Analysis and Applications, Vol. 1, No. 1 (2003) 1–25

World Scientific
www.worldscientific.com

ESTIMATING THE APPROXIMATION ERROR IN LEARNING THEORY

STEVE SMALE* and DING-XUAN ZHOU†

Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
**masmale@math.cityu.edu.hk*
†*mazhou@math.cityu.edu.hk*

Received 19 February 2001
Revised 30 August 2001

Let B be a Banach space and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a dense, imbedded subspace. For $a \in B$, its distance to the ball of $\mathcal{H}$ with radius R (denoted as $I(a, R)$) tends to zero when R tends to infinity. We are interested in the rate of this convergence. This approximation problem arose from the study of learning theory, where B is the L_2 space and $\mathcal{H}$ is a reproducing kernel Hilbert space.

The class of elements having $I(a, R) = O(R^{-r})$ with $r > 0$ is an interpolation space of the couple $(B, \mathcal{H})$. The rate of convergence can often be realized by linear operators. In particular, this is the case when $\mathcal{H}$ is the range of a compact, symmetric, and strictly positive definite linear operator on a separable Hilbert space B. For the kernel approximation studied in Learning Theory, the rate depends on the regularity of the kernel function. This yields error estimates for the approximation by reproducing kernel Hilbert spaces. When the kernel is smooth, the convergence is slow and a logarithmic convergence rate is presented for analytic kernels in this paper. The purpose of our results is to provide some theoretical estimates, including the constants, for the *approximation error* required for the learning theory.

Keywords: Learning theory; approximation error; reproducing kernel Hilbert space; kernel machine learning; interpolation space; logarithmic rate of convergence.

1. Introduction

In the recent study of learning theory [3], the following approximation problem arose:

Let $(B, \|\cdot\|)$ be a Banach space and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a dense subspace with $\|b\| \le \|b\|_{\mathcal{H}}$ for $b \in \mathcal{H}$. Given $a \in B$, what is the **convergence rate** of the function

$$I(a, R) := \inf_{\|b\|_{\mathcal{H}} \le R} \{\|a - b\|\}\,, \qquad R > 0\,, \tag{1.1}$$

as $R \to +\infty$?

A typical setting used in learning theory (see Sec. 2 for more details) is the **kernel representation**: B is $L^2(X)$, the space of square integrable functions over

a bounded domain X, and $\mathcal{H}$ is the range of an integral operator $L_K : L^2(X) \to L^2(X)$ given by

$$L_K f(x) = \int_X K(x,t) f(t) dt\,.$$

The norm for $b \in \mathcal{H}$ is $\|b\|_{\mathcal{H}} = \|L_K^{-1} b\|_{L^2(X)}$, that is, $\|L_K f\|_{\mathcal{H}} = \|f\|_{L^2(X)}$ for $f \in L^2(X)$.

We illustrate the solution to the approximation problem by the following example concerning the Gaussian kernel.

Proposition 1.1. *If $a \in L^2([0,1]^n)$ is not C^∞, then for any $\varepsilon > 0$,*

$$I(a,R) := \inf_{\|b\|_{L^2([0,1]^n)} \le R} \left\{ \left\| a(x) - \int_{[0,1]^n} e^{-\frac{|x-t|^2}{2}} b(t) dt \right\|_{L^2([0,1]^n)} \right\} \neq O(R^{-\varepsilon})\,.$$

Conversely, if $\sigma > 0$ and $a \in H^\sigma([0,1]^n)$, then for $R > C_1$, there holds

$$I(a,R) \le \left\{ \left(\frac{\sqrt{2}\pi}{\ln 2} \sqrt{n} \right)^{\sigma/2} + 64\sqrt{n} \left(\frac{8}{\pi} \right)^n \right\} \|a\|_{\sigma,2} \left(\frac{1}{\ln R} \right)^{\sigma/4},$$

where the constant C_1 will be determined explicitly in Example 6.2.

The first statement in Proposition 1.1 holds for the kernel approximation with other C^∞ kernels. Then we see that the decay of $I(a,R)$ can not be polynomially fast, if a is not C^∞. This is the case for most functions studied in learning theory. For example, for some applications with support vector machines, characteristic functions are used as models for target functions [5, 9, 12]. But these are not in $H^{n/2+\varepsilon}(\varepsilon > 0)$.

The second statement in Proposition 1.1 tells us that for approximated functions in Sobolev spaces, the approximation error $I(a,R)$ decays logarithmically. This is the case for most analytic kernels. Hence the **approximation error** extensively studied in [3] often has **logarithmic convergence rate** for regression functions in Sobolev spaces. This slow convergence is not misleading for learning theory, because the sample error with analytic kernels increases more slowly than that with Sobolev smooth kernels. The numbers of samples we need for the same error and confidence are close for analytic and Sobolev smooth kernels. The detailed analysis for this bias-variance problem is given in Sec. 2.

When the approximated functions have some analytic properties, we may have polynomial decays for the approximation error. The following result deals with kernel representations. The proof will be given in Sec. 4.

Theorem 1.1. *Let A be a compact, symmetric, and strictly positive definite linear operator on a separable Hilbert space B. Then for $0 < \sigma < s$, there holds*

$$\inf_{\|A^{-s} b\| \le R} \|a - b\| \le \left(\frac{1}{R} \right)^{\frac{\sigma}{s-\sigma}} \|A^{-\sigma} a\|^{\frac{s}{s-\sigma}}\,. \tag{1.2}$$

On the other hand, if the kernel function is not smooth, $I(a,R)$ may take polynomial decays for approximated functions in Sobolev spaces. This includes the standard case of Sobolev interpolation spaces. It turns out that for $r>0$, the elements a that satisfy $I(a,R)=O(R^{-r})$ form exactly the interpolation space $(B,\mathcal{H})_{\frac{r}{1+r},\infty}$. This relation has been studied in the interpolation community, see [2], as pointed out to us by Ron DeVore. This is a nice connection. However, except for some well–known function spaces, the K-functional used for defining interpolation spaces (see Sec. 3) may be as hard to estimate as the approximation error. Our analysis also provides the constant in the estimate which plays an important role in learning theory. Let us provide one example to show this situation.

Take X to be an open subset (domain) of $\mathbb{R}^n$, and $B=L^2(X)$. With $s\in\mathbb{Z}_+$, take $\mathcal{H}$ to be the Sobolev space $H^s(X)$ consisting of all functions f in $L^2(X)$ with

$$\|f\|_{H^s(X)} := \sum_{|\alpha|\le s} \left\| \frac{\partial^\alpha f}{\partial x^\alpha} \right\|_{L^2(X)} < \infty\,.$$

When the boundary of X is minimally smooth (for definition, see Stein [11]), $H^s(X)$ can be continuously extended to $H^s(\mathbb{R}^n)$. Hence, by [2], for $0<r<s$, $H^r(X)$ is imbedded in $(L^2(X),H^s(X))_{r/s,\infty}$. Here when r is not an integer, say $r=m+\mu$ with $m\in\mathbb{Z}_+$ and $0<\mu<1$, $H^r(X)$ consists of all functions f in $L^2(X)$ such that the $H^r(X)$ norm is finite:

$$\|f\|_{H^r(X)} := \|f\|_{H^m(X)} + \sum_{|\alpha|=m} \int_{|y|<1} \Bigg(\int_{x,x+y\in X} \Big| D^\alpha f(x+y) - D^\alpha f(x) \Big|^2 dx |y|^{-2\mu-n} dy \Bigg)^{1/2}.$$

The following convergence rate follows from the characterization given in Theorem in Sec. 3.

Proposition 1.2. *Let $0<r<s$ and X be an open subset of $\mathbb{R}^n$ with minimally smooth boundary. Then for $a\in H^r(X)$, we have*

$$\inf_{\|b\|_{H^s(X)}\le R} \{\|a-b\|_{L^2(X)}\} \le \left(\frac{1}{R}\right)^{\frac{r}{s-r}} \left(C_X\|a\|_{H^r(X)}\right)^{\frac{s}{s-r}}\,.$$

Here C_X is a constant depending only on the domain X, which comes from the imbedding property: $\|a\|_{r/s,\infty}\le C_X\|a\|_r$. When X is the whole space $\mathbb{R}^n$, C_X can be taken as $2(s+1)^n$.

In Sec. 3, we characterize elements with polynomially decaying approximation errors in terms of interpolation spaces. This type of convergence rate for $I(a,R)$ can often be realized by elements b depending linearly on a, as shown in Sec. 4. In particular, this is the case when $\mathcal{H}$ is the range of a compact, symmetric, and strictly positive definite linear operator on a separable Hilbert space B, as shown

in Theorem 1.2. Sections 5 and 6 are devoted to a special setting, kernel approximation, where $\mathcal{H}$ will be the range of a Hilbert–Schmidt operator. It turns out that the rate of convergence in this case depends on the regularity of the kernel. When the kernel is analytic (its Fourier transform decays exponentially), we usually have a logarithmic convergence rate.

2. Analysis for Learning Theory

The objective of Learning Theory is to find an unknown function $f : X \to Y$ from random samples $(x_i, y_i)_{i=1}^m$.

Suppose that a probability measure ρ on $Z := X \times Y$ governs the random sampling. Let X be a compact subset of $\mathbb{R}^n$ and $Y = \mathbb{R}$. If we define the (least square) error of f as

$$\mathcal{E}(f) = \int_{X\times Y} (f(x) - y)^2 d\rho\,, \tag{2.1}$$

then the function that minimizes the error is the **regression function** f_ρ:

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x)\,, \qquad x \in X\,.$$

Here $\rho(y|x)$ is the conditional probability measure on $\mathbb{R}$.

