APPROXIMATING FUNCTIONS IN REPRODUCING KERNEL HILBERT SPACES VIA STATISTICAL LEARNING THEORY

MARK A. KON LOUISE A. RAPHAEL

Abstract: New non-asymptotic uniform error bounds for approximating functions in reproducing kernel Hilbert spaces are given using F. Girosi's approach to derive approximation theoretic results from statistical learning theory.

The authors congratulate Professor Charles Chui on the occasion of his sixty fifth birthday.

1. Introduction

Reproducing kernel Hilbert spaces (RKHS) [1] are the settings of choice of applied probabilists and statisticians (Whaba [18]) and kernel machine/statistical learning researchers, (e.g., Cuker & Smale [3], Girosi [5], Poggio & Smale [12], Schölkopf and Smola [13], Shawe-Taylor and Chistianini [14], Vapnik [17], and Zhou [21]). In kernel machine learning for example, one often uses RKHS or their r-balls as hypothesis spaces.

F. Girosi was the first to apply statistical learning theory [SLT] results to obtain approximation theoretic bounds [4]. In this paper we give a RKHS setting for Girosi's results. The bounds depend only on the complexity of the class of reproducing kernels and the number of data points. For functions in RKHS our results yield sup norm, probabilistic, non-asymptotic bounds. Our methods yield L^∞ errors which differ from the L^1 and L^2 norm errors of, e.g., regularization [9] and kernel density estimation [6, 19]. The essence of these methods has to do with methodologies for approximating integrals by sums. This topic itself has a wide literature; for some recent results see [11].

For a general reproducing kernel $K(\mathbf{x}, \mathbf{t})$ ($\mathbf{x}$, $\mathbf{t} \in \mathbb{R}^d$) functions in the RKHS $\mathcal{H}_K$ can be approximated in the $\mathcal{H}_K$ norm by linear combinations $\sum_{i=1}^n c_i K(\mathbf{x}, \mathbf{t}_i)$. We show here that the error bounds can be made uniform in the L^∞ norm provided the kernel is modified by possible weight functions, and the L^2 operator K is replaced by $K^{1/2}$, its operator square root K (which in some cases equals K). Though our results are non-constructive, the existence of such an approximation requires a numerical determination of $\{\mathbf{t}_i\}_{i=1}^n$ using optimization techniques, a problem which remains to be solved.

In Section 2 we review some basic SLT concepts and include the seminal VC bound theorem, needed to prove our main result. In Section 3, we give a modified version of Girosi's result. Section 4 contains our main results involving uniform non-asymptotic bounds for real and complex RKHS. Letting $\langle \cdot, \cdot \rangle$ denote inner product, our results distinguish between two sub-cases:

Date: September 30, 2005.

1991 *Mathematics Subject Classification.* Primary 68T05, 41A65.

Key words and phrases. Reproducing kernel Hilbert space; statistical learning theory; VC dimension; wavelets.

- $\langle \cdot, \cdot \rangle_{RKHS} = \langle \cdot, \cdot \rangle_{L^2}$; and
- $\langle \cdot, \cdot \rangle_{RKHS} \neq \langle \cdot, \cdot \rangle_{L^2}$.

In the first case the reproducing kernel operator K is an L^2 projection onto its closed subspace H_K, e.g., a wavelet space, spline space, or space of bandlimited functions, and $K = K^{1/2}$. In the second case, (for example in the case of a Gaussian kernel), K maps onto a non-closed subspace of L^2, which we will assume to be dense in L^2.

2. SLT Background and Definitions

The goal in a standard SLT paradigm is to find an unknown function $f : X \to Y$ from random samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $X \subset \mathbb{R}^d, \quad Y \subset \mathbb{R}$. As there are noise and other uncertainties, we do not expect $f(\mathbf{x}_i) = y_i$, but rather that they are approximately equal. Thus at best we can only find an approximation to the predictor function f. In SLT this approximation is usually realized by minimizing some loss function which measures the error between y_i and a predicted value $f(\mathbf{x}_i)$.

To be precise, let $\mathbf{X} \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}^1$. Assume $\mathbf{X} \times Y$ is sampled n times under an unknown probability distribution $P(\mathbf{x}, y)$, and denote the data set by $\{(\mathbf{x}_i, y_i) \in \mathbf{X} \times Y\}_{i=1}^n$. Given a hypothesis space H of possible functions relating $\mathbf{x}$ and y, the problem is to find a predictor $f : \mathbf{X} \to Y$ in H such that when $\mathbf{x} \in \mathbf{X}$ is given, $f(\mathbf{x})$ predicts a value for y optimally.

Following Girosi [4], we use Vapnik's probabilistic bound approach involving the VC dimension [16, 10]. For $f \in H$ and $\mathbf{z} = (\mathbf{x}, y)$, let the *loss function*

$$V(y, f(\mathbf{x})) = V(f, \mathbf{z})$$

measure the error between y and its prediction $f(\mathbf{x})$. Two examples are $V(f, \mathbf{z}) = |y - f(\mathbf{x})|^p$, $1 \leq p < \infty$ and the $\{0,1\}$-valued function $V(f, \mathbf{z}) = 1 - \chi_{[-1,1]}(y - f(\mathbf{x}))$.

For $f \in H$, the *expected risk* $R[f]$ is defined as the average of the loss function V, namely

$$\text{(1)} \qquad \text{Expected Risk} \;=\; R[f] = \int V(f, \mathbf{z}) P(\mathbf{z}) d\mathbf{z},$$

where the probability measure $P(\mathbf{z}) = P(\mathbf{x}, y)$ is unknown. Thus the *estimator function*

$$\text{(2)} \qquad f^* = \arg \left\{ \min_{f \in H} R[f] \right\}$$

cannot be found directly.

