \mathcal{M}_A \), we have
\[
\left\| k(\cdot, x) - \prod_{i=1}^d e^{-\sigma_{i,i}(-x_i,x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \leq \int_0^\infty \left\| e^{-\sigma \|x\|^2} - \prod_{i=1}^d e^{-\sigma_{i,i}(-x_i,x_i)^2} \right\|_{L^2(\mathbb{P}_n)} d\Lambda(\sigma) \leq \epsilon d,
\]
and the claim as in (a) follows.

(c) The idea is similar to that of in (b) where for \( k(\cdot, x) = \prod_{i=1}^d \int_0^\infty e^{-\sigma(-x_i)^2} d\Lambda_i(\sigma), \ A_i \in \mathcal{M}_{A_i} \), we have
\[
\left\| k(\cdot, x) - \prod_{i=1}^d e^{-\sigma_{i,i}(-x_i,x_i)^2} \right\|_{L^2(\mathbb{P}_n)} \leq \sum_{i=1}^d \int_0^\infty \left\| e^{-\sigma(-x_i)^2} d\Lambda_i(\sigma) - e^{-\sigma_{i,i}(-x_i,x_i)^2} \right\|_{L^2(\mathbb{P}_n)} d\Lambda_i(\sigma) \leq \epsilon d,
\]
and the claim as in (a) follows.

(d) From \([72]\), we have
\[
k(x, y) = \frac{(e^2/4)^{\beta-\frac{d}{2}}}{\Gamma(\beta - \frac{d}{2})} \int_0^\infty e^{-\sigma \|x-y\|^2} \sigma^{\frac{d}{2} - \beta - 1} e^{-\frac{x^2}{2\sigma}} d\sigma,
\]
which is of the form in (b) where \( d\Lambda(\sigma) = \frac{(e^2/4)^{\beta-\frac{d}{2}}}{\Gamma(\beta - \frac{d}{2})} \sigma^{\frac{d}{2} - \beta - 1} e^{-\frac{x^2}{2\sigma}} d\sigma \) and the result follows from (b).

## A Supplementary Results

In the following, we present supplementary results that are used in the proofs of Theorems\([8,4]\) and\([4]\). 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 \([7]\) 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 some finite subset of \( \mathbb{R}^{\frac{d(d+1)}{2}} \) and \( (\varepsilon_i)_{i=1}^l \) be independent Rademacher variables. For any \( a \in \mathcal{A} \), define \( a := (a_{ij})_{1 \leq i \leq j \leq n} \). Suppose \( \sup_{a \in \mathcal{A}} \|a\|_2 \leq R < \infty \), then for any \( 0 < \theta < 1 \),
\[
\mathbb{E} \sup_{a \in \mathcal{A}} \left| \sum_{1 \leq i < j \leq n} \varepsilon_i \varepsilon_j a_{ij} \right| \leq \frac{eR}{\theta} \log \frac{|\mathcal{A}|}{1 - \theta} \tag{A.1}
\]
and therefore
\[
\mathbb{E} \sup_{a \in \mathcal{A}} \left| \sum_{1 \leq i < j \leq n} \varepsilon_i \varepsilon_j a_{ij} \right| \leq eR \left( 1 + \sqrt{\log |\mathcal{A}|} \right)^2. \tag{A.2}
\]
Proof. For \( \lambda > 0 \), consider

\[
e^{\lambda E \sup_{a \in A} |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|} \leq E e^{\lambda \sup_{a \in A} |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|} = E \sup_{a \in A} e^{\lambda |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|} \leq \sum_{a \in A} E e^{\lambda |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|} = \sum_{a \in A} \sum_{c=0}^{\infty} \frac{\lambda^c |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|^c}{c!}.
\]

By the hypercontractivity of homogeneous Rademacher chaos of degree 2 [7, Theorem 3.2.2], we have

\[
E \left| \sum_{i<j} \varepsilon_i \varepsilon_j a_{ij} \right|^c \leq (c-1)^c \left( E \left| \sum_{i<j} \varepsilon_i \varepsilon_j a_{ij} \right| \right)^{2c/2} \leq (c-1)^c \left( \sum_{i<j} a_{ij}^2 \right)^{c/2}, \quad c \geq 2
\]

and

\[
E \left| \sum_{i<j} \varepsilon_i \varepsilon_j a_{ij} \right| \leq \left( E \left| \sum_{i<j} \varepsilon_i \varepsilon_j a_{ij} \right| ^{2c/2} \right)^{1/2} = \left( \sum_{i<j} a_{ij}^2 \right)^{1/2}
\]

which implies

\[
e^{\lambda E \sup_{a \in A} |\sum_{i<j} \varepsilon_i \varepsilon_j a_{ij}|} \leq \sum_{a \in A} \sum_{c=0}^{\infty} \frac{\lambda^c \|a\|^2}{c!}.
\]

