EXTENDING GIROSI'S APPROXIMATION ESTIMATES FOR FUNCTIONS IN SOBOLEV SPACES VIA STATISTICAL LEARNING THEORY

MARK A. KON LOUISE A. RAPHAEL DANIEL A. WILLIAMS

Abstract. Abstract: An extension of Girosi's sup-norm approximation error bound using the notion of VC dimension in statistical learning theory is derived for kernel spaces, and in particular, L^p-Sobolev spaces $\mathcal{L}_s^p(\mathbb{R}^d)$, $1 \leq p < \infty$, $s > 0$. Applications include non-asymptotic, uniform error bound approximations of $f \in \mathcal{L}_s^p(\mathbb{R}^d)$ by finite linear combinations of weighted Gaussians with different centers and variances for $1 \leq p < \infty$, and by Haar scaling functions when $p = 2$.

1. Introduction

Girosi [8] established an interesting connection between statistical learning theory (SLT) and approximation theory, showing that SLT methods can be used to prove results of a purely approximation theoretic nature. The probabilistic framework used is a powerful one, but it requires existence of L^1 norms.

We show that it is possible to extend these probabilistically-based bounds to analogous ones using more general L^p norms for $1 \leq p < \infty$. In particular, we obtain such approximation bounds for L^p Sobolev spaces $\mathcal{L}_s^p$ for $1 < p < \infty$, where s is the smoothness parameter. We note that approximation by radial basis functions included in the type discussed here has been analyzed extensively in the context of neural network theory (see e.g. [16, 17, 18, 21, 23], and especially [24]); for regularization and neural networks methods (a precursor to regularization in SLT) see [9, 20, 24].

Girosi's novel idea is to exploit the VC dimension-based error bounds on the difference between expected and empirical risks. He presents a straightforward derivation of a non-asymptotic uniform error bound for an approximation of a function f in a kernel space (i.e., a function space defined as the range of an operator kernel).

What is surprising about his general result on $\mathbb{R}^d$ is that the error is of the order $O\left(\sqrt{(h/n)\ln(n/h)}\right)$ where n is the number of data points and h is the VC dimension. In Girosi's application involving approximation of L^1 Sobolev functions by Gaussians, $h = d + 1$, and this uniform bound avoids the so-called curse of dimensionality.

The outline of the paper is as follows. First we recall some notions from SLT, including the seminal VC Bound Theorem of Vapnik and Chervonenkis [29]. Theorem

Date: November 26, 2004.

1991 *Mathematics Subject Classification.* Primary 68T05, 41A65; Secondary 65T60.

Key words and phrases. Approximation; empirical risk minimization principle; Gaussian; Sobolev space; statistical learning theory; VC dimension.

The first author's research was partially supported by the National Science Foundation.

2 of Section 3 is an extension of Girosi's estimates for determining non-asymptotic, uniform VC error bounds for kernel operators. Girosi's application to approximating functions in Sobolev spaces $\mathcal{L}_s^1(\mathbb{R}^d),\ s > 0$, by linear combinations of translates of Bessel potential kernels is extended in Corollary 4 to approximating $f \in \mathcal{L}_s^p(\mathbb{R}^d),\ s > 0$, for $1 \le p < \infty$, with respect to a weighted supremum norm. Other applications for reproducing kernel spaces and Haar wavelets are also given in Section 4.

2. SLT Background and Definitions

Let $\mathbf{X} \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}^1$. Assume the set $\mathbf{X} \times Y$ is sampled n times under an unknown probability distribution $P(\mathbf{x}, y)$, and denote the data set by $\{(\mathbf{x}_i, y_i) \in \mathbf{X} \times Y\}_{i=1}^n$. Here $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$ is defined on $\mathbf{X} \times Y$, with $P(y|\mathbf{x})$ the *conditional probability* and $P(\mathbf{x})$ the *marginal probability*. Given samples $\{(\mathbf{x}_i, y_i) \in \mathbf{X} \times Y\}_{i=1}^n$, an important problem in SLT is, for a given hypothesis space $\mathcal{H}$, to find a function $f : \mathbf{X} \to Y$ in $\mathcal{H}$ such that when $\mathbf{x} \in \mathbf{X}$ is given, f predicts a value for y optimally.

