JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–24

Consistency of Robust Kernel Density Estimators

Robert A. Vandermeulen RVDM@UMICH.EDU
Clayton D. Scott CLAYSCOT@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109

Abstract

The kernel density estimator (KDE) based on a radial positive-semidefinite kernel may be viewed as a sample mean in a reproducing kernel Hilbert space. This mean can be viewed as the solution of a least squares problem in that space. Replacing the squared loss with a robust loss yields a robust kernel density estimator (RKDE). Previous work has shown that RKDEs are weighted kernel density estimators which have desirable robustness properties. In this paper we establish asymptotic L^1 consistency of the RKDE for a class of losses and show that the RKDE converges with the same rate on bandwidth required for the traditional KDE. We also present a novel proof of the consistency of the traditional KDE.

Keywords: Kernel Density Estimation, Robust Estimation, Reproducing Kernel Hilbert Space, Consistency

1. Introduction

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf and $X_1, \ldots, X_n$ be iid samples from f. Let $k_\sigma(x, x')$ be a radial smoothing kernel of the form $k_\sigma(x, x') = \sigma^{-d} q(\|x - x'\|_2 / \sigma)$ for some function $q \geq 0$ such that $q(\|\cdot\|_2)$ is a pdf on $\mathbb{R}^d$. Then

$$\bar{f}_\sigma^n := \frac{1}{n} \sum_{i=1}^{n} k_\sigma(\cdot, X_i)$$

is the well-known kernel density estimator (KDE) (Silverman (1986), Scott (1992), Devroye and Lugosi (2001)). We additionally assume that k_σ is positive-semidefinite. Thus $k_\sigma(x, x') = \langle \Phi_\sigma(x), \Phi_\sigma(x') \rangle_{\mathcal{H}_\sigma}$, where $\mathcal{H}_\sigma$ is the reproducing kernel Hilbert space (RKHS) associated with k_σ (Aronszajn, 1950), and $\Phi_\sigma(x) := k_\sigma(\cdot, x)$ is the canonical feature map (Steinwart and Christmann, 2008). Some kernels satisfying these properties include the multivariate Gaussian, Laplacian, and Student kernels. Note that for radial kernels we have

$$\begin{aligned} \|\Phi_\sigma(x)\|_{\mathcal{H}_\sigma} &= \sqrt{\sigma^{-d} q(\|x - x\|_2 / \sigma)} \\ &= \sqrt{q(0)} \sigma^{-d/2} \end{aligned}$$

which does not depend on x. Because of this, we will abuse notation slightly and let $\|\Phi_\sigma\|_{\mathcal{H}_\sigma} \triangleq \|\Phi_\sigma(x)\|_{\mathcal{H}_\sigma}$. Note that as $\sigma \to 0$, $\|\Phi_\sigma\|_{\mathcal{H}_\sigma}$ grows without bound, a fact we will use frequently.

With this notation, the KDE may be written as

$$\bar{f}_\sigma^n = \frac{1}{n}\sum_{i=1}^n \Phi_\sigma(X_i),$$

the mean of the mapped data. The sample mean is easily shown to be the unique solution of a least squares problem

$$\bar{f}_\sigma^n = \operatorname*{argmin}_{g\in\mathcal{H}_\sigma} \frac{1}{n}\sum_{i=1}^n \|g - \Phi_\sigma(X_i)\|_{\mathcal{H}_\sigma}^2.$$

Replacing the squared loss with a robust loss ρ, yields a *robust kernel density estimator*:

$$f_\sigma^n = \operatorname*{argmin}_{g\in\mathcal{H}_\sigma} \frac{1}{n}\sum_{i=1}^n \rho\left(\|g - \Phi_\sigma(X_i)\|_{\mathcal{H}_\sigma}\right). \tag{1}$$

This construction was first introduced by Kim and Scott (2012) where they established several properties including a representer theorem, a convergent iterative algorithm, and the influence function. The representer theorem states that

$$f_\sigma^n = \sum_{i=1}^n \alpha_i \Phi_\sigma(X_i),$$

where $\alpha_i \geq 0$ and $\sum_1^n \alpha_i = 1$.

In this paper we will establish consistency of the RKDE in the L^1 norm. Throughout the paper σ will implicitly be a function of n, such that $\sigma \to 0$ as $n \to \infty$. We will use f_σ^n to denote the RKDE for a general loss ρ and $\bar{f}_\sigma^n$ to denote the special case corresponding to $\rho(\cdot) = (\cdot)^2$, i.e. the classic KDE.

1.1. Related Work

The consistency of kernel density estimators has been established under the L^1 norm with very weak assumptions on distribution and kernel (Devroye and Lugosi, 2001). Necessary conditions on n and σ for L^1 consistency of the KDE are $n \to \infty$ with $\sigma \to 0$ and rate on bandwidth $n\sigma^d \to \infty$. Sup-norm consistency has also been established for a less general class of kernels and densities requiring more restrictive regularity conditions (Silverman (1978), Stute (1982), Einmahl and Mason (2000), Deheuvels (2000), Giné and Guillou (2002), Gine et al. (2004), Wied and Weissbach (2012)).

Consistency proofs tend to proceed by decomposing the error into a stochastic estimation error and a non-stochastic approximation error, namely

$$\left\|\bar{f}_\sigma^n - f\right\| \leq \left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\| + \left\|\bar{f}_\sigma - f\right\|,$$

where $\bar{f}_\sigma = \int k_\sigma(\cdot, x) f(x)\, dx = \int \Phi_\sigma(x) f(x)\, dx$. The right summand is shown to go to zero analytically and the left summand is shown to go to zero with techniques from empirical process theory. We will show a simple proof of the consistency of the KDE using this decomposition and Bennett's inequality for Hilbert space to control the stochastic term. However, this decomposition is less fruitful for the RKDE, for which f_σ does not have a closed form expression (see Section 4). Instead, we use a completely different technique by investigating the convergent iterative algorithm used to compute the RKDE in Kim and Scott (2012).

2. Novel KDE Consistency Proof

First we will introduce a construction that will be used frequently throughout the paper:

$$\mathcal{D}_\sigma = \left\{ \int \Phi_\sigma(x) d\nu(x) \middle| \nu \text{ is a probability measure} \right\}.$$

Note that this and all Hilbert space valued integrals are Bochner integrals; see Steinwart and Christmann (2008) for a basic introduction to Bochner integrals. For this paper these integrals can be thought of as the convolution of the kernel with a measure. This in turn implies that all elements of $\mathcal{D}_\sigma$ are pdfs. In fact all of the density estimators in this paper will be an element of some $\mathcal{D}_\sigma$.

We will now present a novel proof of L^1 consistency of the kernel density estimator.

Theorem 1 *If $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$ then $\left\| \bar{f}_\sigma^n - f \right\|_1 \xrightarrow{p} 0$.*

Proof Let $\bar{f}_\sigma = \mathbb{E}_{X \sim f} \left[\Phi_\sigma (X) \right]$. By the triangle inequality we have

$$\left\| f - \bar{f}_\sigma^n \right\|_1 \leq \left\| f - \bar{f}_\sigma \right\|_1 + \left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_1.$$

The left term in the sum goes to zero by elementary analysis (Devroye and Lugosi, 2001). We only need to show that $\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_1 \xrightarrow{p} 0$. First we show convergence in the RKHS.

Lemma 2 *Let $\varepsilon > 0$. For sufficiently small σ,*

$$\mathbb{P}\left(\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_{\mathcal{H}_\sigma} \geq \varepsilon \right) \leq \exp\left\{ -\frac{n\varepsilon^2}{4 \left\| \Phi_\sigma \right\|_{\mathcal{H}_\sigma}^2} \right\}.$$

Therefore if $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$, then $\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$.

Proof Sketch Observe that

$$\mathbb{E}\left[\bar{f}_\sigma^n \right] = \mathbb{E}\left[\frac{1}{n} \sum_1^n \Phi_\sigma \left(X_i \right) \right] = \mathbb{E}_{X \sim f} \left[\Phi_\sigma \left(X \right) \right] = \bar{f}_\sigma.$$

This fact combined with Bennett's inequality for Hilbert space yields the inequality in the lemma, after some trivial manipulations. The second part of the lemma is a simple consequence of the inequality. ■

The previous lemma follows from Bennett's inequality for Hilbert space, but Hoeffding's or Bernstein's inequality for Hilbert space would also suffice (Pinelis, 1994). For other examples of simple proofs using concentration inequalities see Caponnetto and Vito (2007) and Bauer et al. (2007). The next lemma allows us to bound L^1 norms over sets of finite Lebesgue measure. Let λ denote Lebesgue measure.