Instead the data set $\{(\mathbf{x}_i, y_i) \in \ \mathbf{X} \times Y\}_{i=1}^n$ is used to find an information-based approximation of the expected risk, called the *empirical risk.* For $f \in H_k \subset H$ define

$$\text{(3)} \qquad \text{Empirical Risk} = R_{emp}[f; n] = \frac{1}{n} \sum_{i=1}^{n} V(f, \mathbf{z}_i).$$

A difficulty in finding a minimizer of the expected risk using the empirical risk arises from the possible existence of many minimizing functions. Moreover, it is possible to pick an f which often has small empirical risk, but large expected risk. An SLT approach to resolving this is to find uniform probabilistic bounds on the difference between the expected and empirical risks.

In applications, H is often too large and so the empirical risk is successively minimized on a nested sequence of increasing subspaces $H_0 \subset H_1 \subset \cdots \subset H_k \subset$

$\cdots \subset H$, where increasing k reflects increasing "capacity" of H_k. Standard examples of spaces H_k include splines with k nodes, and degree k trigonometric polynomials in d variables. The VC bound theorem (below) is stated in terms of H_k.

Vapnik's *empirical risk minimization principle* (*ERMP*) is an approach which gives an approximation in H_k to the minimizer f^* in (2) by first finding a sequence of "locally" minimizing approximations $f_{k,n} \in H_k$ (with n the number of data points) defined by

$$f_{k,n} = \arg\left\{ \min_{f \epsilon H_k} R_{emp}[f;n] \right\}. \tag{4}$$

As $n \to \infty$, we hope that $f_{k,n} \in H_k$ converges to

$$f_k = \arg\left\{ \min_{f \epsilon H_k} R[f] \right\}. \tag{5}$$

and

$$\lim_{n\to\infty} R_{\mathrm{emp}}[f_{k,n};n] = \lim_{n\to\infty} R[f_{k,n}] = R[f_k]. \tag{6}$$

Seminal work of Vapnik and Chervonenkis ([17], Theorem 2.1) shows that, for bounded $R[f]$, (6) is satisfied in H_k when the following uniform convergence in probability holds for all $\epsilon > 0$:

$$\lim_{n\to\infty} P\left\{ \sup_{f\in H_k} \Big(R[f] - R_{emp}[f;n] \Big) > \epsilon \right\} = 0. \tag{7}$$

This in turn leads to uniform non-asymptotic VC bounds which are given in terms of the VC dimension of the class of empirical risks. We now give a definition of VC-dimension.

Definition 1. *The VC dimension of a family H of functions on a space X is the maximum number h of points $\{t_i\}_{i=1}^h$ which can be separated into two classes in all possible ways, using classes of the form:*

- *$f(t_i) - \alpha \geq 0$, and*
- *$f(t_i) - \alpha \leq 0$*

as $f \in \mathcal{H}$ and the parameter $\alpha \in \mathbb{R}$ vary.

Given a kernel $K(\mathbf{x}, \mathbf{t})$ the VC dimension of a family of functions $K(\mathbf{x}, \mathbf{t})$ (in the variable $\mathbf{t}$ and parameter $\mathbf{x}$) is the above-defined VC dimension of the family $\mathcal{H} = \{K(\mathbf{x}, \mathbf{t})\}_{\mathbf{x}\in\mathbb{R}^d}$, where $\mathbf{x}$ parameterizes the family H.

Examples of function classes of VC-dimension $d + 1$ include:

- characteristic functions of half-spaces on R^d
- characteristic functions of d-dimensional balls on R^d
- Gaussian kernels on R^d [4].

The following well-known theorem of Vapnik and Chervonenkis, which gives probabilistic estimates of integrals by finite sums, is needed in the proofs of [4] as well as here.

Theorem 1. *(VC Bound Theorem [17]). Let $V(f, \mathbf{z})$, $\mathbf{z} = (\mathbf{x}, y)$, satisfy $A \leq V(f, \mathbf{z}) \leq B$ for f in a hypothesis space of functions H_k. Let h be the VC dimension of $\{V(f, \mathbf{z})\}_{f\in H_k}$ and n be the number of data points $\mathbf{z}_i$ (chosen with respect to the* *probability distribution $P(\mathbf{z}) = P(\mathbf{x}, y)$). Then for any $0 < \eta < 1$, the following inequality holds simultaneously for all $f \in H_k$, with probability at least $1 - \eta$:*

$$\left| R[f] - R_{emp}[f; n] \right| \leq (B - A) \sqrt{\frac{h \ln \frac{2en}{h} - \ln \frac{\eta}{4}}{n}}. \tag{8}$$

Note above e denotes the base of the natural logarithm.

3. Modifications of Girosi's Results

The so-called curse of dimensionality occurs when a problem's complexity grows exponentially with dimension d. Typically, for a function of smoothness s in dimension d, the number of parameters n needed to achieve an approximation error smaller than some positive ϵ is

$$n \propto \left(\frac{1}{\epsilon} \right)^{d/s}.$$

Letting s change with d improves the error, which is $O(n^{-s/d})$.