As the probability measure ρ is unknown, neither f_ρ nor $\mathcal{E}(f)$ is computable. All we have in hand are the samples $\mathbf{z} := (x_i, y_i)_{i=1}^m$. In Learning Theory, one approximates f_ρ by the function minimizing the **empirical error** $\mathcal{E}_\mathbf{z}$ with respect to the sample $\mathbf{z}$:

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2\,. \tag{2.2}$$

This minimization is taken over functions from a **hypothesis space**. In kernel machine learning, this hypothesis space is often taken to be a ball of a reproducing kernel Hilbert space.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric, and positive definite, *i.e.*, for any finite set $\{x_1, \ldots, x_m\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^m$ is a positive definite matrix. We call K a **Mercer kernel**.

The Reproducing Kernel Hilbert Space $\mathcal{H}_K$ associated with the kernel K is defined (see [1]) to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product satisfying

$$\langle K_x, f\rangle_{\mathcal{H}_K} = f(x)\,, \qquad \forall x \in X, f \in \mathcal{H}_K\,. \tag{2.3}$$

An equivalent definition can be given by means of the square root of a Hilbert–Schmidt operator associated with the kernel K. Let μ be a Borel measure on X.

Define the integral operator L_K as

$$L_K f(x) = \int_X K(x,t) f(t) d\mu(t)\,, \qquad x \in X, f \in L^2_\mu(X)\,. \tag{2.4}$$

Then L_K is a positive, compact operator and its range lies in $C(X)$.

If we denote $\{\lambda_j\}_{j=1}^\infty$ as the nonincreasing sequence of eigenvalues of L_K and let $\{\phi_j\}$ be the corresponding eigenfunctions, then

$$K(x,t) = \sum_{j=1}^\infty \lambda_j \phi_j(x)\phi_j(t)\,, \tag{2.5}$$

where the series converges uniformly and absolutely.

Take $L_K^{1/2}$ to be the linear operator on $L^2_\mu(X)$ satisfying $L_K^{1/2} L_K^{1/2} = L_K$, that is, $L_K^{1/2}(\phi_j) = \sqrt{\lambda_j}\phi_j$ for each j. Then $\mathcal{H}_K = L_K^{1/2}(L^2_\mu(X))$, and $\|f\|_K = \|(L_K^{1/2})^{-1} f\|_{L^2_\mu(X)}$. This space can be imbedded into $C(X)$, and we denote the inclusion as $I_K : H_K \to C(X)$.

Let $R > 0$ and B_R be the ball of $\mathcal{H}_K$ with radius R:

$$B_R := \{f \in H_K : \|f\|_K \le R\}\,.$$

Then $I_K(B_R)$ is a subset of $C(X)$. Denote its closure in $C(X)$ as $\overline{I_K(B_R)}$. Then it is a compact subset of $C(X)$, and we take it as our hypothesis space. For these facts, see [3].

The procedure of regularized empirical minimization in kernel machine learning is as follows.

Given the random samples $\mathbf{z} := (x_i, y_i)_{i=1}^m$, we choose some $R > 0$ and take the hypothesis space as $\overline{I_K(B_R)}$. Then the function $f_\mathbf{z}$ that minimizes the empirical error (2.2) is

$$f_\mathbf{z}(x) = \sum_{j=1}^m c_j K(x, x_j)\,,$$

where the coefficients $(c_j)_{j=1}^m$ is solved by the linear system:

$$\sum_{j=1}^m K(x_i, x_j) c_j = y_i\,, \qquad i = 1, \ldots, m\,.$$

Take $f_\mathbf{z}$ as an approximation of the regression function f_ρ.

The main question for the above learning procedure is:

How many samples do we need to draw to assert, with a confidence greater than $1 - \delta$, *that* $\int_X (f_\mathbf{z} - f_\rho)^2$ *is not more than* ε?

To answer the above question, we decompose the error into two parts: the approximation error and the sample error. Note that [3]

$$\int_X (f_\mathbf{z} - f_\rho)^2 = \mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\,.$$

We only need to analyze $\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)$.

Since $\overline{I_K(B_R)}$ is compact, by [3] there is a function f_R, called the **target function** in $\overline{I_K(B_R)}$, minimizing the error $\mathcal{E}(f)$ over $f \in \overline{I_K(B_R)}$, *i.e.*, an optimizer of

$$\min_{f\in\overline{I_K(B_R)}} \int_{X\times Y}(f(x)-y)^2 d\rho = \inf_{f\in B_R}\int_{X\times Y}(f(x)-y)^2 d\rho\,.$$

The **approximation error** is defined as $\mathcal{E}(f_R)$ and is equal to

$$\mathcal{E}(f_R) = \int_X (f_R - f_\rho)^2 + \mathcal{E}(f_\rho)\,.$$

The approximation error decreases as R becomes larger, and $\mathcal{E}(f_\rho)$ is a fixed constant.

The **sample error** of a function f in $\overline{I_K(B_R)}$ is defined as

$$\mathcal{E}_R(f) := \mathcal{E}(f) - \mathcal{E}(f_R)\,.$$

Thus, the error $\int_X (f_{\mathbf{z}} - f_\rho)^2$ can be decomposed as

$$\int_X (f_{\mathbf{z}} - f_\rho)^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \mathcal{E}_R(f_{\mathbf{z}}) + \int_X (f_R - f_\rho)^2\,.$$

The sample error has been extensively investigated in statistical learning theory [3, 6, 9, 12]. In our situation, we apply the estimate presented by Cucker and Smale [3, Theorem C*] and obtain

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{E}_R(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \mathcal{N}\left(\overline{I_K(B_R)}, \frac{\varepsilon}{24M}\right) e^{-\frac{m\varepsilon}{288M^2}}\,, \qquad \forall \varepsilon > 0\,, \quad (2.6)$$

if $|f(x)-y| \le M$ a.e. for all $f \in \overline{I_K(B_R)}$. Here $\mathcal{N}(\overline{I_K(B_R)}, \frac{\varepsilon}{24M})$ is the covering number, *i.e.*, the minimal integer l such that there exist l disks with radius $\varepsilon/(24M)$ covering the compact set $\overline{I_K(B_R)}$.

The approximation error can be derived from our analysis for the kernel approximation. Recall that

$$\int_X (f_R - f_\rho)^2 = \inf_{\|f\|_K \le R} \|f_\rho - f\|_{L^2}^2 = \inf_{\|L_K^{-1/2} f\|_{L^2}\le R} \|f_\rho - f\|_{L^2}^2\,.$$

Then the approximation error can be analyzed from kernel approximation by the following general result with $\sigma = 1/2, s = 1$.

Theorem 2.1. *Let A be a compact, symmetric, and strictly positive definite linear operator on a separable Hilbert space B. Let $0 < \sigma < s$ and $0 \ne a \in B$. Then for all $R > 0$, there holds*

$$\inf_{\|A^{-\sigma} b\| \le R} \|a - b\| \le \inf_{\|A^{-s} c\| \le 2^{-\frac{s}{2\sigma}} \|a\|^{\frac{\sigma - s}{\sigma}} R^{\frac{s}{\sigma}}} \|a - c\|\,. \quad (2.7)$$

Proof. Let $\{\nu_j\}_{j=1}^{\infty}$ be the non-increasing sequence of eigenvalues of A corresponding to orthonormal eigenvectors $\{u_j\}$. Then $\{u_j\}_{j=1}^{\infty}$ forms an orthonormal basis of B, and each $a \in B$ can be represented as $a = \sum_{j=1}^{\infty} a_j u_j$ with $\|a\| = \sqrt{\sum_{j=1}^{\infty} |a_j|^2}$. For $s > 0$,

$$\|A^{-s}a\| = \left\| \sum_{j=1}^{\infty} a_j \nu_j^{-s} u_j \right\| = \left(\sum_{j=1}^{\infty} \nu_j^{-2s} |a_j|^2 \right)^{1/2} .$$

For $R > 0$, we choose $n \in \mathbb{N}$ such that

$$\nu_1 \geq \nu_2 \geq \cdots \geq \nu_{n-1} \geq (\sqrt{2}\|a\|/R)^{1/\sigma} > \nu_n \geq \nu_{n+1} \geq \cdots$$

(Choose $n = 1$ if $(\sqrt{2}\|a\|/R)^{1/\sigma} > \nu_1$).

For $c = \sum_{j=1}^{\infty} c_j u_j \in B$ with $\|A^{-s}c\| \leq 2^{-s/(2\sigma)} \|a\|^{(\sigma-s)/\sigma} R^{s/\sigma}$, we set

$$b = \sum_{j=1}^{n-1} a_j u_j + \sum_{j=n}^{\infty} c_j u_j .$$

Then

$$\|a - b\| = \left\| \sum_{j=n}^{\infty} a_j u_j - \sum_{j=n}^{\infty} c_j u_j \right\| \leq \|a - c\|$$

and

$$\|A^{-\sigma}b\| \leq \Bigg((\sqrt{2}\|a\|/R)^{(1/\sigma)(-2\sigma)} \sum_{j=1}^{n-1} |a_j|^2 + (\sqrt{2}\|a\|/R)^{(1/\sigma)2(s-\sigma)} \sum_{j=n}^{\infty} \nu_j^{-2s} |c_j|^2 \Bigg)^{1/2} \leq R .$$

Therefore, by taking the infimum over c, we have

$$\inf_{\|A^{-\sigma}b\| \leq R} \|a - b\| \leq \inf_{\|A^{-s}c\| \leq 2^{-s/(2\sigma)} \|a\|^{(\sigma-s)/\sigma} R^{s/\sigma}} \|a - c\| .$$

This proves (2.7) and Theorem 2.1. □

Now we can explain the bias-variance problem for choosing the parameter R. Here we assume that $X = [0,1]^n$ and the marginal probability measure ρ_X of ρ on X is the Lebesgue measure.