Within the framework of SLT, we follow Vapnik's probabilistic bound approach, which involves the VC dimension [29, 30]. We refer the reader to the excellent articles [7, 19], standard SLT references [4, 25, 29], and the comprehensive bibliographic databases [13, 14] .

For $f \in \mathcal{H}$, we will assume a given *loss function* $V(y, f(\mathbf{x})) = V(f, \mathbf{z})$, where $\mathbf{z} = (\mathbf{x}, y)$, which measures the error between y and the predicted value $f(\mathbf{x})$. Two examples are $V(f, \mathbf{z}) = |y - f(\mathbf{x})|^p,\ 1 \le p < \infty$ and the $\{0, 1\}$-valued function $V(f, \mathbf{z}) = 1 - \chi_{[-1,1]}(y - f(\mathbf{x}))$.

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

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

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

$$f^* = \arg \left\{ \min_{f \in \mathcal{H}} R[f] \right\} \tag{2}$$

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 a stochastic approximation of the expected risk, called the *empirical risk*. For a function $f \in H_k \subset \mathcal{H}$ and loss function $V(f, \mathbf{z})$, define

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

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

In applications, $\mathcal{H}$ is often too large and so the empirical risk is successively minimized on a nested sequence of increasing subspaces $H_0 \subset H_1 \subset \cdots \subset H_k \subset \cdots \subset \mathcal{H}$, where the subscript k denotes the "capacity" of the set H_k. Standard examples of

linear H_k spaces include: splines with k nodes, and degree k trigonometric polynomials in d variables. The results in this paper are stated in terms of H_k.

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

$$f_{k,n} = \arg\left\{\min_{f \epsilon H_k} R_{emp}[f;n]\right\}. \tag{3}$$

As $n \to \infty$, ideally $f_{k,n} \in H_k$ converges to

$$f_k = \arg\left\{\min_{f \epsilon H_k} R[f]\right\}. \tag{4}$$

For a hypothesis space H_k the precise requirement is

$$\lim_{n\to\infty} R_{emp}[f_{k,n};n] = \lim_{n\to\infty} R[f_{k,n}] = R[f_k]. \tag{5}$$

Seminal work of Vapnik and Chervonenkis [29] shows that (5) is satisfied in H_k when the following one-sided uniform convergence in probability holds for all $\epsilon > 0$:

$$\lim_{n\to\infty} P\left\{\sup_{f\in H_k} \left(R[f] - R_{emp}[f;n]\right) > \epsilon\right\} = 0. \tag{6}$$

We begin a detailed discussion of non-asymptotic VC bounds by defining the VC dimension.

Definition 1. *The VC dimension of a set of functions $\{V(f,\mathbf{z}) : f \in H\}$ is the maximum number h of vectors $\{\mathbf{z}_i\}_{i=1}^h$ that can be separated by functions in this set into two classes $\{0,1\}$ in all 2^h possible ways using the rules:*

$$\begin{cases} \textit{Class } 1: & \textit{if } V(f,\mathbf{z}_i) - \alpha \geq 0 \\ \textit{Class } 2: & \textit{if } V(f,\mathbf{z}_i) - \alpha < 0 \end{cases}$$

where $\alpha \in \mathbb{R}, f \in H$. If such a separation is possible the set $\{\mathbf{z}_i\}_{i=1}^h$ is said to be shattered by H.

In the above definition we sometimes use the convention that f is the parameter and $\mathbf{z}$ is the variable, since f (along with α) is fixed for each separation in the variable $\mathbf{z}$.

Two sets of real functions with VC dimension $d+1$ are:

- characteristic functions of half-planes on $\mathbb{R}^d$ (in the variable $\mathbf{z}$)
- characteristic functions of circles on $\mathbb{R}^d$.

The following well-known theorem of Vapnik and Chervonenkis, which gives probabilistic estimates of integrals by finite sums, is used in the main results of [8] and this paper.

Theorem 1. *(VC Bound Theorem - [29]). Let $V(y,f(\mathbf{x})) = V(f,\mathbf{z})$, $\mathbf{z} = (\mathbf{x},y)$, satisfy $A \leq V(f,\mathbf{z}) \leq B$ for f in 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 the following inequality holds simultaneously for all $f \in H_k$, with probability at least $1-\eta$:*

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

3. Extending Girosi's Results to $\mathcal{L}_s^p(\mathbb{R}^d),\ p > 1$

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 the smoothness s change with dimension d enables the approximation error to be better than $O(n^{-s/d})$, and some researchers have used this technique to deal with such strong dimensional dependence [2, 3, 15, 17, 21, 22].