We follow Girosi [4] in reinterpreting SLT notions as approximation theory (AT) concepts as follows:

SLT Notation	**AT Notation**
R [risk function]	f
f	$\mathbf{x}$
$\mathbf{z}$	$\mathbf{t}$
V [loss function]	J [kernel]
P [probability distribution]	λ [measure]
H_k [approximation space]	$\mathbb{R}^d$

Under these replacements the expected risk of SLT

$$R[f] = \int V(y, f(\mathbf{x})) P(\mathbf{x}, y) d\mathbf{x} dy = \int V(f, \mathbf{z}) P(\mathbf{z}) d\mathbf{z}$$

becomes in AT

$$f(\mathbf{x}) = \int J(\mathbf{x}, \mathbf{t}) \lambda(\mathbf{t}) d\mathbf{t}, \quad \mathbf{x}, \mathbf{t} \in \mathbf{R}^d. \tag{9}$$

Then the empirical risk in AT has the form

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} J(\mathbf{x}, \mathbf{t}_i). \tag{10}$$

Girosi [4] used the VC bound theorem of Vapnik and Chervonenkis to estimate integrals of the form (9) using sums like (10) when $\lambda(\mathbf{t}) \in L^1(\mathbb{R}^d)$. We will give a modification of this result applied to functions of the form (9) for any bounded kernel.

Proposition 2 (Girosi). *Let f be represented as an integral in the form (9), with $\lambda \in L^1(\mathbb{R}^d)$. If the kernel J satisfies $A \leq J(\mathbf{x}, \mathbf{t}) \leq B$, $\mathbf{x}, \mathbf{t} \in \mathbf{R}^d$, the following probabilistic error bound holds with probability $1-\eta$ for a sample of n points $\{\mathbf{t}_i\}_{i=1}^n$ taken with respect to the probability density $|\lambda(\mathbf{x})| d\mathbf{x}$ (normalized to unit L^1 norm if necessary):*

$$(11)\qquad \left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\lambda(\mathbf{t}_i))J(\mathbf{x},\mathbf{t}_i)\|\lambda\|_{L^1} \right\|_{L^\infty} \leq 4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} - \ln\frac{\eta}{4}}{n}}.$$

where $\tau = B - A$.

The term 4τ follows from the fact that if $|J(\mathbf{x},\mathbf{t})| \leq \tau$, then $B - A = 2\tau$. The additional factor of 2 in the term 4τ is a consequence of writing the coefficients $sgn(\lambda(\mathbf{t}_i))c_i = c_i^+ - c_i^-$, the sum of their negative and positive parts (i.e., $c^+ = \sup(c,0)$ and $c^- = \sup(-c,0)$).

Proposition 2 leads to the following corollary.

Corollary 3. *Under the assumptions of Proposition 2, for every* $\epsilon > 0$ *there exists a sample* $\{\mathbf{t}_i\}_{i=1}^n \in \mathbb{R}^d$ *such that*

$$(12)\qquad \left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\lambda(\mathbf{t}_i))J(\mathbf{x},\mathbf{t}_i)\|\lambda\|_{L^1} \right\|_{\infty} \leq 4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} + \ln 4}{n}} + \epsilon.$$

Proof. Letting $\eta \uparrow 1$, we see that the right hand side of (11) approaches

$$4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} + \ln 4}{n}}$$

from above. That is for any $\epsilon > 0$ we can find an $\eta < 1$ such that

$$4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} - \ln\frac{\eta}{4}}{n}} \leq 4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} + \ln 4}{n}} + \epsilon.$$

Thus for any $\epsilon > 0$ there exists a sample $T = \{\mathbf{t}_i\}_{i=1}^n$ such that the result holds. □

Remarks:

- In SLT for learning to take place the number of data points n must be greater than the VC-dimension h.
- For kernels which are uniformly continuous in $\mathbf{t}$ and the parameter $\mathbf{x}$ (which occurs in many RKHS), equation (12) holds for $\epsilon = 0$. This can be shown through the existence of a limiting data set $\mathbf{t}_1, \ldots, \mathbf{t}_k$ for which the limiting value $\epsilon = 0$ holds in (12). For a proof for generalized Sobolev spaces $\mathcal{L}_s^p(\mathbb{R}^d)$ see ([8], Corollary 3).

4. Main Theorem - Girosi's Result for RKHS

A real RKHS $\mathcal{H}_K$ on a space X can be defined [18] to be a Hilbert space of real valued functions on X with the property that, for each $\mathbf{x} \in \mathbb{R}^d$, the pointwise evaluation functional $K_\mathbf{x}$ which associates f with $f(\mathbf{x})$, $K_\mathbf{x} \to f(\mathbf{x})$ is a bounded linear functional.

Letting T be the transpose, we recall that a symmetric $n \times n$ matrix A is:

- *positive semi-definite* if $\mathbf{x}^T A\mathbf{x} \geq 0$; and
- *positive definite* if $\mathbf{x}^T A\mathbf{x} > 0$.

for all $x \in \mathbb{R}^n$. Our somewhat more restrictive definition follows [3, 21].

Definition 2. *Let $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ be continuous, symmetric and positive semi-definite, meaning that for any finite set of distinct points $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbf{X}$, the matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ is positive semi-definite. It is called positive definite if the matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ is positive definite. The reproducing kernel Hilbert space [RKHS] $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the algebraic span of the set of functions $\{K_\mathbf{x} := K(\mathbf{x}, \cdot) | \mathbf{x} \in \mathbf{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ satisfying*

$$\| \sum_{i=1}^n c_i K_{\mathbf{x}_i} \|_{\mathcal{H}_K}^2 = \Big\langle \sum_{i=1}^n c_i K_{\mathbf{x}_i}, \sum_{i=1}^n c_i K_{\mathbf{x}_i} \Big\rangle_{\mathcal{H}_K} = \sum_{i,j=1}^n c_i K(\mathbf{x}_i, \mathbf{x}_j) c_j.$$

The reproducing kernel property is given by

$$\Big\langle K_\mathbf{x}, f \Big\rangle_{\mathcal{H}_K} = f(\mathbf{x}) \quad \forall\ \mathbf{x} \in \mathbf{X}, \quad f \in \mathcal{H}_K. \tag{13}$$

We note some well known RKHS facts:

- Any positive semi-definite $K(\mathbf{x}, \mathbf{t})$ can be used to construct a RKHS $\mathcal{H}_K$ associated with it.
- Any convolution kernel $K(\mathbf{x}, \mathbf{t}) = K(\mathbf{x} - \mathbf{t})$ with a non-negative Fourier transform is a reproducing kernel and is associated with a RKHS.