Lemma 3 *Let $S \in \mathbb{R}^d$ be a set with finite Lebesgue measure and $g \in \mathcal{H}_\sigma$. Then*

$$\int_S |g(x)| \, dx \leq 2\sqrt{\lambda(S)} \left\| g \right\|_{\mathcal{H}_\sigma}.$$

Proof Sketch We will present a proof for the situation where $g > 0$. For the general case we can split the following integral into two parts corresponding to the subsets of S where g is positive and g is negative. We have,

$$\begin{aligned}
\left(\int_S g(x)\,dx\right)^2 &= \left(\int_S \langle \Phi_\sigma(x), g\rangle_{\mathcal{H}_\sigma}\,dx\right)^2 \\
&= \left(\left\langle \int_S \Phi_\sigma(x)\,dx, g\right\rangle_{\mathcal{H}_\sigma}\right)^2 \\
&\leq \left\|\int_S \Phi_\sigma(x)\,dx\right\|_{\mathcal{H}_\sigma}^2 \|g\|_{\mathcal{H}_\sigma}^2 \\
&= \int_S\int_S \left\langle \Phi_\sigma(x), \Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} dx dx' \|g\|_{\mathcal{H}_\sigma}^2 \\
&= \int_S\int_S k_\sigma(x,x')\,dx dx' \|g\|_{\mathcal{H}_\sigma}^2 \\
&\leq \int_S 1 dx \|g\|_{\mathcal{H}_\sigma}^2 \\
&= \lambda(S)\|g\|_{\mathcal{H}_\sigma}^2 .
\end{aligned}$$

■

For pdfs embedded in RKHSs, Lemma 3 allows us to show that $\mathcal{H}_\sigma$ convergence implies L^1 convergence.

Lemma 4 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf and g_σ^n and h_σ^n be sequences of (possibly random) densities in a sequence of spaces $\mathcal{D}_\sigma$ (again σ is implicitly a function of n). If $\|g_\sigma^n - f\|_1 \xrightarrow{p} 0$ and $\|g_\sigma^n - h_\sigma^n\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$ then $\|g_\sigma^n - h_\sigma^n\|_1 \xrightarrow{p} 0$.*

Proof Sketch Define $B(y,r)$ to be the open ball centered at y with radius r and χ_S to be the indicator function on the set S. Let $\varepsilon > 0$. Choose r large enough that $\int_{B(0,r)^C} f(x)\,dx < \varepsilon/3$ (this is possible by Lemma 11 in the appendix). Since $B(0,r)$ and $B(0,r)^C$ partition $\mathbb{R}^d$ we have

$$\begin{aligned}
\|g_\sigma^n - h_\sigma^n\|_1 &= \left\|(g_\sigma^n - h_\sigma^n)\left(\chi_{B(0,r)} + \chi_{B(0,r)^C}\right)\right\|_1 \\
&= \left\|(g_\sigma^n - h_\sigma^n)\chi_{B(0,r)}\right\|_1 + \left\|(g_\sigma^n - h_\sigma^n)\chi_{B(0,r)^C}\right\|_1 . \qquad (2)
\end{aligned}$$

The left summand goes to zero in probability by Lemma 3 so it becomes bounded by $\varepsilon/3$ with probability going to one. Since $\left\|(f - g_\sigma^n)\chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} 0$ we have $\left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)^C}\right\|_1 < \varepsilon/3$. Since g_σ^n and h_σ^n are densities and both of them are converging to have the same amount of mass in $B(0,r)$, their mass in $B(0,r)^C$ must also be converging. This means $\left|\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1 - \left\|g_\sigma^n\chi_{B(0,r)^C}\right\|_1\right| \xrightarrow{p} 0$ so $\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1$ becomes bounded by $\varepsilon/3$ with probability going to one. Thus the right summand of (2)

becomes bounded by $2\varepsilon/3$ with high probability. Putting these results together we have $\|g_\sigma^n - h_\sigma^n\|_1 < \varepsilon$ with probability going to one.

■

The previous lemma is a bit more general than is necessary for the current theorem, but it will be handy later. In this case g_σ^n in the last lemma is replaced by $\bar{f}_\sigma$ and h_σ^n is replaced with $\bar{f}_\sigma^n$, thus completing our proof of Theorem 1. ■

It is worth noting that Lemma 2 also implies consistency with respect to L^2 and L^∞ norms, assuming suitable conditions ensuring that the approximation error goes to zero. L^2 consistency is implied as long as $k_\sigma(\cdot, x) \in L^2(\mathbb{R}^d)$ for all $x \in \mathbb{R}^d$, (in particular, k_σ need not be a reproducing kernel) because Lemma 2 holds for general Hilbert spaces. L^∞ consistency follows from the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\bar{f}_\sigma^n(x) - \bar{f}_\sigma(x)\right| &= \left|\left\langle \Phi_\sigma(x), \bar{f}_\sigma^n - \bar{f}_\sigma \right\rangle_{\mathcal{H}_\sigma}\right| \\ &\leq \|\Phi_\sigma\|_{\mathcal{H}_\sigma} \left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}.\end{aligned}$$

Unfortunately the $\|\Phi_\sigma\|_{\mathcal{H}_\sigma}$ term in the last line yields a suboptimal rate on the bandwidth, $n\sigma^{2d} \to \infty$.

3. RKDE Consistency

We begin by reviewing some results about the RKDE.

3.1. Previous Results

Before we prove consistency of the RKDE, we will introduce some additional technical background on the RKDE from Kim and Scott (2012). First we will define some properties ρ may have. Let $\rho : [0,\infty) \to [0,\infty)$, $\psi \triangleq \rho'$, and $\varphi(x) \triangleq \psi(x)/x$. Consider the following properties:

(B1) ρ is strictly convex
(B2) ρ is strictly increasing, $\rho(0) = 0$ and $\rho(x)/x \to 0$ as $x \to 0$
(B3) $\varphi(0) := \lim_{x\to 0} \frac{\psi(x)}{x}$ exists and is finite
(B4) ψ is bounded
(B5) ρ'' exists and is nonincreasing on $(0,\infty)$
(B6) φ is nonincreasing.

Some examples of losses satisfying all of these properties are $\rho(x) = \sqrt{x^2+1} - 1$, $\rho(x) = x\arctan(x)$, and $\rho(x) = x - \log(1+x)$. It is easy to show that property (B1) guarantees the existence and uniqueness of f_σ^n (Kim and Scott, 2012). Let f be a pdf and $X_1, \cdots, X_n$ be iid samples from f. Let $J_\sigma^n(\cdot)$ be the empirical risk introduced in (1). Taking the Gateaux derivative of the risk gives us

$$\delta J_\sigma^n(g; h) = -\left\langle \frac{1}{n}\sum_1^n \varphi\left(\|\Phi_\sigma(X_i) - g\|_{\mathcal{H}_\sigma}\right)(\Phi_\sigma(X_i) - g), h \right\rangle_{\mathcal{H}_\sigma}.$$

If (B2) and (B3) are satisfied then a necessary condition for $g = f_\sigma^n$ is that the Gateaux derivative at g is 0 for all directions h, which is equivalent to left term in the inner product

being 0 (Lemma 1 Kim and Scott (2012)). A straightforward algebraic manipulation of the last condition gives us

$$\frac{\sum_1^n \varphi\left(\left\|\Phi_\sigma\left(X_i\right)-g\right\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma\left(X_i\right)}{\sum_1^n \varphi\left(\left\|\Phi_\sigma\left(X_j\right)-g\right\|_{\mathcal{H}_\sigma}\right)}=g.$$

With this in mind we introduce the following functional,

$$\begin{aligned}R_\sigma^n:\mathcal{H}_\sigma\to\mathcal{H}_\sigma:g\mapsto R_\sigma^n(g)&=\frac{\int\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x)d\mu_n(x)}{\int\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)d\mu_n(y)}\\&=\sum_1^n\alpha_i\left(g\right)k_\sigma\left(\cdot,X_i\right)\end{aligned}$$

where

$$\alpha_i\left(g\right)=\frac{\varphi\left(\left\|\Phi_\sigma(X_i)-g\right\|_{\mathcal{H}_\sigma}\right)}{\sum_1^n\varphi\left(\left\|\Phi_\sigma(X_j)-g\right\|_{\mathcal{H}_\sigma}\right)}$$

and μ_n is the empirical measure corresponding to the sample. This function is the Iterated Reweighted Least Squares algorithm (IRWLS) from Kim and Scott (2012), which is used to compute the RKDE in practice. From Corollary 6 in Kim and Scott (2012) it is easy to show that if (B1), (B2), (B3), (B5), and (B6) are satisfied (note that (B4) is used later), the sequence $\{R_\sigma^n\left(0\right),R_\sigma^n\left(R_\sigma^n\left(0\right)\right),\ldots\}$ converges in $\mathcal{H}_\sigma$ to f_σ^n, which is the unique fixed point of R_σ^n.

