Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 715683, 13 pages
http://dx.doi.org/10.1155/2013/715683

Research Article

Density Problem and Approximation Error in Learning Theory

Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

Correspondence should be addressed to Ding-Xuan Zhou; mazhou@cityu.edu.hk

Received 3 May 2013; Accepted 5 August 2013

Academic Editor: Yiming Ying

We study the density problem and approximation error of reproducing kernel Hilbert spaces for the purpose of learning theory. For a Mercer kernel K on a compact metric space (X, d), a characterization for the generated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ to be dense in $C(X)$ is given. As a corollary, we show that the density is always true for convolution type kernels. Some estimates for the rate of convergence of interpolation schemes are presented for general Mercer kernels. These are then used to establish for convolution type kernels quantitative analysis for the approximation error in learning theory. Finally, we show by the example of Gaussian kernels with varying variances that the approximation error can be improved when we adaptively change the value of the parameter for the used kernel. This confirms the method of choosing varying parameters which is used often in many applications of learning theory.

1. Introduction

Learning theory investigates how to find function relations or data structures from random samples. For the regression problem, one usually has some experience and would expect that the (underlying) unknown function lies in some set of functions $\mathcal{H}$ called the *hypothesis space*. Then one tries to find a good approximation in $\mathcal{H}$ of the underlying function f (under certain metric). The best approximation in $\mathcal{H}$ is called the *target function* $f_{\mathcal{H}}$. However, f is unknown. What we have in hand is a set of random samples $\{(x_i, y_i)\}_{i=1}^{\ell}$. These samples are not given by f exactly $(f(x_i) \neq y_i)$. They are controlled by this underlying function f with noise or some other uncertainties $(f(x_i) \approx y_i)$. The most important model studied in learning theory [1] is to assume that the uncertainty is represented by a Borel probability measure ρ on $X \times Y$, and the underlying function $f : X \to Y$ is the regression function of ρ given by

$$f_\rho(x) = \int_Y y \, d\rho(y \mid x), \quad x \in X. \tag{1}$$

Here, $\rho(y \mid x)$ is the conditional probability measure at x. Then, the samples $\{(x_i, y_i)\}_{i=1}^{\ell}$ are independent and identically distributed drawers according to the probability measure ρ. For the classification problem, $Y = \{1, -1\}$ and $\operatorname{sign}(f_\rho)$ is the optimal classifier.

Based on the samples, one can find a function from the hypothesis space $\mathcal{H}$ that best fits the data $\mathbf{z} := \{(x_i, y_i)\}_{i=1}^{\ell}$ (with respect to certain loss functional). This function is called the *empirical target function* $f_{\mathbf{z}}$. When the number ℓ of samples is large enough, $f_{\mathbf{z}}$ is a good approximation of the target function $f_{\mathcal{H}}$ with certain confidence. This problem has been extensively investigated and well developed in the literature of statistical learning theory. See, for example, [1–4].

What is less understood is the approximation of the underlying desired function f by the target function $f_{\mathcal{H}}$. For example, if one takes $\mathcal{H}$ to be a polynomial space of some fixed degree, then f can be approximated by functions from $\mathcal{H}$ only when f is a polynomial in $\mathcal{H}$.

In kernel machine learning such as support vector machines, one often uses reproducing kernel Hilbert spaces or their balls as hypothesis spaces. Here, we take $(X, d(\cdot, \cdot))$ to be a compact metric space and $Y = \mathbb{R}$.

Definition 1. Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric, and positive semidefinite; that is, for any finite set of distinct points $\{x_1, \ldots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite. Such a kernel is called a *Mercer kernel*. It is called

positive definite if the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is always positive definite.

The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with a Mercer kernel K is defined (see [5]) to be the completion of the linear span of the set of functions $\{K_x := K(x,\cdot) : x \in X\}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K} = \langle\cdot,\cdot\rangle_K$ satisfying

$$\left\|\sum_{i=1}^{\ell} c_i K_{x_i}\right\|_K^2 = \left\langle \sum_{i=1}^{\ell} c_i K_{x_i}, \sum_{i=1}^{\ell} c_i K_{x_i} \right\rangle_K = \sum_{i,j=1}^{\ell} c_i K\left(x_i, x_j\right) c_j. \tag{2}$$

The reproducing kernel property is given by

$$\left\langle K_x, g\right\rangle_K = g(x), \quad \forall x \in X,\ g \in \mathcal{H}_K. \tag{3}$$

This space can be embedded into $C(X)$, the space of continuous functions on X.

In kernel machine learning, one often takes $\mathcal{H}_K$ or its balls as the hypothesis space. Then, one needs to know whether the desired function f can be approximated by functions from the RKHS.

The first purpose of this paper is to study the density of the reproducing kernel Hilbert spaces in $C(X)$ (or in $L^2(X)$ when X is a subset of the Euclidean space $\mathbb{R}^n$). This will be done in Section 2 where some characterizations will be provided. Let us mention a simple example with detailed proof given in Section 6.

Example 2. Let $X = [0,1]$ and let K be a Mercer kernel given by

$$K(x,y) = \sum_{j=0}^{+\infty} a_j (x \cdot y)^j, \tag{4}$$

where $a_j \geq 0$ for each j and $\sum_{j=0}^{+\infty} a_j < \infty$. Set $J := \{j \in \mathbb{Z}_+ : a_j > 0\}$. Then, $\mathcal{H}_K$ is dense in $C(X)$ if and only if

$$a_0 > 0, \qquad \sum_{j \in J\setminus\{0\}} \frac{1}{j} = +\infty. \tag{5}$$

When the density holds, we want to study the convergence rate of the approximation by functions from balls of the RKHS as the radius tends to infinity. The quantity

$$I(f,R) := \inf_{\|g\|_K \leq R} \|f - g\| \tag{6}$$

is called the *approximation error* in learning theory. Some estimates have been presented by Smale and Zhou [6] for the L^2 norm and many kernels. The second purpose of this paper is to investigate the convergence rate of the approximation error with the uniform norm as well as the L^2 norm. Estimates will be given in Section 4, based on the analysis in Section 3 for interpolation schemes associated with general Mercer kernels. With this analysis, we can understand the approximation error with respect to marginal probability distribution induced by ρ. Let us provide an example of Gaussian kernels to illustrate the idea. Notice that when the parameter σ of the kernel is allowed to change with R, the rate of the approximation error may be improved. This confirms the method of adaptively choosing the parameter of the kernel, which is used in many applications (see e.g., [7]).

Example 3. Let

$$K_\sigma(x,y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}, \quad x, y \in X = [0,1]^n. \tag{7}$$

There exist positive constants A and B such that, for each $f \in H^s(\mathbb{R}^n)$ and $R \geq A\|f\|_{L^2}$, there holds

$$\inf_{\|g\|_{K_\sigma} \leq R} \|f - g\|_{L^2(X)} \leq B(\log R)^{-s/4} \tag{8}$$

when σ is fixed; while when σ may change with R, there holds

$$\inf_{\|g\|_{K_{\sigma_R}} \leq R} \|f - g\|_{L^2(X)} \leq B(\log R)^{-s}. \tag{9}$$

2. Density and Positive Definiteness

The density problem of reproducing kernel Hilbert spaces in $C(X)$ was raised to the author by Poggio et al. See [8]. It can be stated as follows.

Given a Mercer kernel K on a compact metric space $(X, d(\cdot,\cdot))$, when is the RKHS $\mathcal{H}_K$ dense in $C(X)$?

By means of the dual space of $C(X)$, we can give a general characterization. This is only a simple observation, but it does provide us useful information. For example, we will show that the density is always true for convolution type kernels. Also, for dot product type kernel, we can give a complete nice characterization for the density, which will be given in Section 6.

Recall the Riesz Representation Theorem asserting that the dual space of $C(X)$ can be represented by the set of Borel measures on X. For a Borel measure μ on X, we define the integral operator $L_{K,\mu}$ associated with the kernel as

$$L_{K,\mu}(f)(x) := \int_X K(x,y) f(y)\, d\mu(y), \quad x \in X. \tag{10}$$

This is a compact operator on $L^2_\mu(X)$ if μ is a positive measure.

Theorem 4. *Let K be a Mercer kernel on a compact metric space (X, d). Then, the following statements are equivalent.*

(1) *$\mathcal{H}_K$ is dense in $C(X)$.*

(2) *For any nontrivial positive Borel measure μ, $\mathcal{H}_K$ is dense in $L^2_\mu(X)$.*

(3) *For any nontrivial positive Borel measure μ, $L_{K,\mu}$ has no eigenvalue zero in $L^2_\mu(X)$.*

(4) *For any nontrivial Borel measure μ, as a function in $C(X)$,*

$$\int_X K(\cdot, y)\, d\mu(y) \neq 0. \tag{11}$$

Proof. (1) $\Rightarrow$ (2). This follows from the fact that $C(X)$ is dense in $L^2_\mu(X)$. See, for example, [9].

(2) $\Rightarrow$ (3). Suppose that $\mathscr{H}_K$ is dense in $L^2_\mu(X)$, but $L_{K,\mu}$ has an eigenvalue zero in $L^2_\mu(X)$. Then, there exists a nontrivial function $f \in L^2_\mu(X)$ such that $L_{K,\mu}(f) = 0$; that is,

$$\begin{aligned} L_{K,\mu}(f)(x) &= \int_X K(x,y) f(y)\, d\mu(y) \\ &= \int_X K_x(y) f(y)\, d\mu(y) = 0. \end{aligned} \tag{12}$$

The identity holds as functions in $L^2_\mu(X)$. If the support of μ is X, then this identity would imply that f is orthogonal to each K_x with $x \in X$. When the support of μ is not X, things are more complicated. Here, the support of μ, denoted as supp μ, is defined to be the smallest closed subset F of X satisfying $\mu(X \setminus F) = 0$.