In addition to the standard situation in which H_K is a dense subset of L^2, we will also consider another case of interest in which H_K is a proper closed subspace of L^2, and the operator K corresponding to the kernel $K(\cdot, \cdot)$ is the orthogonal projection from L^2 onto H_K. Here we will loosen the restriction on continuity of $K(\cdot, \cdot)$ in order to admit some cases of interest. We will use the fact that in both cases an $f \in H_K$ can be expressed as a Lebesgue integral against the kernel.

We assume for our first main result that $K(\mathbf{x}, \mathbf{t})$ is positive definite and $\mathcal{H}_K$ is dense in $L^2(\mathbb{R}^d)$ which occurs in most applications (see examples for Theorem 4). Note that $K(\cdot, \mathbf{t})$ is in L^2 for all $\mathbf{t}$.

Define $K : L^2 \to L^2$ to be the self-adjoint closure of the operator

$$Kf = \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mathbf{t}, \tag{14}$$

initially defined on the space of compactly supported infinitely differentiable functions. We assume the self-adjoint operator $K^{1/2}$ has a kernel $K^{1/2}(\mathbf{x}, \mathbf{t})$ with $K^{1/2}(\cdot, \mathbf{t}) \in L^2$ for all $\mathbf{t}$. Consider the self-adjoint operator $K^{-1/2}$ and note that the function $K(\cdot, \mathbf{t}) \in \mathrm{Dom}_{L^2}(K^{-1/2})$. To verify this we show $K(\cdot, \mathbf{t}) \in \mathrm{Ran}_{L^2}(K^{1/2})$ which follows from

$$K^{1/2}[K^{1/2}(\cdot, \mathbf{t})] = \int K^{1/2}(\cdot, \mathbf{y}) K^{1/2}(\mathbf{y}, \mathbf{t}) d\mathbf{y} = K(\cdot, \mathbf{t}) \tag{15}$$

(the distinction between the operator $K^{1/2}$ and its kernel $K^{1/2}(\mathbf{x}, \mathbf{t})$ above is clear). For an alternative proof see [3].

Define the dense subset $\widetilde{\mathcal{H}}_K \subset \mathcal{H}_K$ by

$$\widetilde{\mathcal{H}}_K = \left\{ \sum_{i=1}^n c_i K(\mathbf{x}, \mathbf{t}_i) : n \in \mathbb{N},\, c_i \in \mathbb{R} \right\}.$$

Now we show for $f, g \in \mathcal{H}_K$,

$$\langle f, g \rangle_{\mathcal{H}_K} = \langle K^{-1/2} f, K^{-1/2} g \rangle_{L^2(\mathbb{R}^d)}. \tag{16}$$

Note that if $f = \sum_{i=1}^n c_i K(\mathbf{x}, \mathbf{t}_i)$, $g = \sum_{j=1}^n d_j K(\mathbf{x}, \mathbf{t}_j) \in \widetilde{\mathcal{H}}_K$, then

$$\langle f, g\rangle_{\mathcal{H}_K} = \sum_{i,j=1}^n c_i d_j \langle K(\mathbf{x}, \mathbf{t}_i), K(\mathbf{x}, \mathbf{t}_j)\rangle_{\mathcal{H}_K} = \sum_{i,j=1}^n c_i d_j K(\mathbf{t}_i, \mathbf{t}_j).$$

In addition

$$\langle K^{-1/2} f, K^{-1/2} g\rangle_{L^2(\mathbb{R}^d)}$$

$$= \sum_{i,j=1}^n c_i d_j \langle K^{-1/2} K(\mathbf{x}, \mathbf{t}_i), K^{-1/2} K(\mathbf{x}, \mathbf{t}_j)\rangle_{L^2(\mathbb{R}^d)} = \sum_{i,j=1}^n c_i d_j K(\mathbf{t}_i, \mathbf{t}_j),$$

where

$$K^{-1/2} K(\cdot, \mathbf{t}) = K^{1/2}(\cdot, \mathbf{t})$$

(note we have used (15)).

Since $\widetilde{\mathcal{H}}_K$ is dense in $\mathcal{H}_K$, which is dense in $L^2(\mathbb{R}^d)$, we conclude $\widetilde{\mathcal{H}}_K$ is also dense in $L^2(\mathbb{R}^d)$. Since $K^{-1/2}$ is closed, we form the completion of the space $\widetilde{\mathcal{H}}_K$ on both sides of (16) (with respect to each inner product), concluding that $\mathcal{H}_K = \mathrm{Dom}_{L^2} K^{-1/2}$ (for a more detailed argument, see [3, Section III.3].