3.2. Consistency Theorem and Proof

Theorem 5 *Let $f\in L^2\left(\mathbb{R}^d\right)$ and let ρ satisfy (B1)-(B6). If $n\sigma^d\to\infty$ and $\sigma\to 0$ as $n\to\infty$ then $\left\|f_\sigma^n-f\right\|_1\xrightarrow{p}0$.*

We know that ψ is bounded by (B4). In the proofs that follow it will be assumed, for simplicity, that $\sup_x\psi\left(x\right)=1$. Note that any loss with bounded ψ can be adapted such that $\sup_x\psi\left(x\right)=1$. This is done by dividing ρ by $\sup_x\psi\left(x\right)$ and does not affect the RKDE. The longer and more technical proof sketches are contained in a subsection after this one.

The following lemma helps us establish the behavior of elements in $\mathcal{D}_\sigma$ with large norms.

Lemma 6 *For all $g\in\mathcal{D}_\sigma$, $\left\|g\right\|_{\mathcal{H}_\sigma}^2\leq\left\|g\right\|_\infty$.*

Proof By the definition of $\mathcal{D}_\sigma$, let $g=\int\Phi_\sigma(x)d\nu(x)$, where ν is a probability measure.

$$\begin{aligned}\left\|g\right\|_{\mathcal{H}_\sigma}^2=\left\langle g,g\right\rangle_{\mathcal{H}_\sigma}&=\left\langle\int\Phi_\sigma(x)d\nu(x),g\right\rangle_{\mathcal{H}_\sigma}=\int\left\langle\Phi_\sigma(x),g\right\rangle_{\mathcal{H}_\sigma}d\nu(x)\\&=\int g(x)d\nu(x)\leq\int\left\|g\right\|_\infty d\nu(x)=\left\|g\right\|_\infty.\end{aligned}$$

■

This lemma allows us to show that an element in $\mathcal{D}_\sigma$ with large norm will have most of its mass concentrated around one point. An element of $\mathcal{D}_\sigma$ having most of the mass around one point causes its general risk to be large. The Vapnik-Chervonenkis inequality allows us to show that all such elements will, with high probability, have high empirical risk.

Lemma 7 *If $\sigma \to 0$ and $n \to \infty$ then $\mathbb{P}\left(\|f_\sigma^n\|_{\mathcal{H}_\sigma}^2 \geq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\right) \to 0$.*

The constant $\frac{9}{10}$ was chosen simply for convenience, it could be replaced with any positive value less than one.

The following result will be used to prove Lemma 9 and Theorem 5.

Lemma 8 $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2$.

Proof Using the Cauchy-Schwarz inequality and Young's inequality (Devroye and Lugosi, 2001) we have

$$\begin{aligned}\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}^2 &= \left\langle \int f(x)\Phi_\sigma(x)dx, \int f(y)\Phi_\sigma(y)dy \right\rangle_{\mathcal{H}_\sigma} = \int f(x)\left\langle \Phi_\sigma(x), \int f(y)\Phi_\sigma(y)dy \right\rangle_{\mathcal{H}_\sigma} dx \\ &= \int f(x)\,(f * k_\sigma)\,(x)dx = \langle f, f * k_\sigma\rangle_2 \leq \|f\|_2\,\|f * k_\sigma\|_2 \leq \|f\|_2\,\|f\|_2\,\|k_\sigma\|_1 = \|f\|_2^2\,.\end{aligned}$$

■

Lemma 7 shows that f_σ^n is, with high probability, in a ball of radius $\sqrt{\frac{9}{10}}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}$. Lemma 9 shows that, on that ball, R_σ^n is a contraction mapping.

Lemma 9 *Let $n \to \infty$, $\sigma \to 0$, and $n\sigma^d \to \infty$. There exists C_R such that, with probability going to one, the restriction of R_σ^n to $B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\right)$ is Lipschitz continuous with Lipschitz constant $C_R\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1}$.*

This lemma is the final key to proving Theorem 5.

Proof of Theorem 5 Using the triangle inequality we get

$$\|f - f_\sigma^n\|_1 \leq \left\|f - \bar{f}_\sigma^n\right\|_1 + \left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_1.$$

We know the left term of the summand goes to zero in probability by Theorem 1, so it is sufficient to show that the right summand goes to zero in probability. By Lemma 4 it is sufficient to show that $\left\|f_\sigma^n - \bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma}$ goes to zero in probability.

Notice that $R_\sigma^n(0) = \bar{f}_\sigma^n$ and recall $R_\sigma^n\left(f_\sigma^n\right) = f_\sigma^n$. Using Lemma 7 and 9, with probability going to 1, the following holds

$$\begin{aligned}\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} &= \|R_\sigma^n(0) - R_\sigma^n\left(f_\sigma^n\right)\|_{\mathcal{H}_\sigma} \\ &\leq \|f_\sigma^n - 0\|_{\mathcal{H}_\sigma}\,\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1}\,C_R \\ &< \sqrt{\frac{9}{10}}\,\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\,\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1}\,C_R \\ &= \sqrt{\frac{9}{10}}C_R.\end{aligned}$$

Since $\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$ and $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2 < \infty$ (by Lemma 8), for arbitrary $s > 0$ we have $\left\|\bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma} < \|f\|_2 + s$ with probability going to one. Applying the contraction mapping steps again we get, with probability going to 1, that

$$\begin{aligned}
\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} &= \|R_\sigma^n(0) - R_\sigma^n\left(f_\sigma^n\right)\|_{\mathcal{H}_\sigma} \\
&\leq \|f_\sigma^n - 0\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&\leq \left(\left\|f_\sigma^n - \bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma} + \left\|\bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma}\right) \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&\leq \left(\sqrt{\frac{9}{10}} C_R + \|f\|_2 + s\right) \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R.
\end{aligned}$$

The last line goes to zero as $\sigma \to 0$, completing our proof. ■

3.3. Proof Sketches

Proof Sketch of Lemma 7 We know that $f_\sigma^n \in \mathcal{D}_\sigma$, so to prove this lemma we will show that as $n \to \infty$ and $\sigma \to 0$, all vectors in $\mathcal{D}_\sigma$ with $\mathcal{H}_\sigma$-norm greater than or equal to $\sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}$ will have empirical risk greater than the zero vector. Define $J_\sigma^n : \mathcal{H}_\sigma \to \mathbb{R}$ as the empirical risk function

$$J_\sigma^n(g) = \frac{1}{n} \sum_1^n \rho\left(\|\Phi_\sigma(X_i) - g\|_{\mathcal{H}_\sigma}\right).$$

Let g_σ^n be the minimizer of J_σ^n when restricted to vectors in $\mathcal{D}_\sigma$ with $\mathcal{H}_\sigma$-norm greater than or equal to $\sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}$. By Lemma 6 there must exist x^* such that $g_\sigma^n(x^*) \geq \frac{9}{10} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$, this causes most of of the mass of g_σ^n to reside near x^*. It is possible to show that, given any $r > 0$ and $\varepsilon > 0$, for sufficiently small σ, that $\sup_{x \in B(x^*,r)^C} g_\sigma^n(x) < \frac{3}{20} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 + \varepsilon$. As n gets large, J_σ^n becomes well approximated by J_σ where

$$J_\sigma(g) = \int \rho\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right) f(x)\, dx. \tag{3}$$

We will substitute J_σ for J_σ^n (in the formal proof we work with J_σ^n and invoke the VC inequality to relate it to the population risk). Since ρ is increasing, the following holds for sufficiently small σ,

$$\begin{aligned}
J_\sigma\left(g_\sigma^n\right) &\geq \int_{B(x^*,r)^C} \rho\left(\|\Phi_\sigma(x) - g_\sigma^n\|_{\mathcal{H}_\sigma}\right) f(x)\, dx \\
&\geq \int_{B(x^*,r)^C} \rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 - 2\left\langle g_\sigma^n, \Phi_\sigma(x)\right\rangle_{\mathcal{H}_\sigma} + \|g_\sigma^n\|_{\mathcal{H}_\sigma}^2}\right) f(x)\, dx \\
&\geq \int_{B(x^*,r)^C} \rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 - 2\left(\frac{3}{20} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 + \varepsilon\right) + \|g_\sigma^n\|_{\mathcal{H}_\sigma}^2}\right) f(x)\, dx.
\end{aligned}$$