This problem is also dealt with in [8], where there is a reinterpretation of SLT notions as follows:

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

Under these replacements the expected risk

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

becomes

$$f(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{t})\lambda(\mathbf{t})d\mathbf{t}, \tag{7}$$

and the empirical risk

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

Girosi [8] used the VC Bound Theorem of Vapnik and Chervonenkis, to find estimates of approximation of some integrals of the form (7) when $\lambda(\mathbf{t}) \in L^1(\mathbb{R}^d)$. We now describe a modification of Girosi's result applied to functions of the form (7).

Note that we assume (as in [29] and [8]) that the kernel K is bounded above and below, i.e. $A \le K(\mathbf{x}, \mathbf{t}) \le B$. The following probabilistic error bound holds with probability $1 - \eta$ for $\lambda \in L^1(\mathbb{R}^d)$ (here λ can be both positive and negative) and a sample of n points $\{\mathbf{t}_i\}_{i=1}^n$ taken with respect to the probability density $|\lambda(x)|dx$ (normalized to unit L^1 norm if necessary):

$$\left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\lambda(\mathbf{t}_i))K(\mathbf{x}, \mathbf{t}_i)\|\lambda\|_{L^1} \right\|_{L^\infty} \le 4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} - \ln\frac{\eta}{4}}{n}}. \tag{8}$$

where $\tau = B - A$.

Note that for every positive $\eta < 1$, this implies there exists a sample $\{\mathbf{t}_i\}_{i=1}^n$ such that (8) holds. Letting $\eta \uparrow 1$, we see that the right hand side of (8) 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

$$\left\|f(\mathbf{x})-\frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(\lambda(\mathbf{t}_i))K(\mathbf{x},\mathbf{t}_i)\|\lambda\|_{L^1}\right\|_{\infty}\leq 4\tau\|\lambda\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h}+\ln 4}{n}}+\epsilon. \tag{9}$$

Note that we require the additional $\epsilon>0$ (not given in [8]) on the right side in order for the statement to be true for the most general bounded kernels $K(\mathbf{x},\mathbf{t})$. For kernels which are uniformly continuous in $\mathbf{x}$ as in Corollary 3 below, we show that (9) holds for $\epsilon=0$.

Remarks: It should be emphasized that the probabilistic approach based on the VC Bound Theorem does not give any constructive method for finding a set of vectors $\{\mathbf{t}_i\}_{i=1}^n$.

The term 4τ in the bound arises as follows. First, note that if $|K(\mathbf{x},\mathbf{t})|\leq\tau$, then $A=-\tau\leq K(\mathbf{x},\mathbf{t})\leq\tau=B$, so that the factor $B-A$ becomes 2τ. The additional factor of 2 in the term 4τ is a consequence of writing the coefficients $c_i=c_i^+-c_i^-$, the sum of their negative and positive parts (i.e., $c^+=\sup(c,0)$ and $c^-=\sup(-c,0)$).

In the following we extend Girosi's estimates to the case $\lambda\in L^p(\mathbb{R}^d)$, $1\leq p<\infty$. We recall that a weighted L^∞ norm is defined by

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

Definition 2. *Let $K(\mathbf{x},\mathbf{t})$ be an operator kernel. For fixed $1\leq p<\infty$, we define its range $F_K=\left\{\,f(\mathbf{x})=\int K(\mathbf{x},\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}\;\middle|\;\lambda(t)\in L^p(\mathbb{R}^d)\right\}$.*

We define the VC dimension of K (in the variable $\mathbf{t}$ and parameter $\mathbf{x}$) as in Definition 1. That is the maximum number h of vectors $\{\mathbf{t}_i\}_{i=1}^h$ which can be separated into two classes in all possible ways, using classes of the form $K(\mathbf{x},\mathbf{t}_i)-\alpha\geq 0$ and $K(\mathbf{x},\mathbf{t}_i)-\alpha\leq 0$, as the parameters $\mathbf{x}$ and α vary.

Note that in Theorem 2 below, it is not required that the kernel $K(\mathbf{x},\mathbf{t})$ be bounded as long as (10) holds.