Therefore for $f \in \mathcal{H}_K$, we have $K^{-1/2} f(\mathbf{x}) \in L^2(\mathbb{R}^d)$,

$$f(\mathbf{x}) = \langle (K(\mathbf{x}, \cdot), f(\cdot)\rangle_{\mathcal{H}_K} = \langle (K^{-1/2}) K(\mathbf{x}, \cdot), (K^{-1/2}) f(\cdot)\rangle_{L^2}$$

$$= \int_{\mathbb{R}^d} K^{1/2}(\mathbf{x}, \mathbf{t}) (K^{-1/2} f)(\mathbf{t}) d\mathbf{t}, \tag{17}$$

where, by the above discussion, the integral converges in $L^2(\mathbb{R}^d)$.

Our main results give non-asymptotic uniform error bounds for approximating functions in a RKHS. Such a bound is a function of the number of data points used and the VC dimension, i.e., the richness of the space. What is done here is a generalization of Girosi's results for RKHS, which are inner product spaces; we note that Girosi's results for generalized L^1 Sobolev spaces do not apply to RKHS. However, the usual assumption that the reproducing kernel K is positive definite and continuous excludes some interesting cases. The latter are considered in Proposition 5 and Corollary 6.

We recall that a weighted L^∞ norm is defined by

$$\|f\|_{L^\infty, a(\mathbf{x})} = \text{ess sup}_{\mathbf{x}} |f(\mathbf{x}) a(\mathbf{x})|.$$

Theorem 4. *Given an $L^2(\mathbb{R}^d)-$dense RKHS $\mathcal{H}_K$ with a positive definite reproducing kernel $K(\mathbf{x}, \mathbf{t})$, let $K^{1/2}(\mathbf{x}, \mathbf{t})$ be in $L^2(\mathbb{R}^d)$. Assume there exist positive functions $g(\mathbf{t})$ and $k(\mathbf{x})$ bounded away from 0 with $g \in L^2(\mathbb{R}^d)$ such that*

$$\text{ess sup}_{\mathbf{x}, \mathbf{t}} \left| \frac{K^{1/2}(\mathbf{x}, \mathbf{t})}{g(\mathbf{t}) k(\mathbf{x})} \right| \leq \tau. \tag{18}$$

Let h be the VC dimension of $\dfrac{K^{1/2}(\mathbf{x}, \mathbf{t})}{g(\mathbf{t}) k(\mathbf{x})}$ in the parameter $\mathbf{x}$ and the variable $\mathbf{t}$. Then for every $f \in \mathcal{H}_K$ and every $\epsilon > 0$ there exist $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\} \subset \mathbb{R}^d$, and n coefficients $c_i = \mathrm{sgn}(K^{-1/2} f)(\mathbf{t}_i)) = \pm 1$, such that the weighted L^∞ norm

$$\left\| f(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n c_i \|(K^{-1/2} f) g\|_{L^1} \frac{K^{1/2}(\mathbf{x}, \mathbf{t}_i)}{g(\mathbf{t}_i)} \right\|_{L^\infty, 1/k(\mathbf{x})}$$

$$\leq 4\tau\|f\|_{\mathcal{H}_K}\|g\|_{L^2}\sqrt{\frac{h\ln\frac{2en}{h}+\ln 4}{n}}+\epsilon. \tag{19}$$

Proof. For $f \in \mathcal{H}_K$ and positive $g \in L^2(\mathbb{R}^d)$, we have $(K^{-1/2}f)(\mathbf{x})g(\mathbf{t}) \in L^1$ by Hölder's inequality. Thus recalling (17)

$$\frac{f(\mathbf{x})}{k(\mathbf{x})} = \int \frac{K^{1/2}(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})}(K^{-1/2}f)(\mathbf{t})g(\mathbf{t})d\mathbf{t}.$$

Referring to Proposition 2 and equation (12) define

$$J(\mathbf{x},\mathbf{t}) = \frac{K^{1/2}(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \text{ and } \lambda(\mathbf{t}) = (K^{-1/2}f)(\mathbf{t})g(\mathbf{t}).$$

It now follows by Corollary 3 that for every $\epsilon > 0$ there exist $\{\mathbf{t}_i\}_{i=1}^n$ such that

$$\left\|\frac{f(\mathbf{x})}{k(\mathbf{x})} - \frac{1}{n}\sum_{i=1}^n \frac{K^{1/2}(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)k(\mathbf{x})}\ \|(K^{-1/2}f)(\mathbf{t})g(\mathbf{t})\|_{L^1}\mathrm{sgn}((K^{-1/2}f)(\mathbf{t}_i)g(\mathbf{t}_i))\right\|_{L^\infty}$$

$$\leq \|K^{-1/2}f\|_{L^2}\|g\|_{L^2}\left\|\int \frac{K^{1/2}(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})}\frac{(K^{-1/2}f)(\mathbf{t})g(\mathbf{t})}{\|(K^{-1/2}f)(\mathbf{t})g(\mathbf{t})\|_{L^1}}d\mathbf{t}\right.$$

$$\left. - \frac{1}{n}\sum_{i=1}^n \frac{K^{1/2}(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)k(\mathbf{x})}\mathrm{sgn}((K^{-1/2}f)(\mathbf{t}_i)g(\mathbf{t}_i))\right\|_{L^\infty}$$

$$\leq 4\tau\|f\|_{\mathcal{H}_K}\|g\|_{L^2}\sqrt{\frac{h\ln\frac{2en}{h}+\ln 4}{n}}+\epsilon$$

as $\|f\|_{\mathcal{H}_K} = \|K^{-1/2}f\|_{L^2}$. □

Examples of Theorem 4 for real positive definite reproducing kernels include:

- The Gaussian kernel $e^{\frac{-\|\mathbf{x}-\mathbf{t}\|^2}{2\sigma^2}}$ on $L^2(\mathbb{R}^d)$.
- The Bessel potential kernel associated with a generalized Sobolev space $\mathcal{L}_s^2$, given by

$$K^{1/2}(\mathbf{x},\mathbf{t}) = \frac{(4\pi)^{-s/2}}{\Gamma(\frac{s}{2})}\int_0^\infty \exp\left(-\frac{\pi}{\sigma}|\mathbf{x}|^2\right)\exp\left(-\frac{\sigma}{4\pi}\right)\sigma^{(s-d-2)/2}d\sigma \tag{20}$$

  has Fourier transform (FT) $(1+|\xi|^2)^{-s/2}$ [16]. To find the reproducing kernel $K(\mathbf{x},\mathbf{t})$ square the Fourier transform giving $(1+|\xi|^2)^{-s}$, and take the inverse FT yielding $K(\mathbf{x},\mathbf{t})$, obtainable by substituting $2s$ for s in (20).
- Following [18], denote by W_m^0 the RKHS of functions on $[0,1]$ which have m absolutely continuous derivatives in L^2, with $f^{(\nu)}(0) = 0,\ \nu = 0,1,...,m-1$ and with square norm $\|f\|^2 = \int_0^1 (f^{(m)}(t))^2dt$. Define the Green's function $G_m(x,t) = \frac{(x-t)_+^{m-1}}{(m-1)!}$ for the problem $D^m f = g,\ f \in W_m^0$. For each $f \in \mathcal{W}_m^0$, $f(t) = \int_0^1 G_m(t,u)f^{(m)}(u)du$. The reproducing kernel is

$$R(t,s) = \int_0^1 G_m(t,u)G_m(s,u)du = R_t(s) \tag{21}$$

with

$$\langle f, R_t\rangle_{\mathcal{H}_K} = \int_0^1 G_m(t,u) f^{(m)}(u)du = f(t). \tag{22}$$

For these reproducing kernel spaces with $L^2 \subset L^1$, we can choose $g(\mathbf{t}) = k(\mathbf{x}) = 1$.

For the purpose of including spaces of band-limited functions, wavelet and spline spaces, we now consider an important situation in which K is not positive definite, (nor in some cases continuous). We assume that $K(\mathbf{x}, \mathbf{t})$ is bounded, positive semi-definite, and that $\mathcal{H}_K$ is a closed subspace of L^2 with inner product inherited from L^2, so that $\langle f, g\rangle_{\mathcal{H}_K} = \langle f, g\rangle_{L^2}$. In this case, we have for $f \in \mathcal{H}_K$,

$$f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_K} = \langle f, g\rangle_{L^2} = \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d(\mathbf{t}). \tag{23}$$

We can show as in Theorem 4 the following.

Proposition 5. *Assume K is bounded and positive semi-definite, and that $\mathcal{H}_K$ inherits the L^2 inner product. Assume that there exist positive functions $g(\mathbf{t})$ and $k(\mathbf{x})$, bounded away from 0, with $g \in L^2(\mathbb{R}^d)$ such that*

$$\text{ess sup}_{\mathbf{x},\mathbf{t}} \left| \frac{K(\mathbf{x}, \mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \right| \le \tau. \tag{24}$$

Let h be the VC dimension of $\dfrac{K(\mathbf{x}, \mathbf{t})}{g(\mathbf{t})k(\mathbf{x})}$ in the parameter $\mathbf{x}$ and the variable $\mathbf{t}$. Then for every $f \in \mathcal{H}_K$ and every $\epsilon > 0$ there exist $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\} \subset \mathbb{R}^d$, and n coefficients $c_i = \operatorname{sgn}(f(\mathbf{t}_i)) = \pm 1$, such that the weighted L^∞ norm

$$\left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n c_i \|fg\|_{L^1} \frac{K(\mathbf{x}, \mathbf{t}_i)}{g(\mathbf{t}_i)} \right\|_{L^\infty, 1/k(\mathbf{x})}$$

$$\le 4\tau \|f\|_{L^2} \|g\|_{L^2} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}} + \epsilon. \tag{25}$$

Proof. Note that in this case the operator K is the orthogonal projection onto the closed subspace $\mathcal{H}_K \subset L^2$, so that $K^{1/2} = K$. Our conclusion follows in the same way as that of Theorem 4 by using equation (23), and replacing $K^{1/2}$ by K and $K^{-1/2}f$ by f in the argument of Theorem 4. □

Examples of Proposition 5 for real RKHS include the sinc kernel for bandlimited L^2 functions; and projection reproducing kernels on spaces of splines, frames, and wavelets ([2, 6, 19]). Here we consider the Haar case on $\mathbb{R}^d$.