Since ε can be set to be arbitrarily small and $\|g_\sigma^n\|_{\mathcal{H}_\sigma}^2 \geq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$ the last term has an approximate lower bound of

$$\begin{aligned}
&\gtrsim \int_{B(x^*,r)^C} \rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 - \frac{6}{20}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 + \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2}\right) f(x)\,dx \\
&\geq \rho\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{32}{20}}\right) \inf_y \int_{B(y,r)^C} f(x)\,dx.
\end{aligned}$$

Finally r can be chosen to be sufficiently small so that $\inf_y \int_{B(y,r)^C} f(x)\,dx$ is arbitrarily close to one. Thus as $n \to \infty$ and $\sigma \to 0$, with probability going to one

$$J_\sigma^n(g_\sigma^n) \gtrsim \rho\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{32}{20}}\right).$$

Now, notice that

$$\begin{aligned}
J_\sigma^n(0) &= \frac{1}{n}\sum_1^n \rho\left(\|\Phi_\sigma(X_i) - 0\|_{\mathcal{H}_\sigma}\right) \\
&= \rho\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\right).
\end{aligned}$$

It then follows that, with probability going to one, $J_\sigma^n(g_\sigma^n) > J_\sigma^n(0)$. ■

Proof Sketch of Lemma 9 Let $g, h \in B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\right)$. We have

$$\begin{aligned}
&\|R_\sigma^n(g) - R_\sigma^n(h)\|_{\mathcal{H}_\sigma} \\
&= \left\| \frac{\int \varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\Phi(x)\,d\mu_n(x)}{\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)d\mu_n(y)} - \frac{\int \varphi\left(\|\Phi_\sigma(x') - h\|_{\mathcal{H}_\sigma}\right)\Phi(x)\,d\mu_n(x')}{\int \varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right)d\mu_n(y')} \right\|_{\mathcal{H}_\sigma}. \quad (4)
\end{aligned}$$

Note that all integrals are over the same measure. Consider the situation if the integrals were evaluated at one point,

$$\begin{aligned}
&\left| \frac{\varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)}{\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)} - \frac{\varphi\left(\|\Phi_\sigma(x) - h\|_{\mathcal{H}_\sigma}\right)}{\varphi\left(\|\Phi_\sigma(y) - h\|_{\mathcal{H}_\sigma}\right)} \right| \quad (5) \\
&= \left| \frac{\varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\varphi\left(\|\Phi_\sigma(y) - h\|_{\mathcal{H}_\sigma}\right) - \varphi\left(\|\Phi_\sigma(x) - h\|_{\mathcal{H}_\sigma}\right)\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)}{\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)\varphi\left(\|\Phi_\sigma(y) - h\|_{\mathcal{H}_\sigma}\right)} \right|.
\end{aligned}$$

We will now find a lower bound on the denominator. Note that since g and h live in $B_{\mathcal{H}_\sigma}\left(0, \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{9}{10}}\right)$, that $\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}$ and $\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}$ grow without bound as $\sigma \to 0$. Since ρ is convex ψ must be increasing and since ψ has a supremum of 1, $\psi(z)$ is well approximated by 1 for large z. Thus we have, for small σ that the denominator is well

approximated as follows

$$\begin{aligned}\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)=&\frac{\psi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)}{\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}}\frac{\psi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)}{\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}}\\ \approx&\frac{1}{\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}}\\ \geq&\frac{1}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2\left(1+\sqrt{9/10}\right)^2}\\ =&C_D\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}\end{aligned}$$

where $C_D=\left(1+\sqrt{9/10}\right)^{-2}$. We will now find an upper bound on the numerator. By the triangle inequality

$$\begin{aligned}&\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\right|\\ &\leq\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\right|\\ &+\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\right|.\end{aligned}$$

Consider the second summand,

$$\begin{aligned}&\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\right| \qquad (6)\\ &=\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)\right|\\ &\leq\frac{1}{\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}}\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right|\\ &\leq\frac{1}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)}\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right|.\end{aligned}$$

Just as $\varphi(z)$ becomes well approximated by $\frac{1}{z}$ for large z, $\varphi'(z)$ becomes well approximated by $\frac{-1}{z^2}$. Using this it can be shown that there exists $C_L>0$ such that, for sufficiently small σ, $\varphi\left(\left\|\Phi_\sigma(y)-\cdot\right\|_{\mathcal{H}_\sigma}\right)$ is Lipschitz continuous on $B_{\mathcal{H}_\sigma}\left(0,\sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right)$ with Lipschitz constant $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_L$. Now we have

$$\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right|\leq\left\|g-h\right\|_{\mathcal{H}_\sigma}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_L.$$

It now follows that (6) is less than or equal to $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-3}C_N$ for some $C_N>0$. Returning to (5), we can now show that it has an upper bound of $\frac{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-3}C_N}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_D}=\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-1}\frac{2C_N}{C_D}$. This generally describes the behavior of the values found in (4). To take care of the $\int\Phi_\sigma(x)\,d\mu_n(x)$ terms, note that by Theorem 1 $\left\|\int\Phi_\sigma(x)\,d\mu_n(x)-\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}\xrightarrow{p}0$ if $n\sigma^d\to\infty$. By Lemma 8, $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}\leq\left\|f\right\|_2$ so $\left\|\int\Phi_\sigma(x)\,d\mu_n(x)\right\|_{\mathcal{H}_\sigma}$ becomes bounded with high probability, thus completing our proof sketch. ■

4. Discussion

In this work we have shown that the limit of the RKDE, as $n \to \infty$ and $\sigma \to 0$, is the distribution f. Therefore the robustness of the RKDE is not manifested in its asymptotic limit, at least for the class of strictly convex losses we study. Rather, the robustness of the RKDE is manifested for finite sample sizes as demonstrated by Kim and Scott (2012).

A key feature of our work is our nonstandard analysis. Standard analysis proceeds by the decomposition, $\|f - f_\sigma^n\|_1 \leq \|f - f_\sigma\|_1 + \|f_\sigma - f_\sigma^n\|_1$, where f_σ is the minimizer of J_σ (defined in Eqn. (3)). Using proof techniques from Kim and Scott (2012) it is easy to show that there exists a pdf, p_σ, satisfying

$$f_\sigma = \int p_\sigma(x)\, \Phi_\sigma(x)\, dx$$

and

$$p_\sigma(x) = \frac{\varphi\left(\|\Phi_\sigma(x) - f_\sigma\|_{\mathcal{H}_\sigma}\right) f(x)}{\int \varphi\left(\|\Phi_\sigma(y) - f_\sigma\|_{\mathcal{H}_\sigma}\right) f(y)\, dy}.$$

In the case of the classic KDE, φ is a constant so $p_\sigma = f$. For a robust loss however, φ is a non-constant function so p_σ does not have a closed form expression. The fact that f_σ and f_σ^n do not have closed form expressions makes the standard analysis difficult.

The function R_σ^n is of some interest of its own. It is mentioned in Kim and Scott (2012) that the IRWLS algorithm converges to the RKDE after very few iterations. This phenomenon may be explained by the small contraction constant exhibited by R_σ^n in Lemma 9. It is also worth noting that the density estimator generated by applying the IRWLS algorithm a fixed number of times is also consistent. More precisely, let $f_\sigma^{n,k} = R_\sigma^n(\cdots R_\sigma^n(R_\sigma^n(0))\cdots)$, where R_σ^n is applied k times, then, given the same consistency requirements for the RKDE, $\left\|f_\sigma^{n,k} - f\right\|_1 \xrightarrow{p} 0$.

The last line of the proof for Theorem 5 allows us to say something about the RKDE rate of convergence. From the proof, if $n\sigma^d \to \infty$, there exists $C > 0$ such that, with probability going to one, $\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} \leq C\sigma^{d/2}$. Letting $\sigma^{d/2} = \frac{\log(n)}{\sqrt{n}}$ gives us $\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} \frac{\sqrt{n}}{\log(n)} \leq C$, a rate of convergence of the RKDE to the KDE. We anticipate that this result can be extended to L^1 convergence of the RKDE to f and will be a focus of future work.

We also note that just as f_σ^n is a robust version of $\bar{f}_\sigma^n$ so is f_σ a robust version of $\bar{f}_\sigma$. To see this consider the expression for p_σ. For the traditional KDE φ is a constant, yielding $p_\sigma = f$. When using a robust loss φ is a decreasing function causing $p_\sigma(x)$ to be smaller for more outlying x. We can consider p_σ to be a robust version of f since it suppresses low density regions of f.

Appendix A. Proofs of Lemmas

For convenience the proofs have been split up into two subsections, one for proofs from the KDE section and the other for proofs from the RKDE section.

A.1. KDE Consistency Proofs

The following lemma is a Hilbert space version of Bennett's inequality (Smale and Zhou, 2007) and will be used in the proof of Lemma 2.