The property of the RKHS enables us to prove the general case. As the function $L_{K,\mu}(f)$ is continuous, we know from (12) that, for each x in supp μ,

$$\int_{\text{supp }\mu} K_x(y) f(y)\, d\mu(y) = \int_X K_x(y) f(y)\, d\mu(y) = 0. \tag{13}$$

This means for each x in supp μ, $f \perp K_x$ in $L^2_\mu(\text{supp } X)$, where μ has been restricted onto supp μ. When we restrict K onto supp $\mu \times$ supp μ, the new kernel $\widetilde{K}$ is again a Mercer kernel. Moreover, by (1), $\mathscr{H}_{\widetilde{K}} = \mathscr{H}_K|_{\text{supp }\mu}$. It follows that $\text{span}\{K_x|_{\text{supp }\mu} : x \in \text{supp } \mu\}$ is dense in $\mathscr{H}_{\widetilde{K}} = \mathscr{H}_K|_{\text{supp }\mu}$. The latter is dense in $L^2_\mu(\text{supp } \mu)$. Therefore, f is orthogonal to $L^2_\mu(\text{supp } \mu)$; hence, as a function in $L^2_\mu(X)$, f is zero. This is a contradiction.

(3) $\Rightarrow$ (4). Every nontrivial Borel measure μ can be uniquely decomposed as the difference of two mutually singular positive Borel measures: $\mu = \mu^+ - \mu^-$; that is, there exists a Borel set $A \subset X$ such that $\mu^+(A) = \mu^+(X)$ and $\mu^-(A) = 0$. With this decomposition,

$$\begin{aligned} \int_X K(x,y)\, d\mu(y) &= \int_X K(x,y) \{\chi_A(y) - \chi_{X\setminus A}(y)\}\, d|\mu| \\ &= L_{K,|\mu|}(\chi_A - \chi_{X\setminus A})(x). \end{aligned} \tag{14}$$

Here, χ_A is the characteristic function of the set A, and $|\mu| = \mu^+ + \mu^-$ is the absolute value of μ. As $|\mu|$ is a nontrivial positive Borel measure and $\chi_A - \chi_{X\setminus A}$ is a nontrivial function in $L^2_{|\mu|}(X)$, statement (3) implies that, as a function in $L^2_{|\mu|}(X)$, $\int_X K(x,y) d\mu(y) \neq 0$. Since this function lies in $C(X)$, it is nonzero as a function in $C(X)$.

The last implication (4) $\Rightarrow$ (1) follows directly from the Riesz Representation Theorem. □

The proof of Theorem 4 also yields a characterization for the density of the RKHS in $L^2_\mu(X)$.

Corollary 5. *Let K be a Mercer kernel on a compact metric space (X, d) and μ a positive Borel measure on X. Then, $\mathscr{H}_K$ is dense in $L^2_\mu(X)$ if and only if $L_{K,\mu}$ has no eigenvalue zero in $L^2_\mu(X)$.*

The necessity has been verified in the proof of Theorem 4, while the sufficiency follows from the observation that an $L^2_\mu(X)$ function f lying in the orthogonal complement of $\text{span}\{K_x : x \in X\}$ gives an eigenfunction of $L_{K,\mu}$ with eigenvalue zero:

$$\langle K_x, f\rangle_{L^2_\mu(X)} = \int_X K_x(y) f(y)\, d\mu(y) = L_{K,\mu}(f)(x) = 0. \tag{15}$$

Theorem 4 enables us to conclude that the density always holds for convolution type kernels $K(x,y) = k(x-y)$ with $k \in L^2(\mathbb{R}^n)$. The density for some convolution type kernels has been verified by Steinwart [10]. The author observed the density as a corollary of Theorem 4 when $\widehat{k}(\xi)$ is strictly positive. Charlie Micchelli pointed out to the author that, for a convolution type kernel, the RKHS is always dense in $C(X)$. So, the density problem is solved for these kernels.

Corollary 6. *Let $K(x,y) = k(x-y)$ be a nontrivial convolution type Mercer kernel on $\mathbb{R}^n$ with $k \in L^2(\mathbb{R}^n)$. Then, for any compact subset X of $\mathbb{R}^n$, $\mathscr{H}_K$ on X is dense in $C(X)$.*

Proof. It is well known that K is a Mercer kernel if and only if k is continuous and $\widehat{k}(\xi) \geq 0$ almost everywhere. We apply the equivalent statement (4) of Theorem 4 to prove our statement.

Let μ be a Borel measure on X such that

$$\int_X K(x,y)\, d\mu(y) = 0, \quad \forall x \in X. \tag{16}$$

Then, the inverse Fourier transform yields

$$\begin{aligned} &\int_X \int_{\mathbb{R}^n} \widehat{k}(\xi) e^{i\xi\cdot(x-y)} d\xi d\mu(y) \\ &\quad = \int_{\mathbb{R}^n} \widehat{k}(\xi)\, \widehat{\mu}(\xi)\, e^{i\xi\cdot x} d\xi = 0, \quad \forall x \in X. \end{aligned} \tag{17}$$

Here, $\widehat{\mu}(\xi) = \int e^{-i\xi\cdot y} d\mu(y)$ is the Fourier transform of the Borel measure μ, which is an entire function.

Taking the integral on X with respect to the measure μ, we have

$$\int_{\mathbb{R}^n} \widehat{k}(\xi)\, \widehat{\mu}(\xi) \int_X e^{i\xi\cdot x} d\mu(x)\, d\xi = \int_{\mathbb{R}^n} \widehat{k}(\xi) \left|\widehat{\mu}(\xi)\right|^2 d\xi = 0. \tag{18}$$

For a nontrivial Borel measure μ supported on X, $\widehat{\mu}(\xi)$ vanishes only on a set of measure zero. Hence, $\widehat{k}(\xi) = 0$ almost everywhere, which gives $k = 0$. Therefore, we must have $\mu = 0$. This proves the density by Theorem 4. □

After the first version of the paper was finished, I learned that Micchelli et al. [11] proved the density for a class of

convolution type kernels $k(x-y)$ with k being the Fourier transform of a finite Borel measure. Note that a large family of convolution type reproducing kernels are given by radial basis functions; see, for example, [12].

Now we can state a trivial fact that the positive definiteness is a necessary condition for the density.

Corollary 7. *Let K be a Mercer kernel on a compact metric space (X,d). If $\mathscr{H}_K$ is dense in $C(X)$, then K is positive definite.*

Proof. Suppose to the contrary that $\mathscr{H}_K$ is dense in $C(X)$, but there exists a finite set of distinct points $\{x_i\}_{i=1}^{\ell} \subset X$ such that the matrix $A_{\mathbf{x}} := (K(x_i,x_j))_{i,j=1}^{\ell}$ is not positive definite. By the Mercer kernel property, $A_{\mathbf{x}}$ is positive semidefinite. So it is singular, and we can find a nonzero vector $c := (c_1,\ldots,c_\ell)^T \in \mathbb{R}^\ell$ satisfying $A_{\mathbf{x}}c = 0$. It follows that $c^T A_{\mathbf{x}} c = 0$; that is,

$$\left\| \sum_{i=1}^{\ell} c_i K_{x_i} \right\|_K^2 = \left\langle \sum_{i=1}^{\ell} c_i K_{x_i}, \sum_{j=1}^{\ell} c_j K_{x_j} \right\rangle_K = \sum_{i,j=1}^{\ell} c_i c_j K\left(x_i,x_j\right) = 0. \tag{19}$$

Thus,

$$\sum_{i=1}^{\ell} c_i K_{x_i} = 0. \tag{20}$$

Now, we define a nontrivial Borel measure μ supported on $\{x_1,\ldots,x_\ell\}$ as

$$\mu\left(\{x_i\}\right) = c_i, \quad i = 1,\ldots,\ell. \tag{21}$$

Then, for $x \in X$,

$$\int_X K(x,y)\, d\mu(y) = \sum_{i=1}^{\ell} K(x,x_i)\, c_i = \sum_{i=1}^{\ell} c_i K_{x_i}(x) = 0. \tag{22}$$

This is a contradiction to the equivalent statement (4) in Theorem 4 of the density. □

Because of the necessity given in Corollary 7, one would expect that the positive definiteness is also sufficient for the density. Steve Smale convinced the author that this is not the case in general. This motivates us to present a constructive example of C^∞ kernel. Denote $\|g\|_{W_\infty^m} := \sum_{j=0}^{m} \|g^{(j)}\|_{L^\infty}$ as the norm in the Sobolev space W_∞^m.

Example 8. Let $X = [0,1]$. For every $m \in \mathbb{N}$ and every $i \in \{0,1,\ldots,m\}$, choose a real-valued C^∞ function $\psi_{i,m}(x)$ on $[0,1]$ such that

$$\psi_{i,m}(x) = x^i, \quad \forall x \in [0,1] \setminus \left(\frac{1}{m+1}, \frac{1}{m}\right), \qquad \int_0^1 \psi_{i,m}(x)\, dx = 0. \tag{23}$$

Define K on $[0,1]\times[0,1]$ by

$$K(x,y) = \sum_{m=1}^{\infty} 2^{-m} \frac{\sum_{i=0}^{m} \psi_{i,m}(x)\,\psi_{i,m}(y)}{\sum_{i=0}^{m} \|\psi_{i,m}\|_{W_\infty^m}^2}, \quad x,y \in [0,1]. \tag{24}$$

Then, K is a C^∞ Mercer kernel on $[0,1]$. It is positive definite, but the constant function 1 is not in the closure of $\mathscr{H}_K$ in $C(X)$. Hence, $\mathscr{H}_K$ is not dense in $C(X)$.

Proof. The series in (24) converges in W_∞^m for any $m \in \mathbb{N}$. Hence, K is C^∞ and is a Mercer kernel on $[0,1]$.