Let

$$\phi_d(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in [0,1]^d \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

denote the Haar scaling function in $\mathbb{R}^d$ (i.e., a 0th order B-spline). Then at the scale $n = 0$, the corresponding family of wavelets consists of products of the form $\psi_d^\lambda(\mathbf{x}) = \prod_{i=1}^d \eta_i(x_i)$, where $\eta_i(x_i)$ is either

$$\phi(x_i) = \begin{cases} 1 & x_i \in [0,1] \\ 0 & \text{otherwise} \end{cases} \quad \text{or} \quad \psi(x_i) = \begin{cases} 1 & x_i \in [0,1/2] \\ -1 & x_i \in (1/2,1] \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

Thus the total number of basic wavelets is $2^d - 1$. We define the homogeneous Sobolev space for $s \in \mathbb{R}$ by

$$\mathcal{L}^2_{hom,s} = \left\{ f \in \mathcal{L}^2(R^d) \,\middle|\, \|f\|_{\mathcal{L}^2_{hom,s}} = \left\| |\omega|^s \widehat{f}(\omega) \right\|_{L^2} < \infty \right\},$$

where $\widehat{f}$ denotes the Fourier transform of f. Defining the RKHS $V_0 = \{\phi(2^m\mathbf{x} - \mathbf{k}\}_{\mathbf{k}\in\mathbb{Z}}$, we have a reproducing kernel

$$K(\mathbf{x},\mathbf{t}) = \sum_{\mathbf{k}\in\mathbb{Z}} \phi(2^m\mathbf{x} - \mathbf{k})\phi(2^m\mathbf{t} - \mathbf{k}),$$

for V_0; note in this case that $K(\mathbf{x},\mathbf{t}) = K^{1/2}(\mathbf{x},\mathbf{t})$. Assume now that $f(\mathbf{x}) \in \mathcal{L}^2_s(\mathbb{R}^d)$, $\frac{d}{2} < s < \frac{d}{2} + 1$. Then using Proposition 5 we can show that there exists a positive integer $l \le m$ and an approximation of the form $\sum_{i=1}^{l} c_i\phi_d(2^n\mathbf{x} - \mathbf{k}_i)$, $\mathbf{k}_i \in \mathbb{Z}^d$ so that for any fixed $r > d/4$,

$$\left\| f(x) - \sum_{i=1}^{l} c_i\phi_d(2^n\mathbf{x} - \mathbf{k}_i) \right\|_{L^\infty,(1+|\mathbf{x}|^2)^{-r}}$$

is bounded above by a sum of two error bounds.

Specifically

$$\left\| f(x) - \sum_{i=1}^{l} c_i\phi_d(2^n\mathbf{x} - \mathbf{k}_i) \right\|_{L^\infty,(1+|\mathbf{x}|^2)^{-r}}$$

$$\le 2^d \frac{2^{-(n+1)(s-d/2)}}{1 - 2^{(d/2-s)}} \|f\|_{\mathcal{L}^2_s} \mathrm{sup}_\lambda \|\psi^\lambda_d\|_{\mathcal{L}^2_{hom,-s}} + 4\tau\|g\|_{L^2}\|f\|_{L^2}\sqrt{\frac{2\ln em + \ln 4}{m}}. \tag{28}$$

where

$$c_i = \frac{\pm 2^{nd}}{l} \left\| \frac{f_n}{(1+|\mathbf{t}|^2)^r} \right\|_{L^2} (1 + |\mathbf{t}_i|^2)^r$$

for some choice of $\mathbf{t}_i \in \mathbb{R}^d$, $g(\mathbf{t}) = (1+|\mathbf{t}|^2)^{-r}$, $\tau = 2^{nd}(1+2^{-2n}d)^r$, and f_n denotes the projection of f onto V_n. (For proofs see [8].)

The first term in (28) is a standard error bound [7] for the difference between a function and its best approximation f_m (which itself is generally an infinite sum) in the scaling space V_m. The second term in (28) allows the infinite sum defining f_n to be replaced by a finite one with l terms, with an additional cost of $4\tau\|g\|\ \|f\|\sqrt{\frac{2\ln en + \ln 4}{n}}$ for an appropriate L^2 function g. In this case the VC dimension of the family

$$\mathcal{F} = \{K(\mathbf{x},\mathbf{t})\}_{\mathbf{x}\in\mathbb{R}^d} = \{2^{md}\phi_d(2^m\mathbf{t} - \mathbf{k})\}_{\mathbf{k}\in\mathbb{Z}^d}$$

(in the parameter $\mathbf{k}$ and the variable $\mathbf{t}$) is $h = 2$.

The following corollary for a positive semi-definite complex kernel is immediate from Proposition 5 with appropriate modification of the constants. That is, the results in Proposition 5 remain valid when the space $\mathbb{R}^d$ is replaced with $\mathbb{C}^d$ with $\mathbb{C}$ the complex numbers. In the case when the kernel and the function are complex, the constant 4τ in the bound of Theorem 4 is replaced by $8\sqrt{2}\tau$ and $\ln 4$ by $\ln 16$.

Corollary 6. *Under the assumptions of Proposition 5, let f be in a complex RKHS $\mathcal{H}_K$ with the reproducing projection kernel $K(\mathbf{x},\mathbf{t}) = K^{1/2}(\mathbf{x},\mathbf{t})$ with $\mathbf{x},\mathbf{t} \in D \subset \mathbb{C}^d$.*

Let the positive functions $g(\mathbf{t})$, $k(\mathbf{x})$, $g \in L^2(\mathbb{C}^d)$, bounded away from 0*, be such that*

$$\text{ess sup}_{\mathbf{x},\mathbf{t}} \left| \frac{K(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \right| \leq \tau$$

Let h be the VC dimension of the family

$$\frac{K(\mathbf{x},\mathbf{t})}{g(t)k(x)}$$

in the parameter x and variable t on D, i.e., the VC dimension of the real together with the imaginary parts of this function family on D. Then for every $f \in \mathcal{H}_K$ and every $\epsilon > 0$ there exist $\{\mathbf{t}_1, \dots, \mathbf{t}_n\}$ and n coefficients $c_j = a_j + ib_j$ (with $a_j, b_j = 0, \pm 1$) such that the weighted L^∞ norm

$$\left\| f(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^{n} c_i \|fg\|_{L^1} \frac{K(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)} \right\|_{L^\infty, 1/k(x)}$$

$$\leq 8\sqrt{2}\tau \|f\|_{L^2} \|g\|_{L^2} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 16}{n}} + \epsilon.$$