Lemma 10 *Let $\mathcal{H}$ be a Hilbert space and $\{\xi_i\}_{i=1}^m$ be m ($m < \infty$) independent random variables with values in $\mathcal{H}$. Also, assume that for each i, $\|\xi_i\|_{\mathcal{H}} \leq B < \infty$ almost surely. Let $\delta^2 = \sum_{i=1}^m E\left[\|\xi_i\|_{\mathcal{H}}^2\right]$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right\|_{\mathcal{H}} \geq \varepsilon\right) \leq \exp\left\{-\frac{m\varepsilon}{2B}\log\left(1 + \frac{mB\varepsilon}{\delta^2}\right)\right\}, \forall \varepsilon > 0.$$

Proof of Lemma 2 We will apply Lemma 10. From the lemma statement let $\xi_i = \Phi_\sigma(X_i)$ and $m = n$ yielding, for all $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \geq \varepsilon\right) &\leq \exp\left\{-\frac{n\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\log\left(1 + \frac{n\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\varepsilon}{n\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2}\right)\right\} \\ &= \exp\left\{-\frac{n\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\log\left(1 + \frac{\varepsilon}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\right)\right\}.\end{aligned}$$

As $\sigma \to 0$ then $1 + \frac{\varepsilon}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} \to 1$ so for sufficiently small σ

$$\log\left(1 + \frac{\varepsilon}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\right) \geq \frac{\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}$$

and

$$\mathbb{P}\left(\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \geq \varepsilon\right) \leq \exp\left\{-\frac{n\varepsilon^2}{4\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2}\right\}$$

which goes to zero as $\frac{n}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2} \to \infty$, or equivalently $n\sigma^d \to \infty$. So $\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$.

■

Proof of Lemma 3 Let $S^+ = \{s | s \in S, g(s) \geq 0\}$ and $S^- = S \setminus S^+$. We have

$$\begin{aligned}\int_S |g(x)|\, dx &= \int_{S^+} g(x)dx + \int_{S^-} -g(x')dx' \\ &= \int_{S^+} \langle g, \Phi_\sigma(x)\rangle_{\mathcal{H}_\sigma}\, dx + \int_{S^-} \left\langle -g, \Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} dx' \\ &= \left\langle g, \int_{S^+} \Phi_\sigma(x)dx\right\rangle_{\mathcal{H}_\sigma} + \left\langle -g, \int_{S^-} \Phi_\sigma(x')dx'\right\rangle_{\mathcal{H}_\sigma} \\ &\leq \|g\|_{\mathcal{H}_\sigma}\left(\left\|\int_{S^+} \Phi_\sigma(x)dx\right\|_{\mathcal{H}_\sigma} + \left\|\int_{S^-} \Phi_\sigma(x')dx'\right\|_{\mathcal{H}_\sigma}\right). \qquad (7)\end{aligned}$$

Now consider

$$\begin{aligned}\left\|\int_{S^+}\Phi_\sigma(x)dx\right\|^2_{\mathcal{H}_\sigma} &= \left\langle\int_{S^+}\Phi_\sigma(x)dx, \int_{S^+}\Phi_\sigma(x')dx'\right\rangle_{\mathcal{H}_\sigma}\\ &= \int_{S^+}\int_{S^+}\left\langle\Phi_\sigma(x),\Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma}dxdx'\\ &= \int_{S^+}\int_{S^+}k_\sigma\left(x,x'\right)dxdx'\\ &\leq \int_{S^+}1dx'\\ &= \lambda(S^+)\end{aligned}$$

and a similar result can be shown for S^-. Plugging back into (7) we get

$$\begin{aligned}\int_S |g(x)|\,dx &\leq \|g\|_{\mathcal{H}_\sigma}\left(\sqrt{\lambda\left(S^+\right)}+\sqrt{\lambda\left(S^-\right)}\right)\\ &\leq \|g\|_{\mathcal{H}_\sigma}\,2\sqrt{\lambda\left(S\right)}.\end{aligned}$$

■

Lemma 11 *Let f be a pdf, $\varepsilon > 0$, and $y \in \mathbb{R}^d$. There exists $r > 0$ such that*

$$\int_{B(y,r)} f(x)dx \geq 1-\varepsilon.$$

or equivalently

$$\int_{B(y,r)^C} f(x)dx < \varepsilon.$$

Proof We will prove the second statement. Consider the following, where $i \in \mathbb{N}$,

$$\int_{B(y,i)^C} f\left(x\right)dx = \int \chi_{B(y,i)^C}\left(x\right)f\left(x\right)dx.$$

Clearly as $i \to \infty$, $\chi_{B(y,i)^C}f \to 0$ pointwise. Since $\chi_{B(y,i)^C}f$ is dominated by f, $\int \chi_{B(y,i)^C}\left(x\right)f\left(x\right)dx \to \int 0dx = 0$ by the dominated convergence theorem. Thus there exists $n \in \mathbb{N}$ where $\int_{B(y,n)^C} f\left(x\right)dx < \varepsilon$. ■

Proof of Lemma 4 Let $\varepsilon > 0$; by Lemma 11 let $r > 0$ such that $\left\|f\chi_{B(0,r)^C}\right\|_1 < \varepsilon/3$. From Lemma 3 we have

$$\left\|(g^n_\sigma - h^n_\sigma)\chi_{B(0,r)}\right\|_1 \xrightarrow{p} 0.$$

Since $\|g_\sigma^n - f\|_1 \xrightarrow{p} 0$, we have $\left\|g_\sigma^n \chi_{B(0,r)}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)}\right\|_1$, and therefore

$$\begin{aligned}\left|\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 - \left\|f\chi_{B(0,r)^C}\right\|_1\right| &= \left|\left(1 - \left\|h_\sigma^n\chi_{B(0,r)}\right\|_1\right) - \left(1 - \left\|f\chi_{B(0,r)}\right\|_1\right)\right| \\ &= \left|\left\|h_\sigma^n\chi_{B(0,r)}\right\|_1 - \left\|f\chi_{B(0,r)}\right\|_1\right| \\ &\leq \left\|\left(h_\sigma^n - f\right)\chi_{B(0,r)}\right\|_1 \\ &\leq \left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)}\right\|_1 + \left\|\left(g_\sigma^n - f\right)\chi_{B(0,r)}\right\|_1 \\ &\xrightarrow{p} 0.\end{aligned}$$

Thus, $\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)^C}\right\|_1$. Since $\left\|f\chi_{B(0,r)^C}\right\|_1 < \varepsilon/3$, we have

$$\mathbb{P}\left(\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1 \geq \varepsilon 5/12\right) \to 0. \tag{8}$$

Now to finish the proof,

$$\begin{aligned}\mathbb{P}\left(\|h_\sigma^n - g_\sigma^n\|_1 > \varepsilon\right) =& \mathbb{P}\left(\left\|(h_\sigma^n - g_\sigma^n)\chi_{B(0,r)}\right\|_1 + \left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > \varepsilon\right) \\ \leq& \mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)}\right\|_1 \geq \varepsilon/4\right) + \mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right)\end{aligned}$$

We've already shown the left summand goes to zero, now we take care of the right term

$$\begin{aligned}\mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right) &\leq \mathbb{P}\left(\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1 + \left\|g_\sigma^n\chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right) \\ &\leq \mathbb{P}\left(\left\|h_\sigma^n\chi_{B(0,r)^C}\right\|_1 \geq 5\varepsilon/12\right) + \mathbb{P}\left(\left\|g_\sigma^n\chi_{B(0,r)^C}\right\|_1 > \varepsilon/3\right)\end{aligned}$$

The left summand goes to zero by (8). Since $\left\|g_\sigma^n\chi_{B(0,r)^C} - f\chi_{B(0,r)^C}\right\|_1 \to 0$ and $\left\|f\chi_{B(0,r)^C}\right\|_1 < \frac{\varepsilon}{3}$, with probability going to one, we have $\left\|g_\sigma^n\chi_{B(0,r)^C}\right\|_1 \leq \varepsilon/3$ and the right summand goes to zero. This completes our proof. ■

A.2. RKDE Consistency Proofs

Lemma 12 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf. For all $\varepsilon > 0$, there exists $s > 0$ such that $\int_{B(z,s)} f(x)\,dx \leq \varepsilon$ for all $z \in \mathbb{R}^d$.*