To prove the positive definiteness, we let $\{x_i\}_{i=1}^{\ell} \subset [0,1]$ be a finite set of distinct points and $(c_i)_{i=1}^{\ell}$ a nonzero vector. Choose $m \in \mathbb{N}$ such that

$$m \geq \ell - 1, \qquad \frac{1}{m} < \min\left\{x_j :\ x_j > 0, j \in \{1,\ldots,\ell\}\right\}. \tag{25}$$

Then, for each $j \in \{1,\ldots,\ell\}$, either $x_j = 0$ or $x_j > 1/m$. Hence,

$$x_j \in [0,1] \setminus \left(\frac{1}{m+1}, \frac{1}{m}\right), \quad \forall j = 1,\ldots,\ell. \tag{26}$$

By the construction of $\psi_{i,m}$, there holds

$$\psi_{i,m}\left(x_j\right) = x_j^i, \quad \forall i = 0,1,\ldots,m,\ \ j = 1,\ldots,\ell. \tag{27}$$

Then,

$$\begin{aligned} &\sum_{i,j=1}^{\ell} c_i c_j K\left(x_i,x_j\right) \\ &\geq \frac{2^{-m}}{\sum_{i=0}^{m} \|\psi_{i,m}\|_{W_\infty^m}^2} \sum_{i=0}^{m} \left[\sum_{j=1}^{\ell} c_j \psi_{i,m}\left(x_j\right)\right]^2 \\ &\geq \frac{2^{-m}}{\sum_{i=0}^{m} \|\psi_{i,m}\|_{W_\infty^m}^2} \sum_{i=0}^{\ell-1} \left[\sum_{j=1}^{\ell} c_j \psi_{i,m}\left(x_j\right)\right]^2. \end{aligned} \tag{28}$$

Now, the determinant of the matrix $(x_j^i)_{i=0,1,\ldots,\ell-1,\, j=1,\ldots,\ell}$ is a Vandermonde determinant and is nonzero. Since $(c_j)_{j=1}^{\ell}$ is a nonzero vector, we know that $\sum_{j=1}^{\ell} c_j x_j^i \neq 0$ for some $i \in \{0,1,\ldots,\ell-1\}$. It follows that $\sum_{i,j=1}^{\ell} c_i c_j K(x_i,x_j) > 0$. Thus, K is positive definite.

We now prove that **1**, the constant function taking the value 1 everywhere, is not in the closure of $\mathscr{H}_K$ in $C(X)$. In fact, the uniformly convergent series (24) and the vanishing property of $\psi_{i,m}$ imply that

$$\int_0^1 K(x,y)\, dy = \int_0^1 K_x(y)\, dy = 0, \quad \forall x \in X. \tag{29}$$

Since $\operatorname{span}\{K_x : x \in X\}$ is dense in $\mathscr{H}_K$ and $\mathscr{H}_K$ is embedded in $C(X)$, we know that

$$\int_0^1 f(y)\, dy = 0, \quad \forall f \in \mathscr{H}_K. \tag{30}$$

If $\mathbf{1}$ could be uniformly approximated by a sequence $\{f_m\}$ in $\mathscr{H}_K$, then

$$1 = \int_0^1 \mathbf{1}(y)\,dy = \lim_{m\to\infty} \int_0^1 f_m(y)\,dy = 0, \tag{31}$$

which would be a contradiction. Therefore, $\mathscr{H}_K$ is not dense in $C(X)$. □

Combining the previous discussion, we know that the positive definiteness is a nice necessary condition for the density of the RKHS in $C(X)$ but is not sufficient.

3. Interpolation Schemes for Reproducing Kernel Spaces

The study of approximation by reproducing kernel Hilbert spaces has a long history; see, for example, [13, 14]. Here, we want to investigate the rate of approximation as the RKHS norm of the approximant becomes large.

In the following sections, we consider the approximation error for the purpose of learning theory. The basic tool for constructing approximants is a set of nodal functions used in [6, 15, 16].

Definition 9. We say that $\{u_i(x) := u_{i,\mathbf{x}}(x)\}_{i=1}^{\ell}$ is the set of *nodal functions* associated with the nodes $\mathbf{x} := \{x_1, \ldots, x_\ell\} \subset X$ if $u_i \in \operatorname{span}\{K_{x_j}\}_{j=1}^{\ell}$ and

$$u_i(x_j) = \delta_{ij} = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

The nodal functions have some nice minimization properties; see [6, 16].

In [15], we show that the nodal functions $\{u_i\}_{i=1}^{\ell}$ associated with $\mathbf{x}$ exist if and only if the Gramian matrix $A_{\mathbf{x}} := (K(x_i, x_j))_{i,j=1}^{\ell}$ is nonsingular. In this case, the nodal functions are uniquely given by

$$u_i(x) = \sum_{j=1}^{\ell} \left(A_{\mathbf{x}}^{-1}\right)_{i,j} K_{x_j}(x), \quad i = 1, \ldots, \ell. \tag{33}$$

Remark 10. When the RKHS has finite dimension m, then, for any $\ell \le m$ we can find nodal functions $\{u_j\}_{j=1}^{\ell}$ associated with some subset $\mathbf{x} = \{x_1, \ldots, x_\ell\} \subset X$, while for $\ell > m$, no such nodal functions exist. When $\dim\mathscr{H}_K = \infty$, then, for any $\ell \in \mathbb{N}$, we can find a subset $\mathbf{x} = \{x_1, \ldots, x_\ell\} \subset X$ which possesses a set of nodal functions.

The nodal functions are used to construct an interpolation scheme:

$$I_{\mathbf{x}}(f)(x) = \sum_{i=1}^{\ell} f(x_i)\, u_{i,\mathbf{x}}(x), \quad x \in X,\ f \in C(X). \tag{34}$$

It satisfies $I_{\mathbf{x}}(f)(x_i) = f(x_i)$ for $i = 1, \ldots, \ell$. Interpolation schemes have been applied to the approximation by radial basis functions in the vast literature; see, for example, [17–20].

The error $I_{\mathbf{x}}(f) - f$ for $f \in \mathscr{H}_K$ will be estimated by means of a power function.

Definition 11. Let K be a Mercer kernel on a compact metric space (X, d) and $\mathbf{x} = \{x_1, \ldots, x_\ell\} \subset X$. The power function ε_K is defined on $\mathbf{x}$ as

$$\varepsilon_K(\mathbf{x}) := \sup_{x\in X} \left\{ \inf_{w\in\mathbb{R}^\ell} \left\{ K(x,x) - 2\sum_{i=1}^{\ell} w_i K(x, x_i) + \sum_{i,j=1}^{\ell} w_i K(x_i, x_j)\, w_j \right\}^{1/2} \right\}. \tag{35}$$

We know that $\varepsilon_K(\mathbf{x}) \to 0$ when $d_{\mathbf{x}} := \max_{x\in X} \min_{i=1,\ldots,\ell} d(x, x_i) \to 0$. If K is Lipschitz s on X:

$$\left|K(x,y) - K(x,t)\right| \le C(d(y,t))^s, \tag{36}$$

then

$$\varepsilon_K(\mathbf{x}) \le 2C d_{\mathbf{x}}^s. \tag{37}$$

Moreover, higher order regularity of K implies faster convergence of $\varepsilon_K(\mathbf{x})$. For details, see [16].

The error of the interpolation scheme for functions from RKHS can be estimated as follows.

Theorem 12. *Let K be a Mercer kernel and $A_{\mathbf{x}}$ nonsingular for a finite set $\mathbf{x} = \{x_1, \ldots, x_\ell\} \subset X$. Define the interpolation scheme associated with $\mathbf{x}$ as* (34)*. Then, for $f \in \mathscr{H}_K$, there holds*

$$\left\|I_{\mathbf{x}}(f) - f\right\|_{C(X)} \le \|f\|_K \varepsilon_K(\mathbf{x}), \tag{38}$$

$$\left\|I_{\mathbf{x}}(f)\right\|_K \le \|f\|_K. \tag{39}$$

Proof. Let $x \in X$. We apply the reproducing property (3) of the function f in

$$I_{\mathbf{x}}(f)(x) - f(x) = \sum_{i=1}^{\ell} f(x_i)\, u_i(x) - f(x). \tag{40}$$

Then,

$$I_{\mathbf{x}}(f)(x) - f(x) = \sum_{i=1}^{\ell} u_i(x) \left\langle K_{x_i}, f\right\rangle_K - \left\langle K_x, f\right\rangle_K = \left\langle \sum_{i=1}^{\ell} u_i(x) K_{x_i} - K_x, f \right\rangle_K. \tag{41}$$

By the Schwartz inequality in $\mathscr{H}_K$,

$$\left|I_{\mathbf{x}}(f)(x) - f(x)\right| \le \left\| \sum_{i=1}^{\ell} u_i(x) K_{x_i} - K_x \right\|_K \|f\|_K. \tag{42}$$

As $\langle K_s, K_t\rangle_K = K(s,t)$, we have

$$\begin{aligned}\left\|\sum_{i=1}^{\ell} u_i(x) K_{x_i} - K_x\right\|_K^2 &= K(x,x) - 2\sum_{i=1}^{\ell} u_i(x) K(x,x_i) \\ &\quad + \sum_{i,j=1}^{\ell} u_i(x) K(x_i,x_j) u_j(x).\end{aligned} \tag{43}$$

However, the quadratic function

$$\begin{aligned}Q\left((w_i)_{i=1}^{\ell}\right) &:= K(x,x) - 2\sum_{i=1}^{\ell} w_i K(x,x_i) \\ &\quad + \sum_{i,j=1}^{\ell} w_i K(x_i,x_j) w_j\end{aligned} \tag{44}$$

over $\mathbb{R}^{\ell}$ takes its minimum value at $(u_i(x))_{i=1}^{\ell}$. Therefore,

$$\left\|\sum_{i=1}^{\ell} u_i(x) K_{x_i} - K_x\right\|_K \le \varepsilon_K(\mathbf{x}). \tag{45}$$

It follows that

$$\left|I_{\mathbf{x}}(f)(x) - f(x)\right| \le \|f\|_K \varepsilon_K(\mathbf{x}). \tag{46}$$

This proves (38).