Proof. As the kernel and function are complex the constant 4τ must be replaced by $2\sqrt{2}$ times 4τ and the term $\ln 4$ is replaced by $\ln 16$. This follows as the above theorem is used for the real and imaginary parts (each of which consists of two integrals, since f has two components) separately, and the constant η is allowed to approach $\frac{1}{4}$ instead of 1, since we wish in this case to have a non- vanishing probability that the real and imaginary approximations (i.e. four integrals all together) *simultaneously* approximate the function $f(z)$. □

The corollary applies to the projection kernel onto the Haar scaling space in complex $L^2(\mathbb{R}^d)$ [8]; the sinc kernel for the Paley-Weiner space; the Szegö kernel for the Hardy space; and the Bergman kernel for the space of all functions f that are analytic in the open unit disk and have finite L^2 norm on the open unit disk [20]. We note that the VC dimensions of these kernels are unknown.

Remark: We remark that if the kernels $\frac{K^{1/2}(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)k(\mathbf{x})}$ of Theorem 4 and Proposition 5 (with $K^{1/2}(\mathbf{x},\mathbf{t}) = K(\mathbf{x},\mathbf{t})$) are uniformly continuous in $\mathbf{t}$ and the parameter $\mathbf{x}$, then the bounds in this paper hold with $\epsilon = 0$ [see proof in [8], Corollary 3]. Note that the uniform continuity condition is satisfied if $K^{1/2}$ is uniformly continuous in $\mathbf{t}$, g is continuous and g, k are bounded away from zero.

5. Conclusion

We have shown here that there exists a set of points $\mathbf{t}_i$ in the domain of a reproducing kernel Hilbert space $\mathcal{H}_K$ at which an $f \in \mathcal{H}_K$ can be approximated as combination of its values $f(\mathbf{t}_i)$. We conclude by restating that underlying the above approximation results is the important and unsolved optimization problem of finding approximations to the vectors $\{\mathbf{t}_i\}_i$ that yield non-asymptotic uniform error bounds given in (19) and (25).

6. Acknowledgement

We thank Prof. Maria Pensky for helpful references on kernel density estimation.

References

[1] Aronszajin, Theory of reproducing kernels, Trans. of AMS, **68** (1950) 337–404.

[2] Chui, C. K., *An Introduction to Wavelets* Academic Press, 1992.

[3] Cuker, F. and S. Smale. On the mathematical foundations of learning, AMS Bull. **39**, No. 1 (2002) 1–49.

[4] Girosi, F., Approximation error bounds that use VC bounds, in *Proc. International Conference on Artificial Neural Networks*, F. Fogelman-Soulied and P. Gallinari (eds), Paris (1995) 295–302.

[5] Girosi, F., An equivalence between sparse approximation and support vector machines, Neural Computation, **10** (1998) 1455–1480.

[6] Härdle, W., G. Kerkyacharian, D. Picard, A. Tsybakov. Wavelets, Approximation and Statistical Applications, Lecture Notes in Statistics, 129, Springer, 1998.

[7] Kon, M. and L. Raphael. Convergence Rates of Multiscale and Wavelet Expansions. *Wavelet Transforms and Time-Frequency Signal Analysis*, CBMS Conference Proceedings, L. Debnath, Ed., Chapter 2, Birkhauser, 2001, pp. 37-65.

[8] Kon, M., L. Raphael and D. Williams. Extending Girosi's Approximation Estimates for Functions in Sobolev Spaces via Statistical Learning Theory, J. of Analysis and Applications **3**, No. 2 (2005) 67-90.

[9] Lu, F., S. Keles, Wright, S.J., and G. Wahba. A Framework for Kernel Regularization with Application to Protein Clustering. Tech. Rep. 1107, U. of Wisconsin, May, 2005.

[10] Massachusetts Institute of Technology - Artificial Intelligence: http://www. ai.mit.edu/projects/cbcl/projects/NoticesAMS /PoggioSmale.htm.

[11] Mhaskar, H. N. On the tractability of multivariate integration and approximation by neural networks, J. of Complexity, **20**, (2004) 561–590.

[12] Poggio, T. and S. Smale. The mathematics of learning: dealing with data. Notices AMS, 50, No. 5 (2002) 537–544.

[13] Scholkopf, B. and A. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Organization, and Beyond*, MIT, 2002.

[14] Shawe-Taylor, J. and N Cristianini. *Kernel Methods for Pattern Analysis*, Cambridge, 2004.

[15] Smale, S. and D-X. Zhou. Estimating the approximation error in learning theory, Analysis and Applications **1**, No. 1 (2003) 1–25.

[16] Stein, E.M. *Singular Integrals and Differentiability Properties of Functions.* Princeton University, 1970.

[17] Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed., Springer, 2000.

[18] Wahba, G. *Spline Models for Observational Data*, **59** CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1990.

[19] Walter, G, *Wavelets and Other Orthogonal Systems with Applications*, CRC Press, 1994.

[20] Young, R.M., *Nonharmonic Fourier Series*, revised first edition, Academic Press, 2001.

[21] Zhou, D-X. Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49**, No. 7, (2003) 1743–1782.

(Kon) Department of Mathematics and Statistics, Boston University, Boston, MA, 02215 USA

E-mail address: mkon@math.bu.edu

(Raphael) Department of Mathematics, Howard University, Washington, DC 20059 USA

E-mail address: lraphael@howard.edu