Suppose $f_\rho \in H^\sigma(X)$ for some $\sigma > 0$. To find an optimal value for the parameter R, we require that

$$\int_X (f_R - f_\rho)^2 \leq \varepsilon/2 \qquad \text{and} \qquad \mathcal{E}_R(f_{\mathbf{z}}) \leq \varepsilon/2 .$$

When we use the Gaussian kernel with $k(x) = e^{|x|^2/2}$, the above first requirement is satisfied for sufficiently small ε if we choose

$$R = \sqrt{2\|f_\rho\|_2}\,\exp\left\{\frac{1}{2}\left(\frac{\sqrt{2}\|f_\rho\|_{\sigma,2}}{\sqrt{\varepsilon}}\left\{\left(\frac{\sqrt{2}\pi}{\ln 2}\sqrt{n}\right)^{\sigma/2} + 64\sqrt{n}\left(\frac{8}{\pi}\right)^n\right\}\right)^{4/\sigma}\right\}.$$

This follows from Proposition 1.1 and Theorem 2.1 with $a = f_\rho, A = L_K$.

With this choice, we can apply (2.6) to find m. To this end, we need the estimate for the covering number from [14, Proposition 1.1]:

$$\ln \mathcal{N}(\overline{I_K(B_R)}, \eta) \le 4^n(6n+2)\left(\ln\frac{R}{\eta}\right)^{n+1}.$$

This, in connection with (2.6), tells us that $\mathcal{E}_R(f_{\mathbf{z}}) \le \varepsilon/2$ holds with a confidence at least

$$1 - \exp\left\{4^n(6n+2)\left(\ln\frac{48MR}{\varepsilon}\right)^{n+1}\right\} e^{-\frac{m\varepsilon}{576M^2}}.$$

Therefore, to assert, with a confidence greater than $1-\delta$, that $\int_X (f_{\mathbf{z}} - f_\rho)^2 \le \varepsilon$, we only need m samples with

$$m \ge \frac{576M^2}{\varepsilon}\left\{4^n(6n+2)\left(\frac{1}{2}\left(\sqrt{2}\|f_\rho\|_{\sigma,2}\left\{\left(\frac{\sqrt{2}\pi}{\ln 2}\sqrt{n}\right)^{\sigma/2} + 64\sqrt{n}\left(\frac{8}{\pi}\right)^n\right\}\right)^{4/\sigma}\varepsilon^{-\frac{2}{\sigma}}\right.$$
$$\left. + \ln(1/\varepsilon) + \ln\left(48M\sqrt{2\|f_\rho\|_2}\right)\right)^{n+1} - \ln\delta\right\}.$$

Thus, when we use the Gaussian kernel, the number of samples we need to draw is

$$O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2n+2}{\sigma}+1}\right).$$

Now if we use a Mercer kernel $K(x,t) = k(x-t)$ with $k \in C^{h-n}$ and

$$\hat{k}(\xi) \ge C_0(|\xi|+1)^{-h},$$

where C_0 is a constant and $h > \max\{n, 2\sigma\}$, then Theorem 2.1 in connection with our analysis for kernel approximation tells us that

$$\inf_{f\in B_R}\|f_\rho - f\|_2 \le C_h R^{\frac{2(n-h)\sigma}{h(h+\theta_{n,\sigma})}}.$$

Here C_h is a constant independent of R, and

$$\theta_{n,\sigma} := \begin{cases} n/2-\sigma, & \text{if} \quad \sigma > n/2, \\ 0, & \text{if} \quad \sigma < n/2. \end{cases}$$

Therefore, in order that $\int_X (f_R - f_\rho)^2 \le \varepsilon/2$, it is sufficient to choose

$$R = \left(\frac{2C_h^2}{\varepsilon}\right)^{\frac{h(h+\theta_{n,\sigma})}{4(h-n)\sigma}}.$$

The covering number for this kernel can be bounded as

$$\ln \mathcal{N}(\overline{I_K(B_R)}, \eta) \leq C'_h \left(\frac{R}{\eta}\right)^{2n/(h-n)}$$

for a constants $C'_h > 0$ independent of R, η. Then by (2.6), $\mathcal{E}_R(f_{\mathbf{z}}) \leq \varepsilon/2$ holds with a confidence at least

$$1 - \exp\left\{C'_h \left(\frac{48MR}{\varepsilon}\right)^{2n/(h-n)}\right\} e^{-\frac{m\varepsilon}{576M^2}}.$$

Then in the same way, to assert, with a confidence greater than $1-\delta$, that $\int_X (f_{\mathbf{z}} - f_\rho)^2 \leq \varepsilon$, we only need m samples with

$$m \geq \frac{576M^2}{\varepsilon}\left\{C'_h \left(48M\left(2C_h^2\right)^{\frac{h(h+\theta_{n,\sigma})}{4(h-n)\sigma}}\right)^{2n/(h-n)} \left(1/\varepsilon\right)^{\frac{n(h(h+\theta_{n,\sigma})+4(h-n)\sigma)}{2(h-n)^2\sigma}} - \ln\delta\right\}.$$

Thus, the number m of samples we need to draw is

$$O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{n(h(h+\theta_{n,\sigma})+4(h-n)\sigma)}{2(h-n)^2\sigma}+1}\right).$$

We have seen that the numbers of samples we need to draw would be close when we use Gaussian or other Sobolev smooth kernels, though the intermediate parameter R used has totally different orders.

3. Approximation Error and Interpolation Spaces

The elements with polynomially decaying approximation errors can be characterized by interpolation spaces. This was verified in the interpolation community [2]. For completeness, we state this result (Theorem 3.1) and give a detailed proof here. More general structures were considered in [8] where two or more error criteria were controlled simultaneously; and properties of optimal elements were discussed.

The interpolation spaces are defined by means of the K-functional. The **K-functional** of the couple $(B, \mathcal{H})$ is defined by Peetre [10] for $a \in B$ as

$$K(a,t) := \inf_{b\in\mathcal{H}} \{\|a-b\| + t\|b\|_{\mathcal{H}}\}, \quad t > 0. \tag{3.1}$$

It can be easily seen that for the fixed $a \in B$, the function $K(a,t)$ is continuous, non-decreasing, bounded by $\|a\|$, and tends to zero as $t \to 0$. The interpolation spaces are defined according to the convergence rate of this function. For $0 < \theta < 1$ and $1 \leq p \leq \infty$, the interpolation space $(B, \mathcal{H})_{\theta,p}$ consists of all the elements $a \in B$ such that the norm

$$\|a\|_{\theta,p} := \begin{cases} \sup_{t>0}\{K(a,t)/t^\theta\}, & \text{if } p = \infty \\ \left\{\int_0^\infty \left(K(a,t)/t^\theta\right)^p dt/t\right\}^{1/p}, & \text{if } 1 \leq p < \infty \end{cases}$$

is finite. As the norm $\|a\|_{\theta,p}$ is equivalent to the l_p-norm of the sequence $\{K(a,2^j)/2^{j\theta}\}_{j\in\mathbb{Z}}$, we know that $(B,\mathcal{H})_{\theta,p}$ is imbedded in $(B,\mathcal{H})_{\theta,\infty}$.

With the concept of interpolation spaces, the statement about the polynomial decay of the approximation error (see [2]) can be stated as follows.

Theorem 3.1. *Let $(B, \|\cdot\|)$ be a Banach space and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a dense subspace with $\|b\| \le \|b\|_{\mathcal{H}}$ for $b \in \mathcal{H}$. Let $0 < \theta < 1$. If $a \in (B, \mathcal{H})_{\theta,\infty}$, then*

$$I(a, R) = \inf_{\|b\|_{\mathcal{H}} \le R} \{\|a - b\|\} \le \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}} \left(\|a\|_{\theta,\infty}\right)^{\frac{1}{1-\theta}}. \tag{3.2}$$

Conversely, if $I(a, R) \le C(1/R)^{\theta/(1-\theta)}$ for all $R > 0$, then $a \in (B, \mathcal{H})_{\theta,\infty}$ and

$$\|a\|_{\theta,\infty} \le 2C^{1-\theta}.$$

Proof. Consider the function $f(t) := K(a, t)/t$. It is continuous on $(0, +\infty)$. As $K(a, t) \le \|a\|$ (by taking $b = 0$), $\inf_{t>0}\{f(t)\} = 0$.

Fix $R > 0$. Let us first prove (3.2) under the assumption that

$$\sup_{t>0}\{f(t)\} \ge R.$$

In this case, for any $0 < \varepsilon < 1$ there exists some $t_{R,\varepsilon} \in (0, +\infty)$ such that

$$f(t_{R,\varepsilon}) = \frac{K(a, t_{R,\varepsilon})}{t_{R,\varepsilon}} = (1 - \varepsilon)R.$$

By the definition of the K-functional, we can find $b_\varepsilon \in \mathcal{H}$ such that

$$\|a - b_\varepsilon\| + t_{R,\varepsilon}\|b_\varepsilon\|_{\mathcal{H}} \le K(a, t_{R,\varepsilon})/(1 - \varepsilon).$$

It follows that

$$\|b_\varepsilon\|_{\mathcal{H}} \le \frac{K(a, t_{R,\varepsilon})}{(1 - \varepsilon)t_{R,\varepsilon}} = R$$

and

$$\|a - b_\varepsilon\| \le \frac{K(a, t_{R,\varepsilon})}{1 - \varepsilon}.$$

Observe that

$$\frac{K(a, t_{R,\varepsilon})}{t_{R,\varepsilon}^{\theta}} \le \|a\|_{\theta,\infty}.$$

Then

$$\|a - b_\varepsilon\| \le \left[\frac{K(a, t_{R,\varepsilon})}{(1 - \varepsilon)t_{R,\varepsilon}}\right]^{\frac{-\theta}{1-\theta}} \left[\frac{K(a, t_{R,\varepsilon})}{(1 - \varepsilon)t_{R,\varepsilon}^{\theta}}\right]^{\frac{1}{1-\theta}} \le R^{\frac{-\theta}{1-\theta}} \left(\frac{1}{1 - \varepsilon}\right)^{\frac{1}{1-\theta}} \left(\|a\|_{\theta,\infty}\right)^{\frac{1}{1-\theta}}.$$

Thus,

$$I(a,R) \le \inf_{0<\varepsilon<1}\{\|a-b_\varepsilon\|\} \le \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}} \left(\|a\|_{\theta,\infty}\right)^{\frac{1}{1-\theta}},$$

i.e., (3.2) holds under the assumption (3.3).