Using \( c^c/c! \leq e^c \) and choosing \( \lambda = \frac{\theta}{\sqrt{4n}} \) for some \( 0 < \theta < 1 \), we obtain the desired result in (A.1). Using 

\[-\log((1-\theta)/(1+\theta)) \quad \text{for } 0 < \theta < 1 \]

in (A.1) and taking infinite over \( \theta \in (0,1) \) (where the infinite is obtained at \( \theta = \sqrt{\log |A|/(1 + \sqrt{\log |A|})} \)) yields (A.2). \( \square \)

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 refer to [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 \( \beta \geq 1 \) in Theorem 3) in contrast to the entropy integral bound in [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 [20].

**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 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 \( 0 < \theta < 1 \),

\[
E \left| \sum_{i<j} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| \leq 2\sqrt{2n^2} \left( \inf_{\alpha > 0} \left\{ \alpha + \frac{3e}{\theta} \int_0^\beta \frac{1}{n} \frac{N(g, \rho, \varepsilon)}{\sqrt{1-\theta}} \, de \right\} \right) + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0),
\]

where for any \( g_1, g_2 \in \mathcal{G} \),

\[
\rho(g_1, g_2) = \sqrt{2n^2 \sum_{i<j} (g_1(x_i, x_j) - g_2(x_i, x_j))^2}
\]

and therefore

\[
E \left| \sum_{i<j} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| < 2\sqrt{2n^2} \left( \inf_{\alpha > 0} \left\{ \alpha + \frac{3e}{n} \int_0^\beta \left( 1 + \sqrt{\log N(g, \rho, \varepsilon)} \right) \, de \right\} \right) + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0).
\]

**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 \( \mathcal{G}_l := \{g_1, \ldots, g_l N(G, \rho, \delta_l)\} \) be a \( \rho \)-cover of \( \mathcal{G} \) at scale \( \delta_l \). For any \( M \), any \( g \in \mathcal{G} \) can be expressed as

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

35
where \( g_i \in \mathcal{G}_i \) and \( \mathcal{G}_0 := \mathcal{G} \). Note that \( \rho(g_i, g_{i-1}) \leq \rho(g, g_i) + \rho(g, g_{i-1}) \leq \delta_i + \delta_{i-1} = 3\delta_i \). Consider

\[
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| \leq \mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j (g(x_i, x_j) - g_M(x_i, x_j)) \right| + \mathbb{E} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g_0(x_i, x_j) \right|
+ \sum_{l=1}^M \mathbb{E} \sup_{g \in \mathcal{G}_l, g_{l-1} \in \mathcal{G}_{l-1} \atop \rho(g, g_{l-1}) \leq 3\delta_l} \left| \sum_{i<j} \varepsilon_i \varepsilon_j (g_l(x_i, x_j) - g_{l-1}(x_i, x_j)) \right|. \tag{A.3}
\]

Note that

\[
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j (g(x_i, x_j) - g_M(x_i, x_j)) \right| \leq \mathbb{E} \sum_{j=1}^n \varepsilon_j^2 \mathbb{E} \sup_{g \in \mathcal{G}} \left( \sum_{i<j} (g(x_i, x_j) - g_M(x_i, x_j))^2 \right) = \frac{n^2 \sup_{g \in \mathcal{G}} \rho(g, g_M)}{\sqrt{2}}, \tag{A.4}
\]

\[
\mathbb{E} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g_0(x_i, x_j) \right| \leq \left( \mathbb{E} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g_0(x_i, x_j) \right|^2 \right)^{1/2} \leq \left( \sum_{i<j} g_0^2(x_i, x_j) \right)^{1/2} \leq \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0). \tag{A.5}
\]

and by Lemma [A.1]

\[
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j (g_l(x_i, x_j) - g_{l-1}(x_i, x_j)) \right| \leq \frac{3\varepsilon_i n \log N(\mathcal{G}, \rho, \delta_i)}{\sqrt{2}} \frac{N(\mathcal{G}, \rho, \delta_{i+1})}{1 - \theta} \leq \frac{6\varepsilon_i n \log N(\mathcal{G}, \rho, \delta_i)}{\sqrt{1 - \theta}} \tag{A.6}
\]