Theorem 2. *Let $1\leq p<\infty$. Assume $f\in F_K$ and that there exist positive functions g and k with $g(\mathbf{t})\in L^q$, $\frac{1}{p}+\frac{1}{q}=1$ such that*

$$\text{ess sup}_{\mathbf{x},\mathbf{t}}\left|\frac{K(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})}\right|\leq\tau. \tag{10}$$

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 writing

$$f(\mathbf{x})=\int K(\mathbf{x},\mathbf{t})\lambda(\mathbf{t})d\mathbf{t},\quad\lambda(\mathbf{t})\in L^p(\mathbb{R}^d),$$

for every $\epsilon>0$ there exist $\{\mathbf{t}_1,\ldots,\mathbf{t}_n\}\subset\mathbb{R}^d$, and n coefficients $c_i=\operatorname{sgn}(\lambda(\mathbf{t}_i))=\pm 1$, such that the weighted L^∞ norm

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

$$\leq 4\tau \|g\|_{L^q} \|\lambda\|_{L^p} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}} + \epsilon. \tag{11}$$

Proof: For positive $g \in L^q(\mathbb{R}^d)$, $\frac{1}{p}+\frac{1}{q} = 1$, $\lambda(\mathbf{t})g(\mathbf{t}) \in L^1$ by Hölder's inequality. Thus

$$\frac{f(\mathbf{x})}{k(\mathbf{x})} = \int \frac{K(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \lambda(\mathbf{t})g(\mathbf{t})d\mathbf{t}$$

replacing $\lambda(\mathbf{t})$ by $\lambda(\mathbf{t})g(\mathbf{t})$ (since $\lambda(\mathbf{t})g(\mathbf{t}) \in L^1$). Replacing $K(\mathbf{x},\mathbf{t})$ by $K(\mathbf{x},\mathbf{t})/g(\mathbf{t})$, we conclude that by (9) for every $\epsilon > 0$ there exist $\mathbf{t}_i$ such that

$$\begin{aligned}
&\left\| \frac{f(\mathbf{x})}{k(\mathbf{x})} - \frac{1}{n}\sum_{i=1}^{n} \frac{K(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)k(\mathbf{x})} \ \|\lambda(\mathbf{t})g(\mathbf{t})\|_{L^1} \mathrm{sgn}(\lambda(\ \mathbf{t}_i)g(\mathbf{t}_i)) \right\|_{L^\infty} \\
&\quad \leq \|\lambda\|_{L^p}\|g\|_{L^q} \left\| \int \frac{K(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \frac{\lambda(\mathbf{t})g(\mathbf{t})}{\|\lambda(\mathbf{t})g(\mathbf{t})\|_{L^1}} d\mathbf{t} \right. \\
&\qquad \left. - \frac{1}{n}\sum_{i=1}^{n} \frac{K(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)k(\mathbf{x})} \mathrm{sgn}(\lambda(\mathbf{t}_i)g(\mathbf{t}_i)) \right\|_{L^\infty} \\
&\quad \leq 4\tau \|\lambda\|_{L^p}\|g\|_{L^q} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}} + \epsilon
\end{aligned}$$

where h is the VC dimension of the kernel $K(\mathbf{x},\mathbf{t})/(g(\mathbf{t})k(\mathbf{x}))$. Note that the first inequality above uses Hölder's inequality. ■

We remark that this proposition applies to kernel spaces such as wavelet and Sobolev spaces (see below).

We now give an immediate application of these results to approximation of functions in Sobolev spaces, extending Girosi's L^1 results. The generalized Sobolev space $\mathcal{L}_s^p(\mathbb{R}^d)$ for $s > 0$ and $p \geq 1$, using the notation of Stein [28], is now defined.

Definition 3. *For $1 \leq p \leq \infty$, $s \in \mathbb{R}^+$ and defining*

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(\mathbf{x}) \exp(2\pi i \mathbf{x}\cdot\xi) d\mathbf{x},$$

we define the generalized Sobolev space as:

$$\mathcal{L}_s^p(\mathbb{R}^d) \equiv \left\{ f \in L^p(\mathbb{R}^d) : (I-\Delta)^{s/2} f \in L^p(\mathbb{R}^d) \right\},$$

with norm $\|f\|_{\mathcal{L}_s^p} = \|(I-\Delta)^{s/2} f\|_{L^p(\mathbb{R}^d)}$, where I is the identity operator, Δ is the Laplacian defined via

$$\widehat{\Delta f(\mathbf{x})} = -4\pi^2|\xi|^2 \widehat{f}(\xi),$$

and $\xi = (\xi_1, ..., \xi_d)$ denotes the Fourier variable dual to $\mathbf{x} = (x_1, ..., x_d)$.