Turn to the case when (3.3) is not true. Then for any $0 < \varepsilon < 1 - \sup_{u>0}\{f(u)\}/R$ and any $t > 0$, there exists some $b_{t,\varepsilon} \in \mathcal{H}$ such that

$$\|a - b_{t,\varepsilon}\| + t\|b_{t,\varepsilon}\|_{\mathcal{H}} \le K(a,t)/(1-\varepsilon)\,.$$

This implies that

$$\|b_{t,\varepsilon}\|_{\mathcal{H}} \le \frac{K(a,t)}{(1-\varepsilon)t} \le \frac{1}{1-\varepsilon}\sup_{u>0}\{f(u)\} < R$$

and

$$\|a - b_{t,\varepsilon}\| \le \frac{K(a,t)}{1-\varepsilon}\,.$$

Hence

$$I(a,R) \le \inf_{t>0}\{\|a-b_{t,\varepsilon}\|\} \le \inf_{t>0}\{K(a,t)\}/(1-\varepsilon) = 0\,.$$

This proves (3.2) when (3.3) does not hold. Therefore, (3.2) is always valid.

Conversely, suppose that $I(a,R) \le C(1/R)^{\theta/(1-\theta)}$ for $R > 0$. Let $t > 0$. Choose $R_t = (C/t)^{1-\theta}$. Then for any $\varepsilon > 0$ we can find $b_{t,\varepsilon} \in \mathcal{H}$ such that

$$\|b_{t,\varepsilon}\|_{\mathcal{H}} \le R_t \qquad \text{and} \qquad \|a - b_{t,\varepsilon}\| \le C(1/R_t)^{\theta/(1-\theta)}(1+\varepsilon)\,.$$

It follows that

$$K(a,t) \le \|a - b_{t,\varepsilon}\| + t\|b_{t,\varepsilon}\|_{\mathcal{H}} \le C(1/R_t)^{\theta/(1-\theta)}(1+\varepsilon) + tR_t \le 2(1+\varepsilon)C^{1-\theta}t^{\theta}\,.$$

Since ε can be arbitrarily small, we have

$$K(a,t) \le 2C^{1-\theta}t^{\theta}\,.$$

Thus,

$$\|a\|_{\theta,\infty} = \sup_{t>0}\{K(a,t)/t^{\theta}\} \le 2C^{1-\theta} < \infty\,.$$

The proof of Theorem 3.1 is complete. □

Remark 3.1. *The above proof shows that if $a \in \mathcal{H}$, then $I(a,R) = 0$ for $R > \|a\|_{\mathcal{H}}$. Also, $I(a,R) = O((1/R)^{\theta/(1-\theta)})$ if and only if $a \in (B,\mathcal{H})_{\theta,\infty}$.*

4. Linear Realization

The approximation error (3.2) can often be realized by linear operators in many circumstances. To see this, let us first give the proof of Theorem 1.2.

Proof of Theorem 1. Let $\{\nu_j\}, \{u_j\}$ be as in the proof of Theorem 2.1. Set $\mathcal{H} = A^s B$ and $\|A^s a\|_{\mathcal{H}} = \|a\|$, *i.e.*, $\|b\|_{\mathcal{H}} = \|A^{-s} b\|$ for $b \in \mathcal{H}$.

We define a sequence of linear operators as the orthogonal projections $\{P_n\}$:

$$P_n\Big(\sum_{j=1}^{\infty} a_j u_j\Big) = \sum_{j=1}^{n} a_j u_j\,.$$

Let $R > 0$ and $0 \neq a \in B$ such that $\|A^{-\sigma} a\| < \infty$. Then there exists some $n \in \mathbb{N}$ such that

$$\nu_n^{\sigma-s}\|A^{-\sigma}a\| \leq R < \nu_{n+1}^{\sigma-s}\|A^{-\sigma}a\|\,.$$

(Set $n = 0$ and $P_0 a := 0$ when $R < \nu_1^{\sigma-s}\|A^{-\sigma}a\|$.)

By the properties of the orthogonal projections, we see that

$$\|a - P_n a\| = \sqrt{\sum_{j=n+1}^{\infty} |a_j|^2} \leq \nu_{n+1}^{\sigma} \sqrt{\sum_{j=n+1}^{\infty} \nu_j^{-2\sigma}|a_j|^2} \leq \nu_{n+1}^{\sigma}\|A^{-\sigma}a\|$$

and

$$\|P_n a\|_{\mathcal{H}} = \sqrt{\sum_{j=1}^{n} \nu_j^{-2s}|a_j|^2} \leq \nu_n^{\sigma-s}\|A^{-\sigma}a\|\,.$$

Therefore, $\|A^{-s}P_n a\| = \|P_n a\|_{\mathcal{H}} \leq R$, and

$$\|a - P_n a\| \leq \left(\frac{\|A^{-\sigma}a\|}{R}\right)^{\frac{\sigma}{s-\sigma}} \|A^{-\sigma}a\| = \left(\frac{1}{R}\right)^{\frac{\sigma}{s-\sigma}} \|A^{-\sigma}a\|^{\frac{s}{s-\sigma}}\,.$$

This proves that (1.2) holds when we choose b to be the orthogonal projection $P_n a$.

From the above proof, we can see that the approximation error stated in Theorem 1.2 is realized by the linear operators: orthogonal projections. To establish this for more general linear operators, we need Jackson and Bernstein inequalities.

Let $\{L_n\}_{n\in\mathbb{N}}$ be a sequence of linear operators on B (into $\mathcal{H}$), and $\{\lambda_n\}_{n\in\mathbb{N}}$ be a non-increasing sequence of positive numbers tending to zero. With a fixed constant $C > 0$, the Jackson inequality takes the form

$$\|L_n b - b\| \leq C\lambda_{n+1}\|b\|_{\mathcal{H}}, \qquad \forall b \in \mathcal{H}\,. \tag{4.1}$$

The Bernstein inequality we need is

$$\|L_n a\|_{\mathcal{H}} \leq \begin{cases} C\lambda_n^{-1}\|a\|, & \text{if } a \in B\,, \\ C\|a\|_{\mathcal{H}}, & \text{if } a \in \mathcal{H}\,. \end{cases} \tag{4.2}$$

When the Jackson and Bernstein inequalities hold, the approximation error (3.2) can be realized by the linear operators $\{L_n\}$ (up to a constant). The following result is derived from techniques in approximation theory, *e.g.*, [4].

Theorem 4.1. *Let $C > 0$ and $\{\lambda_n\}_{n\in\mathbb{N}}$ be a non-increasing sequence of positive numbers tending to zero. Suppose that a sequence of linear operators $\{L_n\}$ on B satisfies* (4.1) *and* (4.2)*, and $\|L_n\| \le C$. Let $0 < \theta < 1$ and $a \in (B, \mathcal{H})_{\theta,\infty}$. Then for $R \ge C\lambda_1^{\theta-1}\|a\|_{\theta,\infty}$, there exists some $n \in \mathbb{N}$ such that*

$$\|L_n a\|_{\mathcal{H}} \le R$$

and

$$\|L_n a - a\| \le (C+1)C^{\theta/(1-\theta)}(1/R)^{\theta/(1-\theta)}\|a\|_{\theta,\infty}^{1/(1-\theta)}.$$

Proof. Let $a \in (B, \mathcal{H})_{\theta,\infty}$ and $b \in \mathcal{H}$. By the Jackson inequality (4.1), for $n \in \mathbb{N}$,

$$\|L_n a - a\| \le \|L_n(a-b)\| + \|L_n b - b\| + \|a-b\| \le (C+1)\{\|a-b\| + \lambda_{n+1}\|b\|_{\mathcal{H}}\}.$$

Taking the infimum over $b \in \mathcal{H}$, we obtain

$$\|L_n a - a\| \le (C+1)K(a, \lambda_{n+1}) \le (C+1)\lambda_{n+1}^{\theta}\|a\|_{\theta,\infty}.$$

In the same way, by the Bernstein inequality,

$$\|L_n a\|_{\mathcal{H}} \le \|L_n(a-b)\|_{\mathcal{H}} + \|L_n b\|_{\mathcal{H}} \le C\lambda_n^{-1}\{\|a-b\| + \lambda_n\|b\|_{\mathcal{H}}\}.$$

Taking the infimum again, we have

$$\|L_n a\|_{\mathcal{H}} \le C\lambda_n^{-1}K(a, \lambda_n) \le C\lambda_n^{\theta-1}\|a\|_{\theta,\infty}.$$

Now let $R \ge C\lambda_1^{\theta-1}\|a\|_{\theta,\infty}$, then there exists some $n \in \mathbb{N}$ such that

$$C\lambda_n^{\theta-1}\|a\|_{\theta,\infty} \le R < C\lambda_{n+1}^{\theta-1}\|a\|_{\theta,\infty}.$$

For this n, there holds

$$\|L_n a\|_{\mathcal{H}} \le C\lambda_n^{\theta-1}\|a\|_{\theta,\infty} \le R.$$

Moreover, $\lambda_{n+1} \le \left(\frac{C\|a\|_{\theta,\infty}}{R}\right)^{\frac{1}{1-\theta}}$. Hence

$$\|L_n a - a\| \le (C+1)\lambda_{n+1}^{\theta}\|a\|_{\theta,\infty} \le (C+1)C^{\theta/(1-\theta)}(1/R)^{\theta/(1-\theta)}\|a\|_{\theta,\infty}^{1/(1-\theta)}.$$

Thus, Theorem 4.1 is true. □

Notice that the orthogonal projections $\{P_n\}$ in the proof of Theorem 1.2 satisfy the Jackson and Bernstein inequalities with $\lambda_n = \nu_n^s$ and $C = 1$. Hence Theorem 4.1 would yield the desired rate of convergence in Theorem 1.2. But to see our exact estimate with the constant 1, we need more refined analysis as given before.