As $I_{\mathbf{x}}(f) \in \mathscr{H}_K$ and $I_{\mathbf{x}}(f)(x_i) = f(x_i)$ for $i = 1,\dots,\ell$, we know that

$$I_{\mathbf{x}}(f)(x_i) - f(x_i) = \langle K_{x_i}, I_{\mathbf{x}}(f) - f\rangle_K = 0, \quad i = 1,\dots,\ell. \tag{47}$$

This means that $I_{\mathbf{x}}(f) - f$ is orthogonal to $\operatorname{span}\{K_{x_i}\}_{i=1}^{\ell}$. Hence, $I_{\mathbf{x}}(f)$ is the orthogonal projection of f onto $\operatorname{span}\{K_{x_i}\}_{i=1}^{\ell}$. Thus, $\|I_{\mathbf{x}}(f)\|_K \le \|f\|_K$. □

The regularity of the kernel in connection with Theorem 12 yields the rate of convergence of the interpolation scheme. As an example, from the estimate for $\varepsilon_K(\mathbf{x})$ given in [16, Proposition 2], we have the following.

Corollary 13. *Let $X = [0,1]$, $s \le N \in \mathbb{N}$, and $K(x,y)$ be a C^s Mercer kernel such that $A_{\mathbf{x}}$ is nonsingular for $\mathbf{x} = \{j/N\}_{j=0}^{N-1}$. Then, for $f \in \mathscr{H}_K$, there holds*

$$\begin{aligned}&\left\|I_{\mathbf{x}}(f) - f\right\|_{C(X)} \\ &\quad \le \|f\|_K \left\{\frac{(4s)^s(1+s2^s)}{(s-1)!}\left\|\frac{\partial^s}{\partial y^s}K\right\|_{\infty}\right\} N^{-s}.\end{aligned} \tag{48}$$

For convolution type kernels, the power function can be estimated in terms of the Fourier transform of the kernel function. This is of particular interest when the kernel function is analytic. Let us provide the details.

Assume that k is a symmetric function in $L^2(\mathbb{R}^n)$ and $\widehat{k}(\xi) > 0$ almost everywhere on $\mathbb{R}^n$. Consider the Mercer kernel

$$K(x,y) = k(x-y), \quad x,y \in [0,1]^n. \tag{49}$$

For $N \in \mathbb{N}$, we define the following function to measure the regularity:

$$\begin{aligned}\lambda_k(N) &:= n\left(1 + \frac{1}{2^N}\right)^{n-1} \\ &\quad \times \max_{1\le j\le n}\left\{(2\pi)^{-n}\int_{\xi\in[-N/2,N/2]^n} \widehat{k}(\xi)\left(\frac{|\xi_j|}{N}\right)^N d\xi\right\} \\ &\quad + \left(1 + \left(N2^N\right)^n\right)^2 (2\pi)^{-n}\int_{\xi\notin[-N/2,N/2]^n} \widehat{k}(\xi)\, d\xi.\end{aligned} \tag{50}$$

Remark 14. This function involves two parts. The first part is $\xi \in [-N/2,N/2]^n$, where $(|\xi_j|/N)^N \le 2^{-N}$; hence, it decays exponentially fast as N becomes large. The second part is $\xi \notin [-N/2,N/2]^n$, where ξ is large. Then, the decay of $\widehat{k}$ (which is equivalent to the regularity of k) yields the fast decay of the second part.

The power function $\varepsilon_K(\mathbf{x})$ can be bounded by $\lambda_k(N)$ on the regular points:

$$\mathbf{x} := \left(\frac{\alpha}{N}\right)_{\alpha\in\{0,1,\dots,N-1\}^n}. \tag{51}$$

Proposition 15. *For the convolution type kernel* (49) *and* $\mathbf{x}$ *given by* (51)*, one has*

$$\varepsilon_K(\mathbf{x}) \le \lambda_k(N). \tag{52}$$

In particular, if

$$\widehat{k}(\xi) \le C_0 e^{-\lambda|\xi|}, \quad \forall \xi \in \mathbb{R}^n \tag{53}$$

for some constants $C_0 > 0$ and $\lambda > 4 + 2n\ln 4$, then there holds

$$\varepsilon_K(\mathbf{x}) \le \lambda_k(N) \le 4C_0\left(\max\left\{\frac{1}{e\lambda}, \frac{4^n}{e^{\lambda/2}}\right\}\right)^{N/2}. \tag{54}$$

Proof. Choose $\{w_\alpha := w_{\alpha,\mathbf{x}}(x)\}_{\alpha\in\mathbf{x}}$ as the Lagrange interpolation polynomials on $\mathbf{x}$. It is a vector in $\mathbb{R}^{N^n}$ for each $x \in X$. Then, $\varepsilon_K(\mathbf{x}) \le \sup_{x\in X} Q_N(x)$, where

$$\begin{aligned}Q_N(x) &:= k(0) - 2\sum_{\alpha\in\mathbf{x}} w_{\alpha,\mathbf{x}}(x) k(x-\alpha) \\ &\quad + \sum_{\alpha,\beta\in\mathbf{x}} w_{\alpha,\mathbf{x}}(x)\ k(\alpha-\beta) w_{\beta,\mathbf{x}}(x).\end{aligned} \tag{55}$$

In the proof of Theorem 2 in [16], we showed that $Q_N(x) \le \lambda_k(N)$ for each $x \in [0,1]^n$. Therefore, $\varepsilon_K(\mathbf{x}) \le \lambda_k(N)$.

The estimate for $\lambda_k(N)$ in the second part was verified in the proof of Theorem 3 in [16]. □

For the Gaussian kernels

$$K(x,y)=\exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\},\quad x,y\in[0,1]^n, \tag{56}$$

it was proved in [16, Example 4] that, for $N\geq 80n\log 2/\sigma^2$, there holds

$$\varepsilon_K(\mathbf{x})\leq\lambda_k(N)\leq 2\sqrt{e}\left(\frac{1}{16n}\right)^{N/2}+\frac{4}{\sigma\sqrt{\pi}}2^{-nN}. \tag{57}$$

4. Approximation Error in Learning Theory

Now, we can estimate the approximation error in learning theory by means of the interpolation scheme (34).

Consider the convolution type kernel (49) on $X=[0,1]^n$. As in [6], we denote

$$\Lambda_k(r):=\left\{\inf_{\xi\in[-r\pi,r\pi]^n}\widehat{k}(\xi)\right\}^{-1/2},\quad r>0. \tag{58}$$

The approximation error (6) can be realized as follows.

Theorem 16. *Let $k\in L^2(\mathbb{R}^n)$ be a symmetric function with $\widehat{k}(\xi)>0$, and let the kernel on $X=[0,1]^n$ be $K(x,y)=k(x-y)$. For $f\in L^2(\mathbb{R}^n)$ and $M\leq N\in\mathbb{N}$, we set $f_M\in L^2(\mathbb{R}^n)$ by*

$$\widehat{f}_M(\xi)=\begin{cases}\widehat{f}(\xi), & \text{if } \xi\in[-M\pi,M\pi]^n,\\ 0, & \text{otherwise.}\end{cases} \tag{59}$$

Then, with $\mathbf{x}=\{0,1/N,\ldots,(N-1)/N\}^n$, one has

(i) $\|I_{\mathbf{x}}(f_M)\|_K\leq\|f\|_{L^2}\Lambda_k(N)$;

(ii) $\|f_M-I_{\mathbf{x}}(f_M)\|_{C(X)}\leq\|f\|_{L^2}\Lambda_k(M)\varepsilon_K(\mathbf{x})\leq\|f\|_{L^2}\Lambda_k(M)\lambda_k(N)$;

(iii) $\|f-f_M\|^2_{L^2(X)}\leq(2\pi)^{-n}\int_{\xi\notin[-M\pi,M\pi]^n}|\widehat{f}(\xi)|^2d\xi\to 0$ $(as\ M\to\infty)$.

Proof. (i) For $i,j\in X_N:=\{0,1,\ldots,N-1\}^n$ and $x_i=i/N$, expression (33) gives

$$\begin{aligned}\left\langle u_i,u_j\right\rangle_K&=\sum_{s,t\in X_N}\left(A_{\mathbf{x}}^{-1}\right)_{is}\left(A_{\mathbf{x}}^{-1}\right)_{jt}\left\langle K_{x_s},K_{x_t}\right\rangle_K\\&=\sum_{s,t\in X_N}\left(A_{\mathbf{x}}^{-1}\right)_{is}\left(A_{\mathbf{x}}^{-1}\right)_{jt}(A_{\mathbf{x}})_{ts}=\left(A_{\mathbf{x}}^{-1}\right)_{ij}.\end{aligned} \tag{60}$$

Then for $g\in C(X)$ we have

$$\begin{aligned}\left\|I_{\mathbf{x}}(g)\right\|_K^2&=\left\|\sum_{i\in X_N}g(x_i)\,u_i(x)\right\|_K^2\\&=\sum_{i,j\in X_N}g(x_i)\,g(x_j)\left\langle u_i,u_j\right\rangle_K\\&=\left(g|_{\mathbf{x}}\right)^TA_{\mathbf{x}}^{-1}g|_{\mathbf{x}},\end{aligned} \tag{61}$$

where $g|_{\mathbf{x}}$ is the vector $(g(x_i))_{i\in X_N}\in\mathbb{R}^{N^n}$. It follows that

$$\begin{aligned}\left\|I_{\mathbf{x}}(g)\right\|_K^2&=\left\langle g|_{\mathbf{x}},A_{\mathbf{x}}^{-1}g|_{\mathbf{x}}\right\rangle_{\ell^2}\\&\leq\left\|A_{\mathbf{x}}^{-1}\right\|_2\|g|_{\mathbf{x}}\|_{\ell^2}^2\\&=\left\|A_{\mathbf{x}}^{-1}\right\|_2\sum_{i\in X_N}|g(x_i)|^2,\end{aligned} \tag{62}$$

where $\|A_{\mathbf{x}}^{-1}\|_2$ denotes the (operator) norm of the matrix $A_{\mathbf{x}}^{-1}$ in $(\mathbb{R}^{N^n},\ell^2)$.