Thus when $f \in \mathcal{L}_s^p(\mathbb{R}^d)$, there exists a $\lambda \in L^p$ such that

$$f = (I-\Delta)^{-s/2}\lambda \ \ and \ \ \|f\|_{\mathcal{L}_s^p} = \|\lambda\|_{L^p}.$$

Note that the above can be rewritten

$$\begin{aligned} f &= (I-\Delta)^{-s/2}\lambda \\ &= (\widehat{G}_s\hat{\lambda})^\vee \\ &= G_s * \lambda \end{aligned} \tag{12}$$

where

$$\widehat{G}_s(\xi) = \frac{1}{(1+4\pi^2|\xi|^2)^{s/2}}, \tag{13}$$

$u^\vee$ denotes the inverse Fourier transform of u, and * denotes convolution. We will use the integral representation of G_s ([28] p. 132) given by

$$G_s(\mathbf{x}) = \frac{(4\pi)^{-s/2}}{\Gamma(\frac{s}{2})} \int_0^\infty \exp\left(-\frac{\pi}{\sigma}|\mathbf{x}|^2\right) \exp\left(-\frac{\sigma}{4\pi}\right) \sigma^{(s-d-2)/2} d\sigma. \tag{14}$$

to prove Corollary 3 and derive an application of Theorem 2 in Section 4.

The next application of Theorem 2 is an extension of Proposition 3.1 in [8] to $\mathcal{L}_s^p$ spaces with $1 \le p < \infty$. Note that the condition $s > d$ guarantees that the kernel $G_s(\mathbf{x}-\mathbf{t})$ is continuous at the origin, and in fact is uniformly continuous in $\mathbf{x}$ as $\mathbf{t}$ varies — see the proof of the Corollary below.

Corollary 3. *Let $1 \le p < \infty$. Let $f \in \mathcal{L}_s^p(\mathbb{R}^d)$ with $s > d$. Let G_s be the kernel of $(I-\Delta)^{-s/2}$ (see (14)) and $\lambda = (I-\Delta)^{s/2} f$. Assume there exist positive functions g and k with $g(\mathbf{t}) \in L^q$, $\frac{1}{p} + \frac{1}{q} = 1$ such that*

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

Let h_s be the VC dimension of

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

in the parameter $\mathbf{x}$ and variable $\mathbf{t}$. Then for some $m \le n$ there exist $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\} \subset \mathbb{R}^d$, and m coefficients $c_i = \text{sgn}(\lambda(\mathbf{t}_i)) = \pm 1$, such that the weighted L^∞ norm

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

$$\le 4\tau \|g\|_{L^q} \|f\|_{\mathcal{L}_s^p} \sqrt{\frac{h_s \ln \frac{2en}{h_s} + \ln 4}{n}}. \tag{15}$$

Proof: Since $f \in \mathcal{L}_s^p(\mathbb{R}^d)$, there exists $\lambda(\mathbf{t}) \in L^p$ such that

$$f(\mathbf{x}) = [G_s * \lambda](\mathbf{x}) = \int G_s(\mathbf{x}-\mathbf{t})\lambda(\mathbf{t}) d\mathbf{t}$$

and, in addition, $\|f\|_{\mathcal{L}_s^p} = \|\lambda\|_{L^p}$, ([28], p.134). For $g \in L^q(\mathbb{R}^d)$, $\frac{1}{p} + \frac{1}{q} = 1$, we have $\|\lambda(\mathbf{t})g(\mathbf{t})\|_{L^1} \le \|\lambda(\mathbf{t})\|_{L^p} \|g(\mathbf{t})\|_{L^q}$.