5. Kernel Approximation

Good algorithms generated by kernels play an important role in the learning theory [3, 5, 9, 12]. Let us investigate kernel approximation in what follows.

Let X be a complete metric space and μ be a Borel measure on X. Denote $L^2_\mu(X)$ as the Hilbert space of (real) square integrable functions with the inner product

$$\langle f, g\rangle = \int_X f(x)g(x)d\mu(x)\,.$$

Suppose that $K : X \times X \to \mathbb{R}$ is symmetric and positive definite, *i.e.*, $\Big(K(x_i, x_j)\Big)_{i,j=1}^m$ is a positive definite matrix for any finite set $\{x_1, \ldots, x_m\} \subset X$. Assume that $K \in L^2_{\mu\times\mu}$, *i.e.*, $\int_X \int_X |K(x,t)|^2 d\mu(x)d\mu(t) < \infty$. Then the Hilbert–Schmidt linear operator $L_K : L^2_\mu(X) \to L^2_\mu(X)$ defined by

$$L_K f(x) = \int_X K(x,t)f(t)d\mu(t)$$

is symmetric, compact, and positive definite.

As we saw in Section 2, Learning Theory raises the problem of estimating $I(a, R)$, where $B = L^2_\mu(X), \mathcal{H} = L_K(L^2_\mu(X))$ and $\|L_K f\|_\mathcal{H} = \|f\| = \|L_K^{-1}(L_K f)\|$. In particular, we want to know whether $I(a, R) = O((1/R)^{\theta/(1-\theta)})$ for some $\theta > 0$ when a is the characteristic function of a regular domain. For this purpose, we need to understand the interpolation space $(B, \mathcal{H})_{\theta,\infty}$, and see whether it contains characteristic functions.

It turns out that the problem is deeply involved with the regularity of the kernel K. To see this, let us compare the interpolation spaces with some well–known function spaces (Besov, Sobolev spaces).

Theorem 5.1. *Let $(W, \|\cdot\|_W)$ be a Banach space of functions over X. Suppose that*

$$\|K\|_{W\times\mu} := \left\{ \int_X \|K(\cdot, t)\|^2_W d\mu(t) \right\}^{1/2} < \infty\,.$$

Then for $f \in L^2_\mu(X)$,

$$\|L_K f\|_W \le \|K\|_{W\times\mu}\|f\| = \|K\|_{W\times\mu}\|L_K f\|_\mathcal{H}\,.$$

Hence $\mathcal{H}$ is imbedded in W, and for $0 < \theta < 1$, $(B, \mathcal{H})_{\theta,\infty} \subset (B, W)_{\theta,\infty}$.

Proof. By the property of a norm, we have

$$\|L_K f\|_W \le \int_X |f(t)|\,\|K(\cdot, t)\|_W d\mu(t)\,.$$

Then the desired inequality follows from the Schwartz inequality:

$$\|L_K f\|_W \le \sqrt{\int_X |f(t)|^2 d\mu(t)}\sqrt{\int_X \|K(\cdot, t)\|^2_W d\mu(t)} = \|K\|_{W\times\mu}\|f\|\,.$$

Since $\mathcal{H} = L_K(L^2_\mu(X))$ and $\|L_K f\|_\mathcal{H} = \|f\|$, the imbedding property follows. □

As a corollary, we choose μ to be the Lebesgue measure over a domain $X \subset \mathbb{R}^n$, and let W be the Sobolev space $H^s(X)$. Combining Theorem 5 with Theorem 3, we obtain.

Corollary 5.1. *Let $s > 0$, $K : X \times X \to \mathbb{R}$ be symmetric and positive definite such that*

$$\|K\|_{H^s \times L_2} := \left\{ \int_X \|K(\cdot, t)\|^2_{H^s(X)} dt \right\}^{1/2} < \infty .$$

If $0 < \theta < 1$, and $a \in L^2(X)$ satisfies

$$I(a, R) = \inf_{\|f\|_{L^2(X)} \le R} \{\|a - L_K f\|_{L^2(X)}\} = O\left(\left(\frac{1}{R} \right)^{\frac{\theta}{1-\theta}} \right) \quad (R \to \infty) ,$$

then $a \in (L^2(X), H^s(X))_{\theta,\infty}$.

When X has a minimally smooth boundary, $(L^2(X), H^s(X))_{\theta,\infty} \subset H^{s\theta-\eta}(X)$ for any $\eta > 0$.

Let's turn to the examples studied in [3] for the learning theory. Consider only $X = [0, 1]^n$.

Assume that $k : [-1, 1]^n \to \mathbb{R}$ is C^∞ and symmetric about the origin. Set

$$K(x, t) = k(x - t), \qquad x, t \in [0, 1]^n .$$

Special examples of this kind include the Gaussian $k(x) = e^{-\frac{|x|^2}{c^2}}$, and $k(x) = (c^2 + |x|^2)^{-\alpha}$ with $c > 0, \alpha > 0$.

Choose s to be a positive integer. Then

$$\|K\|_{H^s([0,1]^n) \times L^2} = \left\{ \int_{[0,1]^n} \|k(\cdot - t)\|^2_{H^s([0,1]^n)} dt \right\}^{1/2} \le \|k\|_{H^s([-1,1]^n)} < \infty .$$

It follows from Corollary 1 that $I(a, R) = O(R^{-r})$ implies $a \in (L^2, H^s)_{\frac{r}{1+r},\infty}$ which is imbedded in $H^{\frac{rs}{1+r}-\eta}([0, 1]^n)$ for any $\eta > 0$. Thus, if $a \in L^2([0, 1]^n)$ is not C^∞, then for any $\varepsilon, \eta > 0$, we can choose some $s \in \mathbb{N}$ such that $a \notin H^{\frac{\varepsilon s}{1+\varepsilon}-\eta}([0, 1]^n)$, hence $I(a, R) \neq O(R^{-\varepsilon})$. As a corollary, we have proved the first statement of Proposition 1.1.

Consider the characteristic function of a proper rectangular subset $\Pi_{j=1}^n [c_j, d_j]$. Its Fourier transform is $\Pi_{j=1}^n \{[e^{-ic_j\xi_j} - e^{-id_j\xi_j}]/(i\xi_j)\}$. Hence it lies in $H^\sigma(\mathbb{R}^n)$ and then in $H^\sigma([0, 1]^n)$ with $0 < \sigma < 1/2$, but not in $H^{1/2}([0, 1]^n)$. Therefore, for any $r > 0$, $I(a, R) \neq O(R^{-r})$. This is a negative result. In Sec. 6, positive results providing estimates for the decay of $I(a, R)$ will be presented.

Thus we see that the kernel approximation is slow when the kernel function is smooth. Things are different when the kernel function is not smooth.

Example 5.1. *Let k be the characteristic function of the cube $[-1/2, 1/2]^n$. Then for $f \in L^2([0, 1]^n)$,*

$$L_K f(x) = \int_{[0,1]^n} \chi_{[-\frac{1}{2},\frac{1}{2}]^n}(x - t) f(t) dt = \int_{I(x_1)} \cdots \int_{I(x_n)} f(t_1, \ldots, t_n) dt_1 \cdots dt_n ,$$

where $x = (x_1, \ldots, x_n) \in [0,1]^n$, and for $x_j \in [0,1]$,

$$I(x_j) = \begin{cases} \left[0,\, x_j + \dfrac{1}{2}\right], & \text{if} \quad x_j \in \left[0, \dfrac{1}{2}\right], \\ \left[x_j - \dfrac{1}{2}, 1\right], & \text{if} \quad x_j \in \left(\dfrac{1}{2}, 1\right]. \end{cases}$$

It follows that $\|\frac{\partial^n}{\partial x_1 \cdots \partial x_n}(L_K f)\|_{L^2([0,1]^n)} = \|f\|_{L^2([0,1]^n)}$, *and for each* $j = 1, \ldots, n$, *there holds* $\|\frac{\partial}{\partial x_j}(L_K f)\|_{L^2([0,1]^n)} \le \|f\|_{L^2([0,1]^n)}$. *Thus,* $\|g\|_{H^1([0,1]^n)} \le (n+1)\|g\|_{\mathcal{H}}$ *for any* $g \in \mathcal{H} = L_K(L^2([0,1]^n))$, *and* $\|g\|_{\mathcal{H}} \le \|g\|_{H^n([0,1]^n)}$ *for any* $g \in H^n([0,1]^n)$. *It tells that for* $0 < \theta < 1$, $I(a,R) = O(R^{-\theta/(1-\theta)})$ *implies* $a \in (L^2([0,1]^n), H^1([0,1]^n))_{\theta,\infty}$; *while* $a \in (L^2([0,1]^n), H^n([0,1]^n))_{\theta,\infty}$ *implies* $I(a,R) = O(R^{-\theta/(1-\theta)})$.

In particular, in the univariate case, $I(a,R) = O(R^{-\theta/(1-\theta)})$ if and only if $a \in (L^2([0,1]), H^1([0,1]))_{\theta,\infty}$. Also, if a is the characteristic function of a proper rectangular subset $\Pi_{j=1}^n [c_j, d_j]$, then $I(a,R) = O(R^{-r/(2n-1)})$ for any $0 < r < 1$. By taking higher order splines or compactly supported positive definite radial basis functions with high regularity, we can get examples for which $I(a,R) = O(R^{-\theta/(1-\theta)})$ when $a \in (L^2([0,1]^n), H^s([0,1]^n))_{\theta,\infty}$.

6. Logarithmic Rate for Kernel Approximation

In the last section, it is shown that the convergence of kernel approximation is slow when the kernel is smooth. In this section, we give a method for estimating the convergence rate for this case. By our approach, a typical convergence rate will be logarithmic. This is the case for most analytic kernels.