We apply the previous analysis to the function f_M satisfying

$$\begin{aligned}\sum_{j\in X_N}\left|f_M(x_j)\right|^2&=\sum_{j\in X_N}\left|(2\pi)^{-n}\int_{\xi\in[-N\pi,N\pi]^n}\widehat{f}_M(\xi)\,e^{i\cdot(j/N)\cdot\xi}d\xi\right|^2\\&=\sum_{j\in X_N}\left|(2\pi)^{-n}\int_{\xi\in[-\pi,\pi]^n}\widehat{f}_M(N\xi)\,e^{ij\cdot\xi}N^nd\xi\right|^2\\&\leq(2\pi)^{-n}\int_{\xi\in[-\pi,\pi]^n}\left|\widehat{f}_M(N\xi)\,N^n\right|^2d\xi\\&\leq N^n\|f\|_{L^2}^2.\end{aligned} \tag{63}$$

Then,

$$\left\|I_{\mathbf{x}}(f_M)\right\|_K^2\leq\left\|A_{\mathbf{x}}^{-1}\right\|_2N^n\|f\|_{L^2}^2. \tag{64}$$

Now, we need to estimate the norm $\|A_{\mathbf{x}}^{-1}\|_2$. For convolution type kernels, such an estimate was given in [15, Theorem 2] by means of methods from the radial basis function literature, for example, [17, 21–24]. We have

$$\left\|A_{\mathbf{x}}^{-1}\right\|_2\leq N^{-n}\left(\Lambda_k(N)\right)^2. \tag{65}$$

Therefore,

$$\left\|I_{\mathbf{x}}(f_M)\right\|_K\leq\|f\|_{L^2}\Lambda_k(N). \tag{66}$$

This proves the statement in (i).

(ii) Let $x\in X$. Then

$$\begin{aligned}&f_M(x)-I_{\mathbf{x}}(f_M)(x)\\&\quad=(2\pi)^{-n}\int_{\xi\in[-M\pi,M\pi]^n}\widehat{f}(\xi)\\&\qquad\times\left\{e^{ix\cdot\xi}-\sum_{j\in X_N}u_{j,\mathbf{x}}(x)\,e^{ix_j\cdot\xi}\right\}d\xi.\end{aligned} \tag{67}$$

By the Schwartz inequality,

$$\begin{aligned}&\left|f_M(x)-I_{\mathbf{x}}(f_M)(x)\right|\\&\le\left\{(2\pi)^{-n}\int_{\xi\in[-M\pi,M\pi]^n}\frac{\left|\widehat{f}(\xi)\right|^2}{\widehat{k}(\xi)}d\xi\right\}^{1/2}\\&\times\left\{(2\pi)^{-n}\int_{\mathbb{R}^n}\widehat{k}(\xi)\left|e^{ix\cdot\xi}-\sum_{j\in X_N}u_{j,\mathbf{x}}(x)\,e^{ix_j\cdot\xi}\right|^2d\xi\right\}^{1/2}.\end{aligned}\tag{68}$$

The first term is bounded by $\|f\|_{L^2}\Lambda_k(M)$. The second term is

$$\begin{aligned}&\left\{k(0)-2\sum_{j\in X_N}u_{j,\mathbf{x}}(x)\,k\left(x-x_j\right)\right.\\&\left.+\sum_{i,j\in X_N}u_{i,\mathbf{x}}(x)\,k\left(x_i-x_j\right)u_{j,\mathbf{x}}(x)\right\}^{1/2}\end{aligned}\tag{69}$$

which can be bounded by $\varepsilon_K(\mathbf{x})$, as shown in the proof of Theorem 12. Therefore, by (52),

$$\begin{aligned}\left\|f_M-I_{\mathbf{x}}(f_M)\right\|_{C(X)}&\le\|f\|_{L^2}\Lambda_k(M)\,\varepsilon_K(\mathbf{x})\\&\le\|f\|_{L^2}\Lambda_k(M)\,\lambda_k(N).\end{aligned}\tag{70}$$

(iii) By the Plancherel formula,

$$\left\|f-f_M\right\|^2_{L^2(\mathbb{R}^n)}=(2\pi)^{-n}\int_{\xi\notin[-M\pi,M\pi]^n}\left|\widehat{f}(\xi)\right|^2d\xi.\tag{71}$$

This proves all the statements in Theorem 16. □

Theorem 16 provides quantitative estimates for the approximation error:

$$\begin{aligned}&\left\|f-I_{\mathbf{x}}(f_M)\right\|_{L^2(X)}\\&\le\left\{(2\pi)^{-n}\int_{\xi\notin[-M\pi,M\pi]^n}\left|\widehat{f}(\xi)\right|^2d\xi\right\}^{1/2}\\&\quad+\|f\|_{L^2}\Lambda_k(M)\,\lambda_k(N)\end{aligned}\tag{72}$$

with

$$\left\|I_{\mathbf{x}}(f_M)\right\|_K\le\|f\|_{L^2}\Lambda_k(N).\tag{73}$$

Choose $N=N(M)\ge M$ such that $\Lambda_k(M)\lambda_k(N)\to 0$ as $M\to+\infty$; we have $\|f-I_{\mathbf{x}}(f_M)\|_{L^2(X)}\to 0$ and the RKHS norm of $I_{\mathbf{x}}(f_M)$ is controlled by the asymptotic behavior of $\Lambda_k(N)$.

Denote by Λ_k^{-1} the inverse function of Λ_k:

$$\begin{aligned}\Lambda_k^{-1}(R)&:=\max\left\{r>0:\Lambda_k(r)\le R\right\}\\&=\max\left\{r>0:\widehat{k}(\xi)\ge R^{-2}\ \forall\xi\in[-r\pi,r\pi]^n\right\}.\end{aligned}\tag{74}$$

Then, our estimate for the approximation error can be given as follows.

Corollary 17. *Let $X=[0,1]^n$ and $f\in H^s(\mathbb{R}^n)$. Then, for $R>\|f\|_{L^2}$,*

$$\begin{aligned}&\inf_{\|g\|_K\le R}\|f-g\|_{L^2(X)}\\&\quad\le\inf_{0<M\le N_R}\left\{\|f\|_{H^s}(M\pi)^{-s}+\|f\|_{L^2}\Lambda_k(M)\,\lambda_k(N_R)\right\},\end{aligned}\tag{75}$$

where $N_R:=[\Lambda_k^{-1}(R/\|f\|_{L^2})]$. If $s>n/2$, then

$$\begin{aligned}&\inf_{\|g\|_K\le R}\|f-g\|_{C(X)}\\&\quad\le\inf_{0<M\le N_R}\left\{\frac{\|f\|_{H^s}}{\sqrt{s-n/2}}M^{n/2-s}\right.\\&\qquad\left.+\|f\|_{L^2}\Lambda_k(M)\,\lambda_k(N_R)\right\}.\end{aligned}\tag{76}$$

In particular, when

$$C_1\exp\left\{-\lambda_1|\xi|^d\right\}\le\widehat{k}(\xi)\le C_0\exp\left\{-\lambda|\xi|\right\},\quad\forall\xi\in\mathbb{R}^n\tag{77}$$

for some $C_0,C_1,d,\lambda_1>0$ and $\lambda>4+2n\log 4$, one has

$$\begin{aligned}I(f,R)&=\inf_{\|g\|_K\le R}\|f-g\|_{L^2(X)}\\&\le\left(2^{d(d+1)}n^{d/2}\pi^d\lambda_1\right)^{s/d(d+1)}\\&\quad\times\left(\pi^{-s}\|f\|_{H^s}+\frac{2^{s+2}C_0}{\sqrt{C_1}}\|f\|_{L^2}\right)\\&\quad\times\left\{\log R+\frac{1}{2}\log C_1-\log\|f\|_{L^2}\right\}^{-s/d(d+1)}\end{aligned}\tag{78}$$

provided that with the function $G(r):=(1/\sqrt{n}\pi)((2/\lambda_1)\log(r\sqrt{C_1}/\|f\|_{L^2}))^{1/d}$, R satisfies

$$\begin{aligned}G(R)&\ge\left(\frac{16\lambda_1n^{d/2}\pi^d}{-\log\max\left\{1/e\lambda,4^n/e^{\lambda/2}\right\}}\right)^{d+1},\\\frac{\log G(R)}{G(R)}&\le\frac{\left(-\log\max\left\{1/e\lambda,4^n/e^{\lambda/2}\right\}\right)(d+1)}{2^{d+4}s}.\end{aligned}\tag{79}$$

Proof. The first part is a direct consequence of Theorem 16 when we choose N to be N_R, the integer part of $\Lambda_k^{-1}(R/\|f\|_{L^2})$.

To see the second part, we note that (77) in connection with Proposition 15 implies with $\Lambda:=\max\{1/e\lambda,4^n/e^{\lambda/2}\}$,

$$\begin{aligned}\lambda_k(N)&\le 4C_0\exp\left\{N\frac{\log\Lambda}{2}\right\},\\\Lambda_k(r)&\le\left\{C_1\exp\left\{-\lambda_1\left(\sqrt{n}r\pi\right)^d\right\}\right\}^{-1/2}\\&=\frac{1}{\sqrt{C_1}}\exp\left\{\frac{\lambda_1}{2}\left(\sqrt{n}\pi\right)^dr^d\right\}.\end{aligned}\tag{80}$$

Then, $\Lambda_k^{-1}(R/\|f\|_{L^2})\ge G(R)$.