Proof We will proceed by contradiction. Let $\{x_i\}_1^\infty$ be a sequence in $\mathbb{R}^d$ such that $\int_{B(x_i,1/i)} f(x)\,dx > \varepsilon$. Clearly the sequence must be bounded or else f would not be a pdf. Let x_{i_j} be a convergent subsequence and let x' be its limit. Let $\{r_j\}_1^\infty$ be a sequence in $\mathbb{R}^+$ converging to zero with $B\left(x_{i_j}, 1/i_j\right) \subset B(x', r_j)$. So we have $\int_{B(x',r_j)} f(x)\,dx > \varepsilon$, for all j. We know

$$\int_{B(x',r_j)} f(x)\,dx = \int \chi_{B(x',r_j)}(x) f(x)\,dx$$

and $f\chi_{B(x',r_j)} \to 0$ pointwise. Since $f\chi_{B(x',r_j)}$ is dominated by f, the dominated convergence theorem yields

$$\begin{aligned}\lim_{j\to\infty}\int_{B(x',r_j)} f(x)\,dx &= \lim_{j\to\infty}\int f(x)\,\chi_{B(x',r_j)}(x)\,dx \\ &= \int \lim_{j\to\infty} f(x)\,\chi_{B(x',r_j)}(x)\,dx \\ &= \int 0dx \\ &= 0\end{aligned}$$

but $\int_{B(x',r_j)} f(x)\,dx > \varepsilon$, a contradiction. ■

Corollary 13 *Let $f:\mathbb{R}^d\to\mathbb{R}$ be a pdf with associated measure μ, $\varepsilon>0$ and $r>0$. There exists $s>0$ such that for all $x\in\mathbb{R}^d$, $\mu(B(x,r+s)\setminus B(x,r))<\varepsilon$.*

Proof We will omit a full proof; the general strategy is the same as the previous proof. Find a series of annuli with width decreasing to zero that have probability greater than ε. Next find a convergent subsequence of annuli centers, let its limit be x'. Finally construct a series of annuli centered at x' with probability measure greater than ε and width going to zero and arrive at the same contradiction. ■

Lemma 14 *Let $s>0$. If $\sigma\to 0$ then $\sigma^{-d}q(s/\sigma)\to 0$.*

Proof We will proceed by contradiction. Suppose $\sigma^{-d}q(s/\sigma)$ does not converge to zero, then there exists $C>0$ such that we can find arbitrarily small σ satisfying

$$\sigma^{-d}q(s/\sigma) > C. \tag{9}$$

It is well known that there exists C_d such that the Lebesgue measure of a ball in $\mathbb{R}^d$ of radius r is $C_d r^d$. Since q is nonincreasing (Scovel et al., 2010) this along with (9) implies that there exists arbitrarily small σ satisfying

$$\begin{aligned}\int_{B(0,s)} \sigma^{-d}q(\|x\|_2/\sigma)\,dx &\geq \int_{B(0,s)} \sigma^{-d}q(s/\sigma)\,dx \\ &> C_d s^d C\end{aligned}$$

where the last term must be less than or equal to 1. Now, by Lemma 11, there exists $r>0$ such that

$$\int_{B(0,r)} q(\|x\|_2)\,dx = \int_{B(0,r\sigma)} \sigma^{-d}q(\|x\|_2/\sigma)\,dx \geq 1-\frac{C_d s^d C}{2}.$$

For sufficiently small σ we have

$$\begin{aligned}1 \geq & \int_{B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx \\ \geq & \int_{B(0,r\sigma)} \sigma^{-d} q\left(\left\|x'\right\|_2/\sigma\right) dx' + \int_{B(0,s)\setminus B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx \\ \geq & 1 - \frac{C_d s^d C}{2} + \int_{B(0,s)\setminus B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx.\end{aligned}$$

Because q is nonincreasing this is greater than or equal to

$$1 - \frac{C_d s^d C}{2} + C_d\left(s^d - (r\sigma)^d\right)\sigma^{-d} q\left(s/\sigma\right).$$

As $\sigma \to 0$, $C_d\left(s^d - (r\sigma)^d\right) \to C_d s^d$, so by (9) we can find some σ where the last term is greater than or equal to

$$1 - \frac{C_d s^d C}{2} + C_d s^d C\frac{2}{3}.$$

The last line is greater than 1, a contradiction.

■

Proof of Lemma 7 Let conv be the convex hull operator. Define

$$Q_\sigma^n = \operatorname{conv}\left(\Phi_\sigma(X_1), \ldots, \Phi_\sigma(X_n)\right) \bigcap B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right)^C.$$

Clearly $Q_\sigma^n \subset \mathcal{D}_\sigma$ since $\Phi_\sigma\left(X_i\right)$ is a density for all i. By the representer theorem in Kim and Scott (2012), $f_\sigma^n \in \operatorname{conv}\left(\Phi_\sigma\left(X_1\right), \ldots, \Phi_\sigma\left(X_n\right)\right)$. We also know that f_σ^n is the minimizer of J_σ^n, where $J_\sigma^n : \mathcal{H}_\sigma \to \mathbb{R}$ is the empirical risk function

$$J_\sigma^n\left(g\right) = \frac{1}{n}\sum_{i=1}^n \rho\left(\left\|\Phi_\sigma\left(X_i\right) - g\right\|_{\mathcal{H}_\sigma}\right).$$

From these facts if we can show

$$\mathbb{P}\left(J_\sigma^n(0) < J_\sigma^n(g), \forall g \in Q_\sigma^n\right) \to 0$$

then we have proven the lemma.

Since Q_σ^n is compact and J_σ^n is continuous (Kim and Scott, 2012), $\arg\min_{g\in Q_\sigma^n} J_\sigma^n(g)$ contains at least one element. Let g_σ^n be an arbitrary minimizer of J_σ^n restricted to Q_σ^n. Let μ be the measure associated with f. From Lemma 12 we can choose $r > 0$ such that $\mu\left(B\left(x, r\right)\right) \leq \frac{1}{10}$, for all $x \in \mathbb{R}^d$. Choose $s > 0$ such that $\mu\left(B\left(x, r+s\right)^C\right) \geq \frac{4}{5}$, for all $x \in \mathbb{R}^d$. The previous statement is satisfied by finding s such that, for all x, $\mu\left(B\left(x, r+s\right) \setminus B\left(x, r\right)\right) < \frac{1}{10}$, which is possible by Corollary 13 . By Lemma 6 we know

there exists x^* such that $g_\sigma^n(x^*) \geq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$ (x^* is implicitly a function of n). By the definition of Q_σ^n, let $g_\sigma^n = \sum_{i=1}^n \beta_i \Phi_\sigma(X_i)$ with $\beta_i \geq 0$ and $\sum_1^n \beta_i = 1$. Since $g_\sigma^n(x^*) \geq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$ and q is nonincreasing we have

$$\begin{aligned}
\frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 &\leq \sum_{i=1}^n \beta_i k_\sigma(X_i, x^*) \\
&= \sum_{i:X_i \in B(x^*,r)} \beta_i k_\sigma(X_i, x^*) + \sum_{j:X_j \in B(x^*,r)^C} \beta_j k_\sigma(X_j, x^*) \\
&= \sum_{i:X_i \in B(x^*,r)} \beta_i k_\sigma(X_i, x^*) + \sum_{j:X_j \in B(x^*,r)^C} \beta_j \sigma^{-d} q\left(\|X_j - x^*\|_2 / \sigma\right) \\
&\leq \sum_{i:X_i \in B(x^*,r)} \beta_i \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 + \sigma^{-d} q\left(r/\sigma\right)
\end{aligned}$$

The last line is due to the fact q must be nonincreasing (Scovel et al., 2010). From Lemma 14 we know that $\sigma^{-d}q(r/\sigma) \to 0$ as $\sigma \to 0$, so for sufficiently small σ we have

$$\frac{17}{20}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 < \sum_{i:X_i \in B(x^*,r)} \beta_i \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$$

and thus

$$\frac{17}{20} < \sum_{i:X_i \in B(x^*,r)} \beta_i. \tag{10}$$

Again, since q nonincreasing, for sufficiently small σ

$$\begin{aligned}
\sup_{y \in B(x^*, r+s)^C} g_\sigma^n(y) &= \sup_{y \in B(x^*, r+s)^C} \sum_{i=1}^n \beta_i k_\sigma\left(X_i, y\right) \\
&= \sup_{y \in B(x^*, r+s)^C} \sum_{i:X_i \in B(x^*,r)} \beta_i k_\sigma\left(X_i, y\right) \\
&+ \sum_{j:X_j \in B(x^*,r)^C} \beta_j \left\langle \Phi_\sigma\left(y\right), \Phi_\sigma\left(X_j\right)\right\rangle_{\mathcal{H}_\sigma} \\
&\leq \sigma^{-d} q\left(s/\sigma\right) + \sum_{j:X_j \in B(x^*,r)^C} \beta_j \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 .
\end{aligned}$$