Thus by Theorem 2, for each $j > 0$, there exists a sample $T_j = \{\mathbf{t}_{ij}\}_{i=1}^n$, and $c_{ij} = \pm 1$ such that

$$\left\| f(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n c_{ij} \|\lambda g\|_{L^1} \frac{G_s(\mathbf{x}-\mathbf{t}_{ij})}{g(\mathbf{t}_{ij})} \right\|_{L^\infty, 1/k(\mathbf{x})}$$

$$\le 4\tau \|g\|_{L^q} \|f\|_{\mathcal{L}_s^p} \sqrt{\frac{h_s \ln \frac{2en}{h_s} + \ln 4}{n}} + \frac{1}{j}\ . \tag{16}$$

If the sequence of samples $\{T_j\}_{j=1}^{\infty}$ is bounded, it has a subsequence with a limit $T = \{\mathbf{t}_i\}_{i=1}^{n}$. Further (by taking a sub-sequence if necessary) we may assume that for each i, a limit c_i of the j-sequence $c_{ij} = \mathrm{sgn}(\lambda(t_{ij}))$ exists. In this case inequality (15) holds for this choice $T = \{\mathbf{t}_i\}$ and these c_i. This follows from the uniform continuity of $G_s(\mathbf{x} - \mathbf{t}_i)$ (see below), which shows that as $j \to \infty$, $G_s(\mathbf{x} - \mathbf{t}_{ij}) \to G_s(\mathbf{x} - \mathbf{t}_i)$ in the $L^\infty[\mathbf{x}]$ norm.

On the other hand, if $\{T_j\}_j$ is unbounded, then a subsequence converges to ∞, which implies that $t_{ij} \to \infty$ as $j \to \infty$ for some fixed i. In this case the i^{th} term in the sum in (16) can be eliminated without loss (since for sufficiently large $\mathbf{t}_{ij}$ the term $c_{ij}\|\lambda \cdot g\|_{L^1} \frac{G_s(\mathbf{x}-\mathbf{t}_{ij})}{g(\mathbf{t}_{ij})}$ in (16) can only increase the error on the left side). Successively eliminating all terms in the sum of (16) with i for which $\mathbf{t}_{ij} \to \infty$ as $j \to \infty$ in this way, we have left a set of terms $\{\mathbf{t}_i\}_{i=1}^{m}$ in the sum in (16) which we may, without loss of generality, assume are numbered from 1 to $m \leq n$ and for which (15) holds as desired. Note that this process of elimination must stop before the last term is eliminated (for sufficiently large n), since the zero approximation (i.e., that with no terms in it) cannot approximate a function with arbitrary accuracy.

Finally, to prove the uniform continuity of $G_s(\mathbf{x} - \mathbf{t}_i)$ note that since $s > d$, $G_s(\mathbf{x} - \mathbf{t}_i)$ is continuous in $\mathbf{x}$ [28] (and so uniformly continuous on any compact set), and that it decays at infinity, since $G_s(\mathbf{x} - \mathbf{t}_i) = O(\exp(-c|\mathbf{x}|))$ as $|\mathbf{x}| \to \infty$, ([28], page 132). ■

Remark. In general the weight $1/k(\mathbf{x})$ may be needed to counterbalance the effect of dividing the kernel by an L^q function $g(\mathbf{t})$ in order that the full kernel $K(\mathbf{x}, \mathbf{t})/(g(\mathbf{t})k(\mathbf{x}))$ remains bounded. This weight is needed for that reason in our next example involving Gaussian kernels.

4. Applications

To approximate $f \in \mathcal{L}_s^p(\mathbb{R}^d)$, $1 \leq p < \infty$, by a sum of weighted Gaussians with different centers and variances, we follow Girosi [8] and write the Bessel kernel G_s in its integral representation (14). Recall from Section 3 that for $f \in \mathcal{L}_s^p(\mathbb{R}^d)$, there exists a $\lambda \in L^p$ such that $f = G_s * \lambda$, so

$$f(\mathbf{x}) = \int_0^\infty \int_{\mathbb{R}^d} \exp\left(-\frac{\pi}{\sigma}|\mathbf{x} - \mathbf{t}|^2\right) \Lambda(\mathbf{t}, \sigma) d\mathbf{t} d\sigma.$$

where

$$\Lambda(\mathbf{t}, \sigma) = \frac{(4\pi)^{-s/2}}{\Gamma(\frac{s}{2})} \exp\left(-\frac{\sigma}{4\pi}\right) \sigma^{(s-d-2)/2} \lambda(\mathbf{t}). \tag{17}$$