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

\[
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| \leq \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{6\varepsilon_i n}{\sqrt{2}} \sum_{i=1}^M \delta_i \log \frac{N(\mathcal{G}, \rho, \delta_i)}{\sqrt{1 - \theta}} \leq \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{12\varepsilon_i n}{\sqrt{2}} \sum_{i=1}^M (\delta_i - \delta_{i+1}) \log \frac{N(\mathcal{G}, \rho, \delta_i)}{\sqrt{1 - \theta}} \leq \frac{n^2 \delta_M}{\sqrt{2}} + \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + \frac{12\varepsilon_i n}{\sqrt{2}} \int_{\delta_M}^{\delta_0} \log \frac{N(\mathcal{G}, \rho, \varepsilon)}{\sqrt{1 - \theta}} d\varepsilon. \tag{A.7}
\]

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

\[
\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i<j} \varepsilon_i \varepsilon_j g(x_i, x_j) \right| \leq \frac{n}{\sqrt{2}} \sup_{g \in \mathcal{G}} \rho(g, 0) + 2\sqrt{2}n \alpha + \frac{6\sqrt{2} \varepsilon_i n}{\theta} \int_{\alpha}^{\sup g_1, g_2 \in \mathcal{G}, \rho(g_1, g_2)} \log \frac{N(\mathcal{G}, \rho, \varepsilon)}{\sqrt{1 - \theta}} d\varepsilon.
\]

Since \( \alpha \) is arbitrary, taking infimum over \( \alpha > 0 \) yields the result. \( \square \)

### B Bound on \( \mathcal{N}(K_{\alpha}, \rho_{\alpha}, \varepsilon) \) in Theorem 5(a)

The following result presents a bound on \( \mathcal{N}(K_{\alpha}, \rho_{\alpha}, \varepsilon) \) when

\[
K = \left\{ e^{-\sigma}\|x-y\|^2, \ x, y \in (a_0, b_0)^d, \ -\infty < a_0 < b_0 < \infty : \sigma \in (0, a] \right\},
\]
using which it is easy to check that $\omega_* < 1$ in Theorem 4 and therefore the claims shown in Theorem 5 follow.

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

$$N(K_\alpha, \rho_\alpha, \epsilon) = \frac{C}{\epsilon}$$

where $\rho_\alpha$ is defined in Theorem 4 and $C$ is a constant that depends on $a_0, b_0, d$ and $r$.

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

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

Let $\Sigma(\tau) := \{\sigma_1, \ldots, \sigma_{N(\tau)}\}$ be the $L^1$ cover of $\Sigma$. Define $\tilde{K}_\alpha := \{\partial^{\alpha,\alpha} \psi_\sigma(x - y), x, y \in (a_0, b_0)^d, -\infty < a_0 < b_0 < \infty : \sigma \in \Sigma(\tau)\}$. Using the expression for $\partial^{\alpha,\alpha} \psi_\sigma$ in (62), we have

$$|\partial^{\alpha,\alpha} \psi_\sigma - \partial^{\alpha,\alpha} \psi_{\sigma_1}|(x - y) \leq \sum_{j_1=0}^{a_1-1} \cdots \sum_{j_d=0}^{a_d-1} A_{j_1 \cdots j_d} \left| \sigma + \sum_{i=1}^{d_j} j_i e^{-\sigma \|x-y\|_2^2} - \sigma + \sum_{i=1}^{d_j} j_i e^{-\sigma_1 \|x-y\|_2^2} \right|$$