From this, (10) and because $\sigma^{-d}q(s/\sigma) \to 0$ as $\sigma \to 0$, for arbitrary $\varepsilon > 0$ we have, for sufficiently small σ,

$$\sup_{y \in B(x^*, r+s)^C} g_\sigma^n(y) < \varepsilon + \frac{3}{20}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 .$$

Recall that we assumed that $\sup_x \psi(x) = \sup_x \rho'(x) = 1$ and $\rho(0) = 0$. Because ρ is strictly increasing, for sufficiently small σ,

$$\begin{aligned}
J_\sigma^n(g_\sigma^n) &= \frac{1}{n}\sum_{i=1}^n \rho\left(\|\Phi_\sigma(X_i) - g_\sigma^n\|_{\mathcal{H}_\sigma}\right) \\
&= \frac{1}{n}\sum_{i:X_i\in B(x^*,r+s)} \rho\left(\|\Phi_\sigma(X_i) - g_\sigma^n\|_{\mathcal{H}_\sigma}\right) + \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C} \rho\left(\|\Phi_\sigma(X_j) - g_\sigma^n\|_{\mathcal{H}_\sigma}\right) \\
&\geq \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C} \rho\left(\|\Phi_\sigma(X_j) - g_\sigma^n\|_{\mathcal{H}_\sigma}\right) \\
&= \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C} \rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 - 2g_\sigma^n(X_j) + \|g_\sigma^n\|_{\mathcal{H}_\sigma}^2}\right) \\
&\geq \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C} \rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 - 2\left(\frac{3}{20}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 + \varepsilon\right) + \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2}\right) \\
&= \mu_n\left(B(x^*, r+s)^C\right)\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20} - 2\varepsilon}\right) \\
&\geq \inf_x \mu_n\left(B(x, r+s)^C\right)\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20} - 2\varepsilon}\right).
\end{aligned}$$

Since ρ is strictly convex we know that ψ is strictly increasing. Because ψ has a supremum of 1 and is strictly increasing we know that for any $1 > \varepsilon_\psi > 0$ there exists b_ψ such that for all $x > b_\psi$, $\psi(x) > 1 - \varepsilon_\psi$. Then, for sufficiently small σ,

$$\begin{aligned}
\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20} - 2\varepsilon}\right) &= \int_0^{\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}} \psi(x)\,dx \\
&\geq \int_{b_\psi}^{\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}} \psi(x)\,dx \\
&\geq (1-\varepsilon_\psi)\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20} - 2\varepsilon} - b_\psi\right) \qquad (11)
\end{aligned}$$

For sufficiently small σ we have

$$\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20} - 2\varepsilon} \geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{32}{20}} - 2\varepsilon.$$

Since the complements of all open balls, in this case, all balls with radius $r+s$, have a finite shattering dimension (Devroye and Lugosi, 2001), and by our choice of r and s we know, with probability going to one, that $\inf_x \mu_n\left(B(x, r+s)^C\right) \to \inf_x \mu\left(B(x, r+s)^C\right) \geq 0.8$. Because of this for any $\varepsilon_B > 0$ we have, with probability going to one, that

$\inf_x \mu_n\left(B\left(x, r+s\right)^C\right) \geq 0.8 - \varepsilon_B$. Since $\frac{4}{5}\sqrt{\frac{32}{20}} > 1$, we can choose ε_ψ and ε_B such that $\left(\frac{4}{5} - \varepsilon_B\right)(1-\varepsilon_\psi)\sqrt{\frac{32}{20}} > 1$. Using these facts with (11) we have, for sufficiently small σ, with probability going to one

$$\begin{aligned} J_\sigma^n\left(g_\sigma^n\right) &\geq \inf_x \mu_n\left(B\left(x, r+s\right)^C\right)(1-\varepsilon_\psi)\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 \frac{32}{20} - 2\varepsilon} - b_\psi\right) \\ &\geq \left(\frac{4}{5} - \varepsilon_B\right)(1-\varepsilon_\psi)\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{32}{20}} - 2\varepsilon - b_\psi\right) \\ &> \|\Phi_\sigma\|_{\mathcal{H}_\sigma}. \end{aligned}$$

Now consider

$$\begin{aligned} J_\sigma^n(0) &= \frac{1}{n}\sum_{i=1}^n \rho\left(\|\Phi_\sigma(X_i) - 0\|_{\mathcal{H}_\sigma}\right) \\ &= \rho\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\right) \\ &= \int_0^{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} \psi(x)\,dx + \rho(0) \\ &\leq \int_0^{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} 1 dx \\ &= \|\Phi_\sigma\|_{\mathcal{H}_\sigma}. \end{aligned}$$

So as $n \to \infty$ and $\sigma \to 0$ we have

$$\mathbb{P}\left(J_\sigma^n(g_\sigma^n) \leq J_\sigma^n(0)\right) \to 0,$$

thus finishing the proof. ■

Proof of Lemma 9 Let $g, h \in \mathcal{H}_\sigma$ such that $\|g\|_{\mathcal{H}_\sigma}^2 \leq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$ and $\|h\|_{\mathcal{H}_\sigma}^2 \leq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$. Cross multiplication gives us

$$\begin{aligned} &\|R_\sigma^n(g) - R_\sigma^n(h)\|_{\mathcal{H}_\sigma} \\ &= \left\| \frac{\int \varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x) d\mu_n(x)}{\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)} - \frac{\int \varphi\left(\|\Phi_\sigma(x') - h\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x') d\mu_n(x')}{\int \varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right) d\mu_n(y')} \right\|_{\mathcal{H}_\sigma} \\ &= \left\| \frac{A}{B} \right\|_{\mathcal{H}_\sigma} \end{aligned}$$

where

$$\begin{aligned} A = &\left[\int \varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x)\,d\mu_n(x)\right]\left[\int \varphi\left(\left\|\Phi_\sigma(y') - h\right\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right] \\ &- \left[\int \varphi\left(\left\|\Phi_\sigma(x') - h\right\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x')\,d\mu_n(x')\right]\left[\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right] \end{aligned}$$

and

$$B = \left[\int \varphi\left(\left\|\Phi_\sigma(y') - h\right\|_{\mathcal{H}_\sigma}\right) d\mu_n(y')\right]\left[\int \varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right].$$

Note that $A \in \mathcal{H}_\sigma$ and $B \in \mathbb{R}^+$. We will now find a lower bound on B. As shown in the proof for Lemma 7 there exists $b > 0$ such that $\psi(x) > 1/2$ for all $x \geq b$. By the reverse triangle inequality

$$\begin{aligned}\left\|\Phi_\sigma(y') - h\right\|_{\mathcal{H}_\sigma} &\geq \left|\|\Phi_\sigma\|_{\mathcal{H}_\sigma} - \|h\|_{\mathcal{H}_\sigma}\right| \\ &\geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)\end{aligned}$$

which grows without bound as $\sigma \to 0$. So for sufficiently small σ

$$\begin{aligned}\varphi\left(\left\|\Phi_\sigma(y') - h\right\|_{\mathcal{H}_\sigma}\right) &= \frac{\psi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}} \\ &\geq \frac{1}{2\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}} \\ &\geq \frac{1}{2\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma} + \|h\|_{\mathcal{H}_\sigma}\right)} \\ &\geq \frac{1}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma} 2\left(1 + \sqrt{\frac{9}{10}}\right)}.\end{aligned}$$

A similar result can be shown for $\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)$, so there exists $C_B > 0$ such that, for sufficiently small σ,

$$B \geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2} C_B.$$

Now we will focus on A. To make the following manipulations simpler we will let

$$\varphi\left(\|\Phi_\sigma(z) - k\|_{\mathcal{H}_\sigma}\right) = T_\sigma(z, k).$$

A is equal to

$$\begin{aligned}&\left[\int T_\sigma(x, g)\,\Phi_\sigma(x) d\mu_n(x)\right]\left[\int T_\sigma\left(y', h\right) d\mu_n(y')\right] - \left[\int T_\sigma\left(x', h\right)\Phi_\sigma(x') d\mu_n(x')\right]\left[\int T_\sigma(y, g)\, d\mu_n(y)\right] \\ &= \int\left\{T_\sigma(x, g)\,\Phi_\sigma(x)\left[\int T_\sigma\left(y', h\right) d\mu_n(y')\right] - T_\sigma(x, h)\,\Phi_\sigma(x)\left[\int T_\sigma(y, g)\, d\mu_n(y)\right]\right\} d\mu_n(x) \\ &= \int \Phi_\sigma(x)\left[T_\sigma(x, g)\left[\int T_\sigma\left(y', h\right) d\mu_n(y)\right] - T_\sigma(x, h)\left[\int T_\sigma(y, g)\, d\mu_n(y)\right]\right] d\mu_n(x) \\ &= \int\int \Phi_\sigma(x)\left[T_\sigma(x, g)\, T_\sigma(y, h) - T_\sigma(x, h)\, T_\sigma(y, g)\right] d\mu_n(y) d\mu_n(x).\end{aligned}$$