Letting $\mathbb{R}^+$ denote the non-negative real numbers, f has the form

$$f(\mathbf{x}) = \int_{\mathbb{R}^d \times \mathbb{R}^+} K(\mathbf{x}; \mathbf{t}') \Lambda(\mathbf{t}') d\mathbf{t}'$$

where $K(\mathbf{x}; \mathbf{t}') = K(\mathbf{x}; \mathbf{t}, \sigma) = \exp\left(-\frac{\pi}{\sigma}|\mathbf{x} - \mathbf{t}|^2\right)$ is now the Gaussian. To apply Theorem 2, we have replaced $\mathbf{t} = (t_1, t_2, ..., t_n)$ by $\mathbf{t}' = (t_1, t_2, ..., t_n, \sigma)$. Let

$$B(\sigma) = \frac{(4\pi)^{-s/2}}{\Gamma(\frac{s}{2})} \exp\left(-\frac{\sigma}{4\pi}\right) \sigma^{(s-d-2)/2}. \tag{18}$$

From above if $\Lambda(\mathbf{t}') \in L^p(\mathbb{R}^d \times \mathbb{R}^+)$, it follows from the product form of $\Lambda(\mathbf{t}, \sigma) = B(\sigma)\lambda(\mathbf{t})$ that $\Lambda(\mathbf{t}, \sigma) \in L^p(\mathbb{R}^d \times \mathbb{R}^+)$ if and only if $(\frac{s-d}{2} - 1)p > -1$ or $s > d + 2 - \frac{2}{p}$.

The space $\mathcal{L}_s^p(\mathbb{R}^d)$ is contained in the continuous functions $C(\mathbb{R}^d)$ for $s - \frac{d}{p} > 0$. When $p > 1$, we have $d + 2 - \frac{2}{p} > \frac{d}{p}$, since $dp + 2p - 2 > d$ follows from $dp > d$ and $2p - 2 > 0$. Therefore if we assume that $s > d + 2 - \frac{2}{p}$, it follows $s - \frac{d}{p} > 0$, so that $\mathcal{L}_s^p(\mathbb{R}^d) \subset C(\mathbb{R}^d)$ and of course $\Lambda(\mathbf{t}, \sigma) \in L^p(\mathbb{R}^d \times \mathbb{R}^+)$.

Girosi uses a theorem of Dudley [5], to show that the the VC dimension of the family $K(x,t) = \exp\left(-\frac{\pi}{\sigma}|\mathbf{x} - \mathbf{t}|^2\right)$ is $d + 1$ (d = dimension), so that the supremum norm approximation of an arbitrary function in $\mathcal{L}_s^1(R^d)$ by a linear superposition of scaled Gaussians with different centers and variances has a bound of order

$$\sqrt{\frac{(d+1)\ln\frac{n}{d+1}}{n}},$$

where n is the number of data points. Our next example shows that for $s > d+2-\frac{2}{p}$, a weighted bound for approximating any $\mathcal{L}_s^p(\mathbb{R}^d)$ function f by such superpositions has the same form as Girosi's, with the replacement of $d + 1$ by $d + 2$ for the VC dimension.

Within the proof of Corollary 4 below, we prove [Proposition 5] that the VC dimension of our family of kernels is less than that of the collection of sets bounded by all hyperplanes in $\mathbb{R}^{d+1}$, which is $d + 2$.

Definition 4. *We define*

$$K(\mathbf{x}; \mathbf{t}, \sigma) = \frac{\exp\left(\frac{-\pi|\mathbf{x}-\mathbf{t}|^2}{\sigma}\right)}{g(\mathbf{t})k(\mathbf{x})},$$

where

$$g(\mathbf{t}) = \exp\left(\frac{-J(\sigma)\pi}{3}\right)|\mathbf{t}|^2,$$

with

$$J(\sigma) = \begin{cases} 1 & \text{if } \sigma \le 1 \\ \frac{1}{\sigma} & \text{if } \sigma > 1 \end{cases}$$

$$k(\mathbf{x}) = \exp(\pi|\mathbf{x}|^2).$$

We define

$$\tau = \text{ess sup}_{\mathbf{x},\mathbf{t},\sigma} \left|K(x; t, \sigma)\right|.$$

Corollary 4. *Let $f \in \mathcal{L}_s^p(\mathbb{R}^d)$ with $s > d + 2 - \frac{2}{p}$, and let $\Lambda(\mathbf{t}, \sigma)$ be as in (17). Then for every $\epsilon > 0$ there exists means $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\} \subset \mathbb{R}^d$, variances $\{\sigma_1, \sigma_2, ..., \sigma_n\} \subset \mathbb{R}^+$ and coefficients $c_i = \text{sgn}(\Lambda(\mathbf{t}_i, \sigma)) = \pm 1$ such that*

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

$$\le 4\|f\|_{\mathcal{L}_s^p}\|g\|_{L^q}\|B(\sigma)\|_{L^p}\sqrt{\frac{(d+2)\ln\frac{2en}{d+2} + \ln 4}{n}} + \epsilon$$

where $B(\sigma)$ is as in (18).