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

$$C := \left| \sigma + \sum_{i=1}^{d_j} j_i e^{-\sigma \|x-y\|_2^2} - \sigma + \sum_{i=1}^{d_j} j_i e^{-\sigma_1 \|x-y\|_2^2} \right|$$

can be bounded as

$$C \leq \left| \sigma + \sum_{i=1}^{d_j} j_i - \sigma + \sum_{i=1}^{d_j} j_i \right| + a^{r+\sum_{i=1}^{d_j} j_i} \left| e^{-\sigma \|x-y\|_2^2} - e^{-\sigma_1 \|x-y\|_2^2} \right|$$

$$\leq \left( r + \sum_{i=1}^{d} j_i - 1 \right) a^{r+\sum_{i=1}^{d_j} j_i} \left| \sigma - \sigma_1 \right| + a^{r+\sum_{i=1}^{d_j} j_i} \left| \sigma - \sigma_1 \right| \|x-y\|_2^2 e^{a \|x-y\|_2^2}$$

$$\leq (2r - 1) a^{2r-1} \left| \sigma - \sigma_1 \right| + d (b_0 - a_0)^2 a^{2r} \|x-y\|_2^2 e^{a \cdot (b_0 - a_0)^2} \leq \mu \tau,$$

where $\mu$ is a constant that depends on $a, a_0, b_0, d$ and $r$. Therefore,

$$\rho_\alpha(\partial^{\alpha,\alpha} \psi_\sigma, \partial^{\alpha,\alpha} \psi_{\sigma_1}) \leq \| \partial^{\alpha,\alpha} \psi_\sigma - \partial^{\alpha,\alpha} \psi_{\sigma_1} \|_\infty \leq \mu \tau \sum_{j_1=0}^{a_1} \cdots \sum_{j_d=0}^{a_d} B_{j_1 \cdots j_d},$$

which yields the result. $\square$

**Acknowledgements**

I profusely thank Richard Nickl for many valuable comments and insightful discussions.

**References**

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, UK, 1999.

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

[3] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer Verlag, New York, 1984.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, London, UK, 2004.

P. J. Bickel and Y. Ritov. Nonparametric estimators which can be “plugged-in”. *Annals of Statistics*, 31(4):1033–1053, 2003.

V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer-Verlag, NY, 1999.

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, New York, 1985.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, USA, 1977.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, UK, 1999.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK, 2002.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, 1999.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Rowes, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 473–480, 2009.

E. Giné and R. Nickl. Adaptation on the space of finite signed measures. *Mathematical Methods of Statistics*, 17:113–122, 2008.

E. Giné and R. Nickl. Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141:333–387, 2008.

E. Giné and R. Nickl. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, 143:569–596, 2009.

E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Annals of Probability*, 37:1605–1646, 2009.

E. Giné and R. Nickl. Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli*, 16(4):1137–1163, 2010.

E. Giné and J. Zinn. Empirical processes indexed by Lipschitz functions. *Annals of Probability*, 14:1329–1338, 1986.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

W. Hardle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*. Springer-Verlag, New York, 1998.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwith selectors. *Annals of Statistics*, 25:929–947, 1997.

D. J. Marcus. Relationships between Donsker classes and Sobolev spaces. *Z. Wahrscheinlichkeits-theorie verw. Gebiete*, 69:323–330, 1985.

S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

R. Nickl. Donsker-type theorems for nonparametric maximum likelihood estimators. *Probability Theory and Related Fields*, 138:411–449, 2007.
[27] D. Radulović and M. Wegkamp. Weak convergence of smoothed empirical processes: Beyond Donsker classes. In E. Giné, D. M. Mason, and J. A. Wellner, editors, High dimensional probability II, Progress in Probability, volume 47, pages 89–105. Brikhäuser, Boston, 2000.

[28] W. Rudin. Functional Analysis. McGraw-Hill, USA, 1991.

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

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

[31] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. Electronic Journal of Statistics, 6:1550–1599, 2012.

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

[33] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. Journal of Machine Learning Research, 12:2389–2410, 2011.

[34] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research, 11:1517–1561, 2010.

[35] I. Steinwart and A. Christmann. Support Vector Machines. Springer, 2008.

[36] A. W. van der Vaart. Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics, 21:501–504, 1994.

[37] A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, Cambridge, 1998.

[38] A. W. van der Vaart and J. A. Wellner. Weak Convergence and Empirical Processes. Springer-Verlag, New York, 1996.

[39] H. Wendland. Scattered Data Approximation. Cambridge University Press, Cambridge, UK, 2005.

[40] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. Neural Computation, 22:2858–2886, 2010.

[41] J. E. Yukich. Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics, 19:271–279, 1992.