The Sobolev space $H^\sigma(\mathbb{R}^n)$ has an equivalent norm (fractional Sobolev space norm):

$$\|f\|_{\sigma,2} = \left(\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} (|\xi|^2 + 1)^\sigma |\hat{f}(\xi)|^2 d\xi\right)^{1/2} < \infty\,.$$

The corresponding seminorm is

$$|f|_{\sigma,2} := \left(\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} |\xi|^{2\sigma} |\hat{f}(\xi)|^2 d\xi\right)^{1/2}.$$

Let $X = [0,1]^n$. Given a kernel $k \in L^2(\mathbb{R}^n)$, we consider the linear operator L_K: $L^2(X) \to L^2(X)$ defined by

$$L_K f(x) = k * f(x) = \int_X k(x-t) f(t) dt, \qquad x \in X, f \in L^2(X)\,. \tag{6.1}$$

Assume that k is a symmetric function and $\hat{k}(\xi) > 0$ on $\mathbb{R}^n$. Then L_K is symmetric, compact, and positive definite. We are interested in the convergence rate of

the function

$$I(a,R) := \inf_{\|L_K^{-1}b\|_{L^2(X)}\le R}\{\|a-b\|_{L^2(X)}\} = \inf_{\|b\|_{L^2(X)}\le R}\{\|a-L_Kb\|_{L^2(X)}\}\,, \quad (6.2)$$

where $R>0$ and $a\in L^2(X)$.

To state our main result here, we need the following function (with a parameter $\sigma\ge 0$) measuring the decay of $\hat{k}$:

$$\lambda_{k,\sigma}(r) := \left\{\inf_{|\xi|\le r}\{(|\xi|^2+1)^\sigma \hat{k}(\xi)\}\right\}^{-1}, \quad r>0\,. \quad (6.3)$$

We also need the following quantity involving the regularity of the kernel k:

$$\begin{aligned}\varepsilon_k(N) := \sup_{x\in X}\Bigg\{\inf\Bigg\{k(0) - 2\sum_{j\in\{1,\dots,N-1\}^n} w_j k(x-j/N) \\ + \sum_{j,l\in\{1,\dots,N-1\}^n} w_j k\left(\frac{j-l}{N}\right) w_l :\ w_j\in\mathbb{R}\Bigg\}\Bigg\}.\end{aligned} \quad (6.4)$$

This quantity was employed in the study of radial basis functions and variational principle in multivariate approximation theory [7, 13]. By approximating k with its Taylor polynomials, we can see that $\varepsilon_k(N)=O(N^{-m})$, if $k\in C^m(X)$. When k is analytic, $\varepsilon_k(N)$ usually decays exponentially.

As an example, if k is the Gaussian kernel: $k(x)=e^{-|x|^2/2}$, then for sufficiently large r and N, $\lambda_{k,\sigma}(r)=(\sqrt{2\pi})^n(r^2+1)^{-\sigma}e^{r^2/2}$; and $\varepsilon_k(N)\le$ const $e^{-\delta N}$ with some fixed constant $\delta>0$.

Similar to $\lambda_{k,\sigma}$, $\Lambda_{k,\sigma}$ denotes the following increasing function (hence its inverse function $\Lambda_{k,\sigma}^{-1}$ is well defined over $(0,+\infty)$):

$$\Lambda_{k,\sigma}(r) := \left\{\inf_{|\xi|\le r}\hat{k}(\xi)\right\}^{-1}\left(\int_0^r(\rho^2+1)^{-\sigma}n\rho^{n-1}d\rho\right)^{1/2}, \qquad r>0\,. \quad (6.5)$$

Using these functions measuring the regularity of the kernel function, our estimate is given as follows. Denote $[x]$ as the integer part of $x>0$.

Theorem 6.1. *Let $\sigma>0, a\in H^\sigma(\mathbb{R}^n)$ and L_K be given as above. Then for $R\ge 8^n\|a\|_{\sigma,2}\Lambda_{k,\sigma}(\sqrt{n}\pi)$, there holds*

$$I(a,R)\le \|a\|_{\sigma,2}\inf_{0<R'\le\pi N_R}\left\{\left(\frac{1}{R'}\right)^\sigma + 2^n\left(\varepsilon_k(N_R)\lambda_{k,\sigma}(R')\right)^{1/2}\right\}, \quad (6.6)$$

where

$$N_R = \left[\frac{1}{\sqrt{n}\pi}\Lambda_{k,\sigma}^{-1}\left(\frac{R}{8^n\|a\|_{\sigma,2}}\right)\right].$$

Proof. Define a band-limited function f as

$$\hat{f}(\xi) = \begin{cases}\hat{a}(\xi)\,, & \text{if } |\xi|\le R'\,,\\ 0\,, & \text{otherwise}\,,\end{cases}$$

where the band width $R' > 0$ will be determined later. Then

$$\|a - f\|_{L^2(X)} \le \|a - f\|_{L^2(\mathbb{R}^n)} \le \left\{(2\pi)^{-n} \int_{|\xi|>R'} |\hat{a}(\xi)|^2 d\xi\right\}^{1/2} \le |a|_{\sigma,2} \left(\frac{1}{R'}\right)^{\sigma}. \tag{6.7}$$

We shall use the samples $(f(j/N))$ of the function f with a suitable integer $N \in \mathbb{N}$ to define the function b realizing the error estimate in (6.6). Here N satisfies $N\pi \ge R'$ and will be determined later.

To define b, we need a function $\varphi(x) := \Pi_{j=1}^n \varphi_0(x_j)$ on $\mathbb{R}^n$, where

$$\varphi_0(t) = \begin{cases} 1 - \dfrac{3}{2}|t| + \dfrac{1}{2}|t|^3\,, & \text{if } |t| < 1\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Then φ is symmetric, supported on $[-1,1]^n$, and positive definite on $\mathbb{R}^n$:

$$\hat{\varphi}(\xi) = \Pi_{j=1}^n \frac{3}{\xi_j^2}\left\{1 - \frac{2\sin\xi_j}{\xi_j} + \frac{2(1-\cos\xi_j)}{\xi_j^2}\right\} > 0\,.$$

A simple computation shows that $\hat{\varphi}(\xi) > 4^{-n}$ for $\xi \in [-\pi,\pi]^n$.

The function b realizing the error for a is given by

$$b(t) = \sum_{j\in J} d_j \varphi(Nt - j)\,. \tag{6.8}$$

Here J is the index set $J := \{1, 2, \ldots, N-1\}^n$, and $\{d_j\}_{j\in J}$ is a set of coefficients depending on f. With this form, b is supported on X. This implies that

$$L_K b(x) = \int_{\mathbb{R}^n} k(x-t) \sum_{j\in J} d_j \varphi(Nt - j) dt = \sum_{j\in J} d_j \Phi\left(x - \frac{j}{N}\right).$$

Here

$$\Phi(x) = k * \varphi(N\cdot)(x) = \int k(x-t)\varphi(Nt) dt$$

is symmetric, positive definite, and

$$\hat{\Phi}(\xi) = N^{-n} \hat{k}(\xi) \hat{\varphi}(\xi/N) > 0\,.$$

Since Φ is positive definite, the matrix

$$A_N := \left(\Phi\left(\frac{j}{N} - \frac{l}{N}\right)\right)_{j,l\in J}$$

is positive definite. The coefficient vector $d := (d_j)_{j\in J}$ is uniquely determined by the linear system:

$$A_N d = (f(j/N))_{j\in J}\,. \tag{6.9}$$

That is,

$$\sum_{l\in J}\Phi\Big(\frac{j}{N}-\frac{l}{N}\Big)d_l = f\Big(\frac{j}{N}\Big), \qquad j\in J\,.$$

Thus, b has been defined, by means of N and $(f(j/N))$. To find the norm $\|b\|_2$, we notice that φ is supported on $[-1,1]^n$ and $0\le\varphi(t)\le 1$. Then, by the definition (6.8), for $t\in l/N+y/N$ with $l\in\{0,\dots,N-1\}^n, y\in[0,1)^n$,

$$|b(t)|^2 = \Big|\sum_{j\in J\cap(l+\{0,1\}^n)} d_j\varphi(l+y-j)\Big|^2 \le 2^n \sum_{j\in J\cap(l+\{0,1\}^n)}|d_j|^2\,.$$

Hence

$$\int_{l/N+[0,1)^n/N}|b(t)|^2dt \le 2^n\Big\{\sum_{j\in J\cap(l+\{0,1\}^n)}|d_j|^2\Big\}N^{-n}\,.$$

It follows that

$$\|b\|^2_{L^2(X)} = \sum_{l\in\{0,\dots,N-1\}^n}\int_{l/N+[0,1)^n/N}|b(t)|^2dt \le 4^nN^{-n}\sum_{j\in J}|d_j|^2\,.$$

Thus,

$$\|b\|_{L^2(X)} \le 2^nN^{-n/2}\|\{d_j\}_{j\in J}\|_{l^2}\,.$$

To bound the discrete norm $\|d\|_{l^2}$, we take inner products of d with both sides of (6.9) and obtain

$$d^TA_Nd = \sum_{j,l\in J} d_j\Phi\Big(\frac{j-l}{N}\Big)d_l = \sum_{j\in J} d_jf\Big(\frac{j}{N}\Big)\,.$$

On the left side, we use the inverse Fourier transform,

$$d^TA_Nd = (2\pi)^{-n}\int_{\mathbb{R}^n}\hat\Phi(\xi)\Big|\sum_{j\in J}d_je^{i\frac{j}{N}\cdot\xi}\Big|^2d\xi \ge 4^{-n}\Big\{\inf_{\xi\in[-N\pi,N\pi]^n}\hat k(\xi)\Big\}\|d\|^2_{l^2}\,.$$

On the right side, by the Schwartz inequality,

$$\sum_{j\in J}d_jf\Big(\frac{j}{N}\Big) \le \|d\|_{l^2}N^{n/2}\|f\|_\infty\,.$$

Combining the above two estimates, we have

$$\|d\|_{l^2} \le 4^n\Big\{\inf_{\xi\in[-N\pi,N\pi]^n}\hat k(\xi)\Big\}^{-1}N^{n/2}\|f\|_\infty\,.$$