For $R \geq (\|f\|_{L^2}/\sqrt{C_1})\ \exp\{(\lambda_1/2)(\sqrt{n}\pi)^d\}$, we can choose $M \in \mathbb{N}$ such that

$$\frac{1}{2}\{G(R)\}^{1/(d+1)} \leq M \leq \{G(R)\}^{1/(d+1)}. \tag{81}$$

Choose $N \in \mathbb{N}$ such that

$$\frac{M^{d+1}}{2} \leq N \leq M^{d+1}. \tag{82}$$

Then, $M \leq N$, and by Theorem 16,

$$\begin{aligned}
&\|I_{\mathbf{x}}(f_M)\|_K \leq \frac{\|f\|_{L^2}}{\sqrt{C_1}} \exp\left\{\frac{\lambda_1}{2}(\sqrt{n}\pi)^d M^{d(d+1)}\right\} \leq R,\\
&\Lambda_k(M)\lambda_k(N)\\
&\leq \frac{4C_0}{\sqrt{C_1}} \exp\left\{N\frac{\log\Lambda}{4} + N\right.\\
&\qquad\left.\times\left(\frac{\log\Lambda}{4} + \lambda_1(\sqrt{n}\pi)^d N^{-1/(d+1)}\right)\right\}.
\end{aligned} \tag{83}$$

When

$$\begin{aligned}
N^{1/(d+1)} &\geq \frac{4\lambda_1(\sqrt{n}\pi)^d}{-\log\Lambda},\\
\frac{\log N}{N} &\leq \frac{(-\log\Lambda)(d+1)}{4s},
\end{aligned} \tag{84}$$

there holds

$$\begin{aligned}
\Lambda_k(M)\lambda_k(N) &\leq \frac{4C_0}{\sqrt{C_1}} \exp\left\{N\frac{\log\Lambda}{4}\right\}\\
&\leq \frac{4C_0}{\sqrt{C_1}} N^{-s/(d+1)}.
\end{aligned} \tag{85}$$

Hence,

$$\begin{aligned}
&\|f - I_{\mathbf{x}}(f_M)\|_{L^2(X)}\\
&\leq \|f\|_{H^s}(M\pi)^{-s} + \|f\|_{L^2}\frac{4C_0}{\sqrt{C_1}}\left(\frac{2}{M}\right)^s\\
&\leq \left(\|f\|_{H^s}\pi^{-s} + 2^s\|f\|_{L^2}\frac{4C_0}{\sqrt{C_1}}\right)\\
&\quad\times 2^s(G(R))^{-s/(d+1)}.
\end{aligned} \tag{86}$$

When R satisfies (79), we know that

$$\begin{aligned}
&N^{1/(d+1)} \geq \frac{M}{2} \geq \frac{1}{4}\left\{G(R)^{1/d}\right\}^{1/(d+1)} \geq \frac{4\lambda_1(\sqrt{n}\pi)^d}{-\log\Lambda},\\
&\frac{\log N}{N} \leq \frac{2\log M^{d+1}}{M^{d+1}} \leq \frac{2^{d+2}\log G(R)}{G(R)} \leq \frac{-\log\Lambda\,(d+1)}{4s}.
\end{aligned} \tag{87}$$

Hence, (84) holds true. This proves our statements. □

For the Gaussian kernels, we have the following.

Proposition 18. *Let*

$$K(x,y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}, \quad x, y \in X = [0,1]^n. \tag{88}$$

Denote $C_{\sigma,n} := \sigma^2 n\pi^2/4\min\{\log(4\sqrt{n}), n\log 2\}$ *and* $C_{\sigma,n,s} := (\sigma\sqrt{n}\pi)^{s/2}(C_{\sigma,n}^{s/2} + (\sigma\sqrt{\pi})^{-n/2}(2\sqrt{e} + 4/\sigma\sqrt{\pi}))$. *If* $f \in H^s(\mathbb{R}^n)$, *then one has*

$$\begin{aligned}
I(f,R) &\leq C_{\sigma,n,s}\left(\|f\|_{H^s} + \|f\|_{L^2}\right)\\
&\quad\times\left\{\log R + \frac{n}{2}\log(\sigma\sqrt{\pi}) - \log\|f\|_{L^2}\right\}^{-s/4}
\end{aligned} \tag{89}$$

and when $s > n/2$,

$$\begin{aligned}
&\inf_{\|g\|_K \leq R}\|f - g\|_{C(X)}\\
&\quad\leq C_{\sigma,n,s-n/2}\left(\|f\|_{H^s} + \|f\|_{L^2}\right)\\
&\quad\times\left\{\log R + \frac{n}{2}\log(\sigma\sqrt{\pi}) - \log\|f\|_{L^2}\right\}^{n/8-s/4}
\end{aligned} \tag{90}$$

for R satisfying

$$\begin{aligned}
R &> \|f\|_{L^2}(\sigma\sqrt{\pi})^{-n/2}\\
&\quad\times\exp\left\{\frac{\sigma^2 n\pi^2\left(\max\left\{C_{\sigma,n}, 80n\log 2/\sigma^2\right\} + 1\right)^2}{8}\right\},
\end{aligned} \tag{91}$$

$$\begin{aligned}
&\left(\frac{1}{32C_{\sigma,n}}\log\frac{(\sigma\sqrt{\pi})^{n/2}R}{\|f\|_{L^2}}\right)^{1/2}\\
&\quad\geq \frac{s}{2}\left(\log\frac{2\sqrt{2}}{\sigma\sqrt{n}\pi} + \frac{1}{2}\log\frac{(\sigma\sqrt{\pi})^{n/2}R}{\|f\|_{L^2}}\right).
\end{aligned} \tag{92}$$

Proof. The Fourier transform of $k(x) = \exp\{-|x|^2/\sigma^2\}$ is

$$\widehat{k}(\xi) = (\sigma\sqrt{\pi})^n \exp\left\{-\frac{\sigma^2|\xi|^2}{4}\right\}. \tag{93}$$

Then,

$$\Lambda_k(r) = (\sigma\sqrt{\pi})^{-n/2}\exp\left\{\frac{\sigma^2 n r^2\pi^2}{8}\right\}. \tag{94}$$

For

$$R \geq \|f\|_{L^2}\Lambda_k\left(\max\left\{C_{\sigma,n}, \frac{80n\log 2}{\sigma^2}\right\} + 1\right), \tag{95}$$

we can take $N \in \mathbb{N}$ with $N \geq \max\{C_{\sigma,n}, 80n\log 2/\sigma^2\}$ such that

$$\frac{1}{2}\Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right) \leq N \leq \Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right). \tag{96}$$

Here, Λ_k^{-1} is the inverse function of Λ_k:

$$\Lambda_k^{-1}(r) = \frac{2\sqrt{2}}{\sigma\sqrt{n\pi}}\left(\log\left\{(\sigma\sqrt{\pi})^{n/2} r\right\}\right)^{1/2}. \tag{97}$$

Then, $\|f\|_{L^2}\Lambda_k(N) \le R$. Let $M \le N$. By Theorem 16, $\|I_{\mathbf{x}}(f_M)\|_K \le R$.

By Corollary 17 and (57),

$$\begin{aligned}\left\|f - I_{\mathbf{x}}(f_M)\right\|_{L^2(X)} &\le \|f\|_{H^s}(M\pi)^{-s} \\ &\quad + \|f\|_{L^2}\Lambda_k(M)\left(2\sqrt{e} + \frac{4}{\sigma\sqrt{\pi}}\right)\exp\left\{-NC_0\right\},\end{aligned} \tag{98}$$

where $C_0 := \min\{\log(4\sqrt{n}), n\log 2\}$. Choose $M \in \mathbb{N}$ such that

$$\frac{1}{2}\sqrt{\frac{N}{C_{\sigma,n}}} \le M \le \sqrt{\frac{N}{C_{\sigma,n}}}. \tag{99}$$

With this choice, $\sigma^2 nM^2\pi^2/8 \le C_0 N/2$. Therefore,

$$\begin{aligned}\left\|f - I_{\mathbf{x}}(f_M)\right\|_{L^2(X)} &\le \|f\|_{H^s}\left(\frac{4C_{\sigma,n}}{\pi^2 N}\right)^{s/2} + \|f\|_{L^2}(\sigma\sqrt{\pi})^{-n/2} \\ &\quad \times \left(2\sqrt{e} + \frac{4}{\sigma\sqrt{\pi}}\right)\exp\left\{-\frac{C_0}{2}N\right\} \\ &\le C'_{\sigma,n,s}\left(\|f\|_{H^s} + \|f\|_{L^2}\right) \\ &\quad \times \max\left\{\left(\frac{R}{\|f\|_{L^2}}\right)^{-s/2}, \exp\left\{-\frac{C_0}{4}\Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right)\right\}\right\},\end{aligned} \tag{100}$$

where

$$C'_{\sigma,n,s} = C_{\sigma,n}^{s/2} + (\sigma\sqrt{\pi})^{-n/2}\left(2\sqrt{e} + \frac{4}{\sigma\sqrt{\pi}}\right). \tag{101}$$

When

$$\frac{C_0}{4}\Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right) \ge \frac{s}{2}\log\Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right), \tag{102}$$

there holds

$$\begin{aligned}\left\|f - I_{\mathbf{x}}(f_M)\right\|_{L^2(X)} &\le C'_{\sigma,n,s}\left(\|f\|_{H^s} + \|f\|_{L^2}\right) \\ &\quad \times \left\{\Lambda_k^{-1}\left(\frac{R}{\|f\|_{L^2}}\right)\right\}^{-s/2}.\end{aligned} \tag{103}$$

This yields the first estimate.

When $s > n/2$, the same method gives the error with the uniform norm. □

5. Learning with Varying Kernels

Proposition 18 in the last section shows that, for a fixed Gaussian kernel, the approximation error $I(f, R)$ behaves as

$$I(f, R) \le C(\log R)^{-s/4} \tag{104}$$

for functions f in H^s.

In this section, we consider the learning with varying kernels. Such a method is used in many applications where we have to choose suitable parameters for the reproducing kernel. For example, in [7] Gaussian kernels with different parameters in different directions are considered. Here, we study the case when the variance parameter keeps the same in all directions. Our analysis shows that the approximation error may be improved when the kernel changes with the RKHS norm R of the empirical target function.