We will now bound the inner term. Using the triangle inequality we have

$$\begin{aligned}
&\left|T_\sigma(x,g)\,T_\sigma(y,h) - T_\sigma(x,h)\,T_\sigma(y,g)\right| \qquad (12)\\
&< \left|T_\sigma(x,g)\,T_\sigma(y,h) - T_\sigma(x,g)\,T_\sigma(y,g)\right| + \left|T_\sigma(x,g)\,T_\sigma(y,g) - T_\sigma(x,h)\,T_\sigma(y,g)\right|\\
&= T_\sigma(x,g)\left|T_\sigma(y,h) - T_\sigma(y,g)\right| + T_\sigma(y,g)\left|T_\sigma(x,g) - T_\sigma(x,h)\right|.
\end{aligned}$$

We will bound the second summand in the last equality; a similar technique can bound the first summand.

$$\begin{aligned}
\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) &= \frac{\psi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}}\\
&\le \frac{1}{\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}}\\
&\le \frac{1}{\left|\|\Phi_\sigma\|_{\mathcal{H}_\sigma} - \|g\|_{\mathcal{H}_\sigma}\right|}\\
&\le \frac{1}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)}. \qquad (13)
\end{aligned}$$

A similar result can be shown for $\varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)$.

Consider $z \ge \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)$, then

$$\begin{aligned}
\left|\varphi'(z)\right| &= \left|\left(\frac{\psi(z)}{z}\right)'\right|\\
&= \left|\frac{z\psi'(z) - \psi(z)}{z^2}\right|\\
&\le \frac{|z\psi'(z)| + |\psi(z)|}{z^2}.
\end{aligned}$$

We will now analyze the behaviour of ψ', specifically, there exists sufficiently large r such that $\psi'(x) \le \frac{1}{x}$ for all $x \ge r$. We will proceed by contradiction. Suppose this is not the case. Then there exist positive numbers t_1, t_2 and t_3 such that $\psi'(t_i) > \frac{1}{t_i}$ and $\frac{t_i}{t_{i+1}} < \frac{1}{3}$. We know ψ' is nonincreasing by (B5) and nonnegative; we also know ψ is bounded above by 1 so

$$\begin{aligned}
1 &\ge \int_0^\infty \psi'(x)\,dx\\
&\ge \int_{t_1}^{t_2} \psi'(x)\,dx + \int_{t_2}^{t_3} \psi'(y)\,dy\\
&\ge \frac{t_2 - t_1}{t_2} + \frac{t_3 - t_2}{t_3}\\
&\ge 2 - \frac{2}{3},
\end{aligned}$$

a contradiction. From this we have that for sufficiently large z,

$$\begin{aligned}\frac{|z\psi'(z)| + |\psi(z)|}{z^2} &\leq \frac{z\frac{1}{z} + 1}{z^2} \\ &= \frac{2}{z^2}.\end{aligned}$$

Thus, for sufficiently small σ, on the space $\left[\left(1 - \sqrt{\frac{9}{10}}\right)\|\Phi_\sigma\|_{\mathcal{H}_\sigma}, \infty\right)$, φ is Lipschitz continuous with Lipschitz constant $2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}$. Therefore we have

$$\begin{aligned}|\varphi(\|\Phi_\sigma(x) - g\|) - \varphi(\|\Phi_\sigma(x) - h\|)| \leq& \left|\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma} - \|\Phi_\sigma(x) - h\|_{\mathcal{H}_\sigma}\right| 2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2} \\ &\leq \|g - h\|_{\mathcal{H}_\sigma} 2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}.\end{aligned}$$

Combining the last inequality with (13) we have that for sufficiently small σ, (12) is less than or equal to

$$4\|g - h\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)^{-3}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}.$$

Using this bound we can do the following. Let $\tau \triangleq \left[T_\sigma(x, g)T_\sigma(y, h) - T_\sigma(x, h)T_\sigma(y, g)\right]$, $\tau' \triangleq \left[T_\sigma(x', g)T_\sigma(y', h) - T_\sigma(x', h)T_\sigma(y', g)\right]$, and $\kappa \triangleq 4\|g - h\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)^{-3}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}$, we have

$$\begin{aligned}\|A\|_{\mathcal{H}_\sigma}^2 &= \left\|\int\int \Phi_\sigma(x)\tau d\mu_n(x)d\mu_n(y)\right\|_{\mathcal{H}_\sigma}^2 \\ &= \left\langle \int\int \Phi_\sigma(x)\tau d\mu_n(x)d\mu_n(y), \int\int \Phi_\sigma(x')\tau' d\mu_n(x')d\mu_n(y')\right\rangle_{\mathcal{H}_\sigma} \\ &= \int\int\int\int \tau\tau' \left\langle \Phi_\sigma(x), \Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(y)d\mu_n(y')d\mu_n(x)d\mu_n(x').\end{aligned}$$

Since $\langle \Phi_\sigma(x), \Phi_\sigma(x')\rangle_{\mathcal{H}_\sigma} \geq 0$ for all x, x', for sufficiently small σ, the last line is less than or equal to

$$\begin{aligned}&\int\int\int\int \kappa^2 \left\langle \Phi_\sigma(x), \Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(y)d\mu_n(y')d\mu_n(x)d\mu_n(x') \\ &= \int\int \kappa^2 \left\langle \Phi_\sigma(x), \Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(x)d\mu_n(x') \\ &= \kappa^2 \left\|\int \Phi_\sigma(x)\, d\mu_n(x)\right\|_{\mathcal{H}_\sigma}^2.\end{aligned}$$

Returning to the original notation, this means, for sufficiently small σ

$$\|A\|_{\mathcal{H}_\sigma} \leq \left\| \int \Phi_\sigma(x)\, d\mu_n(x) \right\|_{\mathcal{H}_\sigma} 4\|g-h\|_{\mathcal{H}_\sigma} \left(1-\sqrt{\frac{9}{10}}\right)^{-3} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}.$$

From our proof of the consistency of the KDE we know that $\left\|\int \Phi_\sigma(x) d\mu_n(x) - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$ and from Lemma 8 $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2$ so $\left\|\int \Phi_\sigma(x) d\mu_n(x)\right\|_{\mathcal{H}_\sigma}$ is bounded by some constant with probability going to one. Note that this is the only probabilistic step, which does not depend on g or h, so the result holds over the whole ball in $\mathcal{H}_\sigma$. So there exists $C_A > 0$ such that

$$\|A\|_{\mathcal{H}_\sigma} \leq \|g-h\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3} C_A$$

with probability going to one (we can omit "for sufficiently small σ" since $\sigma \to 0$ as $n \to \infty$). Finally we get with probability going to one as $n\sigma^d \to \infty$

$$\begin{aligned}\left\|\frac{A}{B}\right\|_{\mathcal{H}_\sigma} &= \frac{\|A\|_{\mathcal{H}_\sigma}}{B} \\ &\leq \|g-h\|_{\mathcal{H}_\sigma} \frac{C_A \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}}{C_B \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}} \\ &= \|g-h\|_{\mathcal{H}_\sigma} C_R \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1}.\end{aligned}$$

■

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 1950.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complex.*, 23(1):52–72, February 2007. ISSN 0885-064X. doi: 10.1016/j.jco.2006.07.001. URL http://dx.doi.org/10.1016/j.jco.2006.07.001.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

P. Deheuvels. Uniform limit laws for kernel density estimators on possibly unbounded interval. In *Recent Advances in Reliability Theory*, pages 477–492. Birkhäuser, 2000.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

U. Einmahl and D. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.*, 13:1–37, 2000.

E. Gine, V. Koltchinskii, and J. Zinn. Weighted uniform consistency of kernel density estimators. *Ann. Probab.*, 32:2570–2605, 2004.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002. ISSN 0246-0203. doi: 10.1016/S0246-0203(02)01128-7.

J. Kim and C. Scott. Robust kernel density estimation. *J. Machine Learning Res.*, 13: 2529–2565, 2012.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):pp. 1679–1706, 1994.

D. W. Scott. *Multivariate Density Estimation.* Wiley, New York, 1992.

Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler. Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, 26(6):641 – 660, 2010. ISSN 0885-064X. doi: 10.1016/j.jco.2010.03.002.

B W Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1), 1978.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007. ISSN 0176-4276. 10.1007/s00365-006-0659-y.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

W. Stute. A law of the logarithm for kernel density estimators. *Ann. Probab.*, 10:414–422, 1982.

D. Wied and R. Weissbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53:1–21, 2012.