It follows that

$$\|b\|_{L^2(X)} \le 8^n\Big\{\inf_{\xi\in[-N\pi,N\pi]^n}\hat k(\xi)\Big\}^{-1}\|f\|_\infty\,.$$

The support of $\hat{f}$ tells us that

$$\|f\|_\infty \le \|a\|_{\sigma,2}\left\{(2\pi)^{-n}\int_{|\xi|\le R'}\left(|\xi|^2+1\right)^{-\sigma}d\xi\right\}^{1/2}$$

$$\le \|a\|_{\sigma,2}\left(\int_0^{R'}(r^2+1)^{-\sigma}nr^{n-1}dr\right)^{1/2}.$$

Therefore, with the definition (6.5), if $R' \le N\pi$, the norm $\|b\|_2$ can be bounded as

$$\|b\|_{L^2(X)} \le \left\{\inf_{|\xi|\le\sqrt{n}N\pi}\hat{k}(\xi)\right\}^{-1} 8^n\|a\|_{\sigma,2}\left(\int_0^{R'}(r^2+1)^{-\sigma}nr^{n-1}dr\right)^{1/2} \tag{6.10}$$

$$\le 8^n\|a\|_{\sigma,2}\Lambda_{k,\sigma}(\sqrt{n}N\pi)\,.$$

We turn to the estimate of the error $\|a-L_Kb\| \le \|a-f\|+\|f-L_Kb\|$. Note that the matrix A_N is symmetric, so is its inverse $A_N^{-1} = (A_N^{-1})_{j,l\in J}$. Define a set of nodal functions $\{u_j(x)\}_{j\in J}$ by

$$u_j(x) = \sum_{l\in J}(A_N^{-1})_{j,l}\Phi\left(x-\frac{l}{N}\right), \qquad j\in J\,.$$

This, in connection with the construction of L_Kb, implies

$$L_Kb(x) = \sum_{j\in J}\left\{\sum_{l\in J}\left(A_N^{-1}\right)_{j,l}f(l/N)\right\}\Phi\left(x-\frac{j}{N}\right)$$

$$= \sum_{l\in J}f(l/N)\left\{\sum_{j\in J}\left(A_N^{-1}\right)_{l,j}\Phi\left(x-\frac{j}{N}\right)\right\} = \sum_{l\in J}f(l/N)u_l(x)\,.$$

By the inverse Fourier transform,

$$f(x)-L_Kb(x) = (2\pi)^{-n}\int_{\mathbb{R}^n}\hat{f}(\xi)\left\{e^{ix\cdot\xi}-\sum_{l\in J}u_l(x)e^{i\frac{l}{N}\cdot\xi}\right\}d\xi\,.$$

Since f is band-limited, we have

$$|f(x)-L_Kb(x)| \le \left\{(2\pi)^{-n}\int_{\mathbb{R}^n}\frac{|\hat{f}(\xi)|^2}{\hat{\Phi}(\xi)}d\xi\right\}^{1/2}$$

$$\times\left\{(2\pi)^{-n}\int_{\mathbb{R}^n}\hat{\Phi}(\xi)\left|e^{ix\cdot\xi}-\sum_{l\in J}u_l(x)e^{i\frac{l}{N}\cdot\xi}\right|^2 d\xi\right\}^{1/2}.$$

As Φ is symmetric,

$$(2\pi)^{-n}\int_{\mathbb{R}^n}\hat{\Phi}(\xi)\left|e^{ix\cdot\xi}-\sum_{l\in J}u_l(x)e^{i\frac{l}{N}\cdot\xi}\right|^2 d\xi$$

$$= \Phi(0)-2\sum_{l\in J}u_l(x)\Phi\left(x-\frac{l}{N}\right)+\sum_{j,l\in J}u_j(x)\Phi\left(\frac{j}{N}-\frac{l}{N}\right)u_l(x)\,.$$

Let $x \in X$ be fixed. The quadratic function

$$Q((w_j)_{j\in J}) := \Phi(0) - 2\sum_{l\in J} w_l \Phi\left(x - \frac{l}{N}\right) + \sum_{j,l\in J} w_j \Phi\left(\frac{j}{N} - \frac{l}{N}\right) w_l$$

over $\mathbb{R}^J$ takes the minimum at $(u_j(x))_{j\in J}$. For any $(w_j) \in \mathbb{R}^J$, we have

$$\begin{aligned} Q((u_j(x))) \le Q((w_j)) &= (2\pi)^{-n} \int_{\mathbb{R}^n} \hat{\Phi}(\xi) \left| e^{ix\cdot\xi} - \sum_{l\in J} w_l e^{i\frac{l}{N}\cdot\xi} \right|^2 d\xi \\ &\le N^{-n}(2\pi)^{-n} \int_{\mathbb{R}^n} \hat{k}(\xi) \left| e^{ix\cdot\xi} - \sum_{l\in J} w_l e^{i\frac{l}{N}\cdot\xi} \right|^2 d\xi \\ &= N^{-n}\left\{ k(0) - 2\sum_{l\in J} w_l k\left(x - \frac{l}{N}\right) + \sum_{j,l\in J} w_j k\left(\frac{j}{N} - \frac{l}{N}\right) w_l \right\}. \end{aligned}$$

Taking the infimum over $(w_j) \in \mathbb{R}^J$, we know from the definition (6.4) that for any $x \in X$,

$$Q((u_j(x))) \le N^{-n}\varepsilon_k(N)\,.$$

It follows that

$$\|f - L_K b\|_{L^2(X)} \le \{N^{-n}\varepsilon_k(N)\}^{1/2} \|a\|_{\sigma,2} \left\{ \inf_{|\xi|\le R'} (|\xi|^2 + 1)^\sigma \hat{\Phi}(\xi) \right\}^{-1/2}.$$

Therefore, concerning the error, as $R' \le N\pi$, we obtain the following estimate:

$$\|f - L_K b\|_{L^2(X)} \le 2^n \|a\|_{\sigma,2} \left(\varepsilon_k(N) \lambda_{k,\sigma}(R') \right)^{1/2}. \tag{6.11}$$

We are in a position to prove our conclusion, using the estimate (6.10), for $\|b\|$, the error estimate (6.11) and $\|a - f\|$. Let $R \ge 8^n \|a\|_{\sigma,2} \Lambda_{k,\sigma}(\sqrt{n}\pi)$. Since $\Lambda_{k,\sigma}$ is increasing, there exists a positive integer N such that

$$8^n \|a\|_{\sigma,2} \Lambda_{k,\sigma}(\sqrt{n} N\pi) \le R\,,$$

that is,

$$N \le \frac{1}{\sqrt{n}\pi} \Lambda_{k,\sigma}^{-1}\left(\frac{R}{8^n \|a\|_{\sigma,2}} \right).$$

But $\lim_{r\to\infty} \Lambda_{k,\sigma}(r) = +\infty$. The largest integer satisfying this condition is N_R. Then choose $N = N_R$. By (6.10),

$$\|b\|_{L^2(X)} \le R\,.$$

The conclusion (6.6) follows from the error bounds (6.7) and (6.11) by taking the infimum over $0 < R' \le \pi N_R$. The proof of Theorem 6.1 is complete. □

The proof of Theorem 6.1 also provides a way to find the element b for achieving the approximation error.

To see how to handle the functions measuring the regularity of the kernel, and then to estimate the approximation error, we turn to the example of Gaussian kernels stated in the introduction. Denote $p_n := \max\{1/(4\sqrt{n}), 2^{-n}\}$, and for $\sigma > 0$,

$$C_{\sigma,n} := \begin{cases} 2\sigma/(2\sigma - n)\,, & \text{if } \sigma > n/2\,, \\ 1 + n\,, & \text{if } \sigma = n/2\,, \\ n/(n - 2\sigma)\,, & \text{if } 0 < \sigma < n/2\,. \end{cases}$$

Example 6.1. Let $c > 0$ and

$$k(x) = e^{-\frac{|x|^2}{c^2}}\,, \qquad x \in \mathbb{R}^n\,.$$

If $a \in H^\sigma(\mathbb{R}^n)$ and

$$R \geq \max\left\{ \left(8^n \|a\|_{\sigma,2} \sqrt{C_{\sigma,n}} \max\left\{c^{-\frac{3}{2}n} e^{\max\{\frac{c^2}{2}, 8n^2\}}, c^{-n} n^{n/4} e^{\frac{c^2 n\pi^2}{4}}\right\}\right)^2, \right. \\ \left. e^{n\pi^2 c^2 (80n \ln 2/c^2 + 3)^2}, \quad \left(\frac{2\sigma\sqrt{n}\pi c}{-\ln p_n}\right)^4 \right\}, \tag{6.12}$$

then

$$I(a,R) = \inf_{\|b\|_{L^2([0,1]^n)} \leq R} \{\|a - L_K b\|_{L^2([0,1]^n)}\} \leq \|a\|_{\sigma,2} \left\{ \left(\frac{-2\ln p_n}{c^3\sqrt{n}\pi}\right)^{-\sigma/2} \right. \\ \left. + 2^n \left(32n\sqrt{e} + \frac{4^{n+1}}{c\sqrt{\pi}}\right) p_n^{-1} (c\sqrt{\pi})^{-n} \right\} \left(\frac{1}{\ln R}\right)^{\sigma/4}.$$

Proof. It is well known that

$$\hat{k}(\xi) = (c\sqrt{\pi})^n e^{-\frac{c^2|\xi|^2}{4}}\,.$$

Hence $\hat{k}(\xi) > 0$ for any $\xi \in \mathbb{R}^n$. Then

$$\lambda_{k,\sigma}(r) \leq (c\sqrt{\pi})^{-n} e^{\frac{c^2 r^2}{4}}\,.$$

By a simple computation,

$$\int_0^r (\rho^2 + 1)^{-\sigma} n \rho^{n-1} d\rho \leq \begin{cases} C_{\sigma,n}\,, & \text{if } \sigma > n/2\,, \\ C_{\sigma,n}(1 + \ln r)\,, & \text{if } \sigma = n/2\,, r \geq 1\,, \\ C_{\sigma,n} r^{n-2\sigma}\,, & \text{if } 0 < \sigma < n/2\,, r \geq 1\,. \end{cases}$$