Proposition 19. *Let*

$$K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}, \quad x, y \in X = [0,1]^n. \tag{105}$$

There exist positive constants $A_{n,s}$ and $B_{n,s}$, depending only on n and s, such that for each $f \in H^s(\mathbb{R}^n)$ and $R \ge A_{n,s}\|f\|_{L^2}$, one can find some $\sigma = \sigma_R$ satisfying

$$\inf_{\|g\|_{K_{\sigma_R}} \le R}\|f - g\|_{L^2(X)} \le B_{n,s}(\log R)^{-s}. \tag{106}$$

Proof. Take

$$\begin{aligned}\sigma &= \left(\frac{80n\log 2}{N}\right)^{1/2}, \\ \frac{N}{4n\pi}\left(\frac{1}{5}\min\left\{2 + \frac{\log n}{2\log 2}, n\right\}\right)^{1/2} &\le M \le \frac{N}{2n\pi}\left(\frac{1}{5}\min\left\{2 + \frac{\log n}{2\log 2}, n\right\}\right)^{1/2},\end{aligned} \tag{107}$$

where N depends on R. Denote $C_n := n\pi^2/4\min\{\log(4\sqrt{n}), n\log 2\}$. As in the proof of Proposition 18, we have

$$\begin{aligned}\left\|f - I_{\mathbf{x}}(f_M)\right\|_{L^2(X)} &\le (N\pi)^{-s}\|f\|_{H^s}(320nC_n\log 2)^{s/2} \\ &\quad + \frac{N^{n/2}\|f\|_{L^2}}{(80n\log 2)^{n/4}} \\ &\quad \times \left(2\sqrt{e} + \frac{4\sqrt{N}}{\sqrt{80n\pi\log 2}}\right)\exp\left\{-\frac{C_0}{2}N\right\}.\end{aligned} \tag{108}$$

When N is large enough, with a constant $C'_{n,s}$ depending on n and s, this yields

$$\left\|f - I_{\mathbf{x}}(f_M)\right\|_{L^2(X)} \le C'_{n,s}\left(\|f\|_{H^s} + \|f\|_{L^2}\right)N^{-s}. \tag{109}$$

Finally, we determine N by requiring

$$\begin{aligned}&\|f\|_{L^2}\Lambda_k(N)\\&\quad=\|f\|_{L^2}N^{n/4}(80n\pi\log 2)^{-n/4}\exp\left\{10n^2\pi^2N\log 2\right\}\\&\quad\le R\le\|f\|_{L^2}\Lambda_k(N+1).\end{aligned}\tag{110}$$

There is a constant $A_{n,s}>0$ depending only on n and s such that, for $R\ge A_{n,s}\|f\|_{L^2}$, an integer N satisfying all the previous requirements and

$$(N+1)^{n/4}\le\exp\left\{10n^2\pi^2(N+1)\log 2\right\}\tag{111}$$

exists. This makes all the estimates valid. It follows that

$$\begin{aligned}&R\le\|f\|_{L^2}\Lambda_k(N+1)\\&\quad\le\|f\|_{L^2}(80n\pi\log 2)^{-n/4}\exp\left\{20n^2\pi^2(N+1)\log 2\right\}.\end{aligned}\tag{112}$$

Hence,

$$N+1\ge\frac{1}{20n^2\pi^2\log 2}\log\left\{\frac{R}{\|f\|_{L^2}}(80n\pi\log 2)^{n/4}\right\}.\tag{113}$$

Therefore, there holds $\|I_{\mathbf{x}}(f_M)\|_{K_\sigma}\le R$ and

$$\begin{aligned}&\|f-I_{\mathbf{x}}(f_M)\|_{L^2(X)}\\&\quad\le 2^sC'_{n,s}\left(\|f\|_{H^s}+\|f\|_{L^2}\right)\left(20n^2\pi^2\log 2\right)^s\\&\qquad\times\left\{\log R+\frac{n}{4}\log(80n\pi\log 2)-\log\|f\|_{L^2}\right\}^{-s}.\end{aligned}\tag{114}$$

This verifies our claim for the approximation error in $L^2(X)$. □

Let us mention the following problem concerning learning with Gaussian kernels with changing variances.

Problem 20. What is the optimal rate of convergence of

$$\sup_{\|f\|_{H^s}=1}\inf\left\{\|f-g\|_{L^2(X)}:\ \|g\|_{K_\sigma}\le R\text{ for some }\sigma>0\right\}\tag{115}$$

as R tends to infinity?

6. Dot Product Kernels

In this section, we illustrate our results by the family of dot product type kernels. These kernels take the form

$$K(x,y)=\sum_{j=0}^{+\infty}a_j(x\cdot y)^j,\quad x,y\in\mathbb{R}^n.\tag{116}$$

When $\sum_{j=0}^{+\infty}|a_j|R^{2j}<\infty$ for some $R>0$, the kernel K is a Mercer kernel on $X:=\{x\in\mathbb{R}^n:|x|\le R\}$ if and only if $a_j\ge 0$ for each $j\ge 0$. See [25–28]. Here, we will characterize the density for this family as [29]. Denote $x^\alpha:=\Pi_{i=1}^n x_i^{\alpha_i}$ and $\binom{|\alpha|}{\alpha}:=(\alpha_1+\cdots\alpha_n)!/(\Pi_{i=1}^n\alpha_j!)$ for $x=(x_1,\ldots,x_n)\in\mathbb{R}^n$ and $\alpha=(\alpha_1,\ldots,\alpha_n)\in\mathbb{Z}_+^n$.

Corollary 21. *Let $R>0$, $X:=[0,R]^n$, and the kernel K be given by* (116)*, where $a_j\ge 0$ for each $j\in\mathbb{Z}_+$ and $\sum_{j=0}^{+\infty}a_jR^{2j}<\infty$. Set $J:=\{\alpha\in\mathbb{Z}_+^n:a_{|\alpha|}>0\}$. Then, $\mathscr{H}_K$ is dense in $C(X)$ if and only if span$\{x^\alpha:\alpha\in J\}$ is dense in $C(X)$. Thus, the density depends only on the location of positive coefficients in* (116)*. In particular, when $n=1$, $\mathscr{H}_K$ is dense in $C[0,R]$ if and only if*

$$a_0>0,\qquad\sum_{j\in J\setminus\{0\}}\frac{1}{j}=+\infty.\tag{117}$$

Proof. Note that

$$\begin{aligned}K_x(y)&=K(x,y)\\&=a_0+\sum_{j=1}^{+\infty}a_j\sum_{|\alpha|=j}\binom{|\alpha|}{\alpha}x^\alpha y^\alpha\\&=\sum_{\alpha\in J}a_{|\alpha|}\binom{|\alpha|}{\alpha}x^\alpha y^\alpha.\end{aligned}\tag{118}$$

Sufficiency. Suppose that span$\{x^\alpha:\alpha\in J\}$ is dense in $C(X)$, but $\mathscr{H}_K$ is not dense in $C(X)$. Then, by Theorem 4 there exists a nontrivial Borel measure μ on X such that

$$\int_X K(x,y)\,d\mu(y)=0,\quad\forall x\in X.\tag{119}$$

Taking the integral with respect to μ and using (118), we have

$$\begin{aligned}0&=\iint_X K(x,y)\,d\mu(x)\,d\mu(y)\\&=\sum_{\alpha\in J}a_{|ga|}\binom{|\alpha|}{\alpha}\left[\int_X x^\alpha d\mu(x)\right]^2.\end{aligned}\tag{120}$$

Since $a_{|\alpha|}>0$ for each $\alpha\in J$, there holds

$$\int_X x^\alpha d\mu(x)=0,\quad\forall\alpha\in J.\tag{121}$$

That is, μ annihilates each x^α for $\alpha\in J$. But span$\{x^\alpha:\alpha\in J\}$ is dense in $C(X)$; μ also annihilates all functions in $C(X)$, which is a contradiction.

Necessity. If span$\{x^\alpha:\alpha\in J\}$ is not dense in $C(X)$, then there exists a nontrivial Borel measure μ annihilating each x^α; that is, $\int_X x^\alpha d\mu(x)=0$ for each $\alpha\in J$. Then, (118) tells us that, for each $x\in X$,

$$\int_X K(x,y)\,d\mu(y)=\sum_{\alpha\in J}a_{|ga|}\binom{|\alpha|}{\alpha}x^\alpha\int_X y^\alpha d\mu(y)=0.\tag{122}$$

This in connection with Theorem 4 implies that $\mathscr{H}_K$ is not dense in $C(X)$. This proves the first statement of Corollary 21.

The second statement follows from the classical Muntz Theorem in approximation theory (see [30]): for a strictly increasing sequence of nonnegative numbers $\lambda_0<\lambda_1,\ldots$, span$\{x^{\lambda_j}:j\in\mathbb{Z}_+\}$ is dense in $C[0,R]$ if and only if $\lambda_0=0$ and $\sum_{j=1}^{+\infty}1/\lambda_j=+\infty$. □

The conclusion in Example 2 follows directly from Corollary 21.

By Corollary 21, we can provide more examples of dot product positive definite kernels whose corresponding RKHS is not dense. The following is such an example. However, compared with Example 8, it is not constructive, in the sense that no function outside the closure of $\mathscr{H}_K$ is explicitly given.

Example 22. Let $X = [0, 1]$ and define

$$K(x, y) = 1 + \sum_{k=1}^{+\infty} 2^{-k} \sum_{m=0}^{k-1} (x \cdot y)^{2^k+m}. \tag{123}$$

Then, K is a positive definite Mercer kernel on X, but $\mathscr{H}_K$ is not dense in $C(X)$.

Proof. Observe that the assumption in Corollary 21 holds for K, $a_0 = 1 > 0$ and

$$J \setminus \{0\} = \bigcup_{k=1}^{+\infty} \left\{2^k, 2^k + 1, \ldots, 2^k + k - 1\right\}. \tag{124}$$

Since $\sum_{j\in J\setminus\{0\}} (1/j) = \sum_{k=1}^{+\infty} \sum_{i=0}^{k-1} (1/(2^k + i)) \le \sum_{k=1}^{+\infty} (k/2^k) < +\infty$, Corollary 21 tells us that $\mathscr{H}_K$ is not dense in $C(X)$.