Then for $r \geq 1$,

$$\left(\int_0^1 (\rho^2 + 1)^{-\sigma} n \rho^{n-1} d\rho\right)^{1/2} (c\sqrt{\pi})^{-n} e^{\frac{c^2 r^2}{4}} \leq \Lambda_{k,\sigma}(r) \leq (c\sqrt{\pi})^{-n} \sqrt{C_{\sigma,n}}\, r^{n/2} e^{\frac{c^2 r^2}{4}}\,.$$

It follows that for $r \geq \max\{4n/c, 1\}$,

$$\Lambda_{k,\sigma}(r) \leq (c\sqrt{\pi})^{-n} \sqrt{C_{\sigma,n}} \left(\frac{4}{c^2}\right)^{n/4} e^{\frac{c^2 r^2}{2}} \leq c^{-\frac{3}{2}n} \sqrt{C_{\sigma,n}}\, e^{\frac{c^2 r^2}{2}}\,.$$

Hence for $r \geq c^{-\frac{3}{2}n}\sqrt{C_{\sigma,n}}e^{\max\{\frac{c^2}{2},8n^2\}}$,

$$\Lambda_{k,\sigma}^{-1}(r) \geq \frac{\sqrt{2}}{c}\sqrt{\ln r + \frac{3}{2}n\ln c - \frac{1}{2}\ln C_{\sigma,n}}\,.$$

Let

$$R \geq 8^n\|a\|_{\sigma,2}\sqrt{C_{\sigma,n}}\max\left\{c^{-\frac{3}{2}n}e^{\max\{\frac{c^2}{2},8n^2\}}, c^{-n}n^{n/4}e^{\frac{c^2n\pi^2}{4}}\right\}.$$

Under this restriction, (6.6) holds, and $N_R \geq [N_R'] \geq 1$, where

$$N_R' := \frac{\sqrt{2}}{\sqrt{n}\pi c}\sqrt{\ln R + \frac{3}{2}n\ln c - \ln(8^n\|a\|_{\sigma,2}) - \frac{1}{2}\ln C_{\sigma,n}}\,.$$

To get the desired constant, we cite the estimate for $\varepsilon_k(N)$ from [14, Example 3]:

$$\varepsilon_k(N) \leq 2\sqrt{e}\left(\frac{1}{16n}\right)^{(N-2)/2} + \frac{4}{c\sqrt{\pi}}2^{-n(N-2)}, \qquad \text{when } N \geq \frac{80n\ln 2}{c^2} + 2\,.$$

Then restrict R further such that $N_R' \geq \frac{80n\ln 2}{c^2} + 3$. We have

$$\varepsilon_k(N_R) \leq \left(32n\sqrt{e} + \frac{4^{n+1}}{c\sqrt{\pi}}\right)\left(\max\left\{\frac{1}{4\sqrt{n}}, \frac{1}{2^n}\right\}\right)^{N_R} \leq \left(32n\sqrt{e} + \frac{4^{n+1}}{c\sqrt{\pi}}\right)p_n^{-1}p_n^{N_R'}\,.$$

Now under the restriction

$$R \geq \max\left\{\left(8^n\|a\|_{\sigma,2}\sqrt{C_{\sigma,n}}c^{-\frac{3}{2}n}\right)^2, e^{n\pi^2c^2(80n\ln 2/c^2+3)^2}\right\},$$

we know that (6.6) holds and

$$N_R' \geq \frac{1}{\sqrt{n}\pi c}\sqrt{\ln R} \geq \frac{80n\ln 2}{c^2} + 3\,.$$

Finally, we choose

$$R' = \sqrt{\frac{-2\ln p_n}{c^3\sqrt{n}\pi}}\left(\ln R\right)^{1/4} > 0\,.$$

It can be easily checked that $R' \leq \sqrt{\ln R}/(\sqrt{n}\pi c) \leq \pi N_R$.

We can see that all the above restrictions on R hold when (6.12) is valid. Therefore, with (6.12), (6.6) tells us that

$$I(a,R) \leq \|a\|_{\sigma,2}\left\{\left(\frac{-2\ln p_n}{c^3\sqrt{n}\pi}\sqrt{\ln R}\right)^{-\sigma/2} + 2^n\left(\left(32n\sqrt{e} + \frac{4^{n+1}}{c\sqrt{\pi}}\right)p_n^{-1}(c\sqrt{\pi})^{-n}p_n^{\frac{\sqrt{\ln R}}{2\sqrt{n}\pi c}}\right)^{1/2}\right\}.$$

The choice $R \geq \left(\frac{2\sigma\sqrt{n}\pi c}{-\ln p_n}\right)^4$ implies that

$$p_n^{\frac{\sqrt{\ln R}}{2\sqrt{n}\pi c}} \leq (\ln R)^{-\sigma/4}\,.$$

Hence

$$I(a,R) \le \|a\|_{\sigma,2}\left\{\left(\frac{-2\ln p_n}{c^3\sqrt{n}\pi}\right)^{-\sigma/2} + 2^n\left(32n\sqrt{e}+\frac{4^{n+1}}{c\sqrt{\pi}}\right)p_n^{-1}(c\sqrt{\pi})^{-n}\right\}\left(\ln R\right)^{-\frac{\sigma}{4}}.$$

This yields the desired approximation error for the Gaussian kernels. □

By taking $c=\sqrt{2}$, the second statement of Proposition 1.1 follows from the estimates in Example 6.2.

The next example deals with multiquadric kernels which are C^∞ kernels.

Example 6.2. Let $c>0$, $\alpha>n$, and

$$k(x) = (c^2+|x|^2)^{-\alpha/2}, \qquad x\in\mathbb{R}^n.$$

If $\sigma>0$ and $a\in H^\sigma(\mathbb{R}^n)$, then

$$I(a,R) = \inf_{\|b\|_{L^2([0,1]^n)}\le R}\{\|a-L_Kb\|_{L^2([0,1]^n)}\} = O\left(\left(\frac{1}{\ln R}\right)^\sigma\right).$$

Proof. For any $\varepsilon>0$, there are positive constants C_1, C_2 such that

$$C_1e^{-(c+\varepsilon)|\xi|} \le \hat{k}(\xi) \le C_2e^{-c|\xi|} \qquad \forall\xi\in\mathbb{R}^n.$$

Then

$$\lambda_{k,\sigma}(r) \le \frac{1}{C_1}e^{(c+\varepsilon)r}$$

and for $r\ge 1$,

$$\sqrt{C_{\sigma,n}}\frac{1}{C_2}e^{cr} \le \Lambda_{k,\sigma}(r) \le \sqrt{C_{\sigma,n}}r^{n/2}\frac{1}{C_1}e^{(c+\varepsilon)r} \le C_1'e^{(c+2\varepsilon)r}.$$

Hence for sufficiently large R,

$$(\ln R+\ln C_2-\ln C_{\sigma,n}/2)/c \le \Lambda_{k,\sigma}^{-1}(R) \le (\ln R-\ln C_1')/(c+2\varepsilon).$$

It follows that for sufficiently large R,

$$N_R \ge \left[\frac{1}{\sqrt{n}\pi(c+2\varepsilon)}\{\ln R-\ln(8^n\|a\|_{\sigma,2}C_1')\}\right] \ge \frac{1}{2\sqrt{n}\pi(c+2\varepsilon)}\ln R.$$

The estimate in [7] provides a bound for ε_k: with some fixed constants $\delta>0$ and $C_2'>0$,

$$\varepsilon_k(N) \le C_2'e^{-\delta N} \qquad \forall N\in\mathbb{N}.$$

Combining the above estimates, we know from Theorem 6.1, that

$$I(a,R) \le \|a\|_{\sigma,2}\inf_{0<R'\le\pi N_R}\left\{\left(\frac{1}{R'}\right)^\sigma + 2^n\left(C_2'e^{-\delta N_R}\frac{1}{C_1}e^{(c+\varepsilon)R'}\right)^{1/2}\right\}.$$

Choosing $R'=d\ln R$ for sufficiently small d yields

$$I(a,R) \le \|a\|_{\sigma,2}\left\{(d\ln R)^{-\sigma}+2^n\frac{C_2'}{C_1}e^{\{(c+\varepsilon)d-\delta/(2\sqrt{n}\pi(c+2\varepsilon))\}\ln R}\right\} = O\left(\left(\frac{1}{\ln R}\right)^\sigma\right).$$

This is the expected estimate for the approximation error. □

Acknowledgements

The first author is supported by CERG grant No. 9040457 and City University grant No. 8780043. The second author is supported by CERG grant No. 9040536 and City University grant No. 7001029.

References

[1] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337–404.
[2] J. Bergh and J. Löfström, *Interpolation Spaces, an Introduction*, Springer-Verlag, 1976.
[3] F. Cucker and S. Smale, On the mathematical foundations of learning, to appear in *Bull. Amer. Math. Soc.*.
[4] R. DeVore and G. Lorentz, *Constructive Approximation*, Springer-Verlag, 1993.
[5] T. Evgeniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.
[6] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Information and Computation* **100** (1992), 78–150.
[7] W. R. Madych and S. A. Nelson, Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation, *J. Approx. Theory* **70** (1992), 94–114.
[8] C. A. Micchelli and A. Pinkus, Variational problems arising from balancing several error criteria, *Rend. Mat. Appl.* **14**(7) (1994), 37–86.
[9] P. Niyogi, *The Informational Complexity of Learning*, Kluwer, 1998.
[10] J. Peetre, *New thoughts on Besov spaces*, Duke Univ. Math. Series, Durham Univ., 1976.
[11] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, 1970.
[12] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
[13] Z. Wu and R. Schaback, Local error estimates for radial basis function interpolation of scattered data, *IMA J. Numer. Anal.* **13** (1993), 13–27.
[14] D. X. Zhou, The covering number in learning theory, preprint, 2001.