What is left is to show that the Mercer kernel K is positive definite. Suppose to the contrary that there exist a finite set of distinct points $I \subset X$ and a nonzero vector $c = (c_s)_{s\in I}$ such that

$$\sum_{s,t\in I} c_s c_t K(s, t) = 0. \tag{125}$$

Denote

$$\widetilde{K}(x, y) = \sum_{k=1}^{+\infty} 2^{-k} \sum_{m=0}^{k-1} (x \cdot y)^{2^k+m}. \tag{126}$$

Then,

$$\sum_{s,t\in I} c_s c_t K(s, t) = \left(\sum_{s\in I} c_s\right)^2 + \sum_{s,t\in I\setminus\{0\}} c_s c_t \widetilde{K}(s, t) = 0. \tag{127}$$

Hence, $\sum_{s\in I} c_s = 0$ which implies that $I \setminus \{0\} \neq \emptyset$ and $(c_s)_{s\in I\setminus\{0\}}$ is a nonzero vector. Also,

$$\begin{aligned} 0 &= \sum_{s,t\in I\setminus\{0\}} c_s c_t \widetilde{K}(s, t) \\ &= \sum_{k=0}^{+\infty} 2^{-k} \sum_{m=0}^{k-1} \left(\sum_{s\in I\setminus\{0\}} s^{2^k+m} c_s \right)^2. \end{aligned} \tag{128}$$

It follows that

$$\sum_{s\in I\setminus\{0\}} s^{2^k+m} c_s = 0, \quad \forall k \in \mathbb{N},\ m = 0, 1, \ldots, k - 1. \tag{129}$$

Choose an integer k which is not less than $\#(I \setminus \{0\})$, the number of elements in the set $I \setminus \{0\}$. Then, we know that the linear system

$$\sum_{s\in I\setminus\{0\}} s^{2^k+m} x_s = 0, \quad m = 0, 1, \ldots, \#(I \setminus \{0\}) \tag{130}$$

has a nonzero solution $(c_s)_{s\in I\setminus\{0\}}$. Hence, the matrix $(s^{2^k+m})_{s\in I\setminus\{0\}, m=0,1,\ldots,\#(I\setminus\{0\})}$ is singular. So, there exists a nonzero vector $(d_m)_{m=0}^{\#(I\setminus\{0\})}$ such that

$$\sum_{m=0}^{\#(I\setminus\{0\})} s^{2^k+m} d_m = 0, \quad \forall s \in I \setminus \{0\}\ . \tag{131}$$

As each element s in $I \setminus \{0\}$ is nonzero, we have

$$\sum_{m=0}^{\#(I\setminus\{0\})} s^m d_m = 0, \quad \forall s \in I \setminus \{0\}. \tag{132}$$

However, the determinant of the matrix $(s^m)_{s\in I\setminus\{0\}, m=0,1,\ldots,\#(I\setminus\{0\})}$ is a Vandermonde determinant and is nonzero. This is a contradiction, as the linear system having this matrix as the coefficient matrix has a nonzero solution. Therefore, the Mercer kernel K is positive definite. □

An alternative simpler proof for the positive definiteness of the kernel in Example 22 can be given by means of the recent results in [25, 26].

After characterizing the density, we can then apply our analysis in Section 3 and provide some estimates for the convergence rate of the approximation error under the assumption that all the coefficients a_j in (116) are strictly positive. We will not provide details here, but only show the application of the interpolation scheme (34) to polynomials.

If $f(x) = \sum_{|\alpha|\le M} c_\alpha \binom{|\alpha|}{\alpha} x^\alpha$, then

$$\begin{aligned} &I_{\mathbf{x}}(f)(x) - f(x) \\ &\quad = \sum_{|\alpha|\le M} c_\alpha \binom{|\alpha|}{\alpha} \left\{ \sum_{j=1}^{\ell} u_j(x)\, x_j^\alpha - x^\alpha \right\}. \end{aligned} \tag{133}$$

It follows from the Schwartz inequality that

$$\begin{aligned} &\left|I_{\mathbf{x}}(f)(x) - f(x)\right|^2 \\ &\quad \le \left\{ \sum_{|\alpha|\le M} \frac{|c_\alpha|^2 \binom{|\alpha|}{\alpha}}{a_{|\alpha|}} \right\} \\ &\quad \times \left\{ \sum_{|\alpha|\le M} a_{|\alpha|} \binom{|\alpha|}{\alpha} \left| \sum_{j=1}^{\ell} u_j(x)\, x_j^\alpha - x^\alpha \right|^2 \right\}. \end{aligned} \tag{134}$$

The first term can be bounded by

$$\left\{ \sum_{|\alpha|\le M} \binom{|\alpha|}{\alpha} |c_\alpha|^2 \right\} \left(\min_{j=0,1,\ldots,M} a_j \right)^{-1} \tag{135}$$

while the second is bounded by

$$\begin{aligned} &\sum_{\alpha\in\mathbb{Z}_+^n} a_{|\alpha|} \binom{|\alpha|}{\alpha} \left\{ \sum_{i,j=1}^{\ell} u_i(x)\, u_j(x)\, x_i^\alpha x_j^\alpha \right. \\ &\qquad \left. -2 \sum_{j=1}^{\ell} u_j(x)\, x_j^\alpha x^\alpha + x^\alpha \cdot x^\alpha \right\} = \varepsilon_K(\mathbf{x}). \end{aligned} \tag{136}$$

Thus, the approximation error can be given in terms of the regularity of the kernel K. The regularity of the approximated function yields the rate of approximation by polynomials $f = f_M$ while the asymptotic behavior of the coefficients a_j in (116) provides the control of the RKHS norm of $I_{\mathbf{x}}(f_M)$.

Acknowledgments

The author would like to thank Charlie Micchelli for proving Corollary 6 in the general form, Allan Pinkus for clarifying Example 2, Tommy Poggio for raising the density problem, Steve Smale for suggestions on positive definiteness and approximation error in learning theory, and Grace Wahba for knowledge on earlier work of approximation by reproducing kernel Hilbert spaces. The work described in this paper is partially supported by a Grant from the Research Grants Council of Hong Kong (Project no. CityU 104710).

References

[1] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[3] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[4] G. Wahba, *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics (SIAM), 1990.

[5] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[6] S. Smale and D.-X. Zhou, "Estimating the approximation error in learning theory," *Analysis and Applications*, vol. 1, no. 1, pp. 17–41, 2003.

[7] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, pp. 131–159, 2002.

[8] T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri, "Uncertainty," in *Goemetric Computations*, J. Winkler and M. Niranjan, Eds., pp. 131–141, Kluwer, 2002.

[9] P. Malliavin, *Integration and Probability*, Springer, 1995.

[10] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *Journal of Machine Learning Research*, vol. 2, no. 1, pp. 67–93, 2002.

[11] C. A. Micchelli, Y. Xu, and P. Ye, "Cucker Smale learning theory in Besov spaces," in *Advances in Learning Theory: Methods, Models and Applications*, J. Suykens, G. Horvath, S. Basu, C. A. Micchelli, and J. Vandewalle, Eds., pp. 47–68, IOS Press, Amsterdam, The Netherlands, 2003.

[12] C. A. Micchelli, "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, no. 1, pp. 11–22, 1986.

[13] F. Girosi and T. Poggio, "Networks and the best approximation property," *Biological Cybernetics*, vol. 63, no. 3, pp. 169–176, 1990.

[14] G. Wahba, "Practical approximate solutions to linear operator equations when the data are noisy," *SIAM Journal on Numerical Analysis*, vol. 14, no. 4, pp. 651–667, 1977.

[15] D.-X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1743–1752, 2003.

[16] D.-X. Zhou, "The covering number in learning theory," *Journal of Complexity*, vol. 18, no. 3, pp. 739–767, 2002.

[17] R. Schaback, "Reconstruction of multivariate functions from scattered data," monograph manusript, 1997.

[18] M. D. Buhmann and M. J. D. Powell, "Radial basis function interpolation on an infinite regular grid," in *Algorithms for Approximation, II*, pp. 146–169, Chapman and Hall, London, UK, 1990.

[19] K. Jetter, J. Stöckler, and J. D. Ward, "Error estimates for scattered data interpolation on spheres," *Mathematics of Computation*, vol. 68, no. 226, pp. 733–747, 1999.

[20] Z. M. Wu and R. Schaback, "Local error estimates for radial basis function interpolation of scattered data," *IMA Journal of Numerical Analysis*, vol. 13, no. 1, pp. 13–27, 1993.

[21] K. Ball, "Eigenvalues of Euclidean distance matrices," *Journal of Approximation Theory*, vol. 68, no. 1, pp. 74–82, 1992.

[22] F. J. Narcowich and J. D. Ward, "Norms of inverses and condition numbers for matrices associated with scattered data," *Journal of Approximation Theory*, vol. 64, no. 1, pp. 69–94, 1991.

[23] F. J. Narcowich and J. D. Ward, "Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices," *Journal of Approximation Theory*, vol. 69, no. 1, pp. 84–109, 1992.

[24] R. Schaback, "Lower bounds for norms of inverses of interpolation matrices for radial basis functions," *Journal of Approximation Theory*, vol. 79, no. 2, pp. 287–306, 1994.

[25] F. Y. Lu and H. W. Sun, "Positive definite dot product kernels in learning theory," *Advances in Computational Mathematics*, vol. 22, no. 2, pp. 181–198, 2005.

[26] A. Pinkus, "Strictly positive definite functions on a real inner product space," *Advances in Computational Mathematics*, vol. 20, no. 4, pp. 263–271, 2004.

[27] A. J. Smola, B. Schölkopf, and K. R. Müller, "The connection between regularization operators and support vector kernels," *Neural Networks*, vol. 11, pp. 637–649, 1998.

[28] D. X. Zhou, "Conditionally reproducing kernel spaces in learning theory," preprint.

[29] W. Dahmen and C. A. Micchelli, "Some remarks on ridge functions," *Approximation: Theory and Applications*, vol. 3, pp. 139–143, 1987.

[30] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, 1966.