Proof: We note that the norm $\|\Lambda g\|_{L^1}$ is taken in the variables $(\mathbf{t}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+$. By Theorem 2, we need to show:

(1) the kernel $K(\mathbf{x}; \mathbf{t}, \sigma)$ is bounded (in this case by the bound $\tau = 1$)

(2) $d+2$ is an upper bound of the VC dimension of the class of functions $K(\mathbf{x};\mathbf{t},\sigma)$ in the parameter $\mathbf{x}$ and variables $\mathbf{t},\sigma$.

Step 1. To show that $K(\mathbf{x};\mathbf{t},\sigma)$ is bounded by 1 for $\mathbf{x},\ \mathbf{t}\in\mathbb{R}^d, \sigma\in\mathbb{R}^+,\ \ d\geq 1$, it suffices to show

$$\frac{-1}{\sigma}|\mathbf{x}-\mathbf{t}|^2+\frac{J(\sigma)|\mathbf{t}|^2}{3}-|\mathbf{x}|^2<0,\quad \mathbf{x},\ \mathbf{t}\in\mathbb{R}^d,\ \textit{and}\ \ \sigma\in\mathbb{R}^+$$

Note that for $d\geq 1$,

$$\frac{-1}{\sigma}|\mathbf{x}-\mathbf{t}|^2+\frac{J(\sigma)|\mathbf{t}|^2}{3}-|\mathbf{x}|^2\leq\left(\frac{-1}{\sigma}-1\right)|\mathbf{x}|^2+\left(\frac{-1}{\sigma}+\frac{J(\sigma)}{3}\right)|\mathbf{t}|^2+\frac{2|\mathbf{x}||\mathbf{t}|}{\sigma}.$$

For a fixed σ, the polynomial on the right is a quadratic form in $|\mathbf{x}|$ and $|\mathbf{t}|$. The matrix of the form is

$$S=\begin{bmatrix} -\frac{1}{\sigma}-1 & \frac{1}{\sigma} \\ \frac{1}{\sigma} & -\frac{1}{\sigma}+\frac{J(\sigma)}{3}\end{bmatrix}.$$

The trace of S is $\frac{-2}{\sigma}-1+\frac{J(\sigma)}{3}$, which is negative for both cases of J. The determinant of S is $\frac{1}{\sigma}-\frac{J(\sigma)}{3\sigma}-\frac{J(\sigma)}{3}$, which is positive for both cases. As the determinant equals the product of the eigenvalues, we conclude that the eigenvalues of S are negative, and S is negative definite. Thus K is bounded uniformly by 1.

Step 2: We use the following proposition to show that the VC dimension of the class of functions determined by our weighted Gaussian kernel $K(\mathbf{x};\mathbf{t},\sigma)$ is less than the VC dimension of the subsets bounded by hyperplanes in $\mathbb{R}^{d+1}$.

Proposition 5. *(Bound on VC dimension of Weighted Gaussian Kernel) The VC dimension of the family*

$$K(\mathbf{x};\mathbf{t},\sigma)=\exp\left(\frac{-\pi|\mathbf{x}-\mathbf{t}|^2}{\sigma}+\frac{\pi J(\sigma)|\mathbf{t}|^2}{3}-\pi|\mathbf{x}|^2\right) \tag{19}$$

in the parameter $\mathbf{x}$ *and variables* $(\mathbf{t},\sigma)$ *is bounded by* $d+2$*, with* d *the dimension.*

Proof: Since the exponential function is one-to-one, it suffices to show that this VC dimension bound holds for the family of functions.

$$\left\{\frac{-\pi|\mathbf{t}-\mathbf{x}|^2}{\sigma}+\frac{\pi J(\sigma)|\mathbf{t}|^2}{3}-\pi|\mathbf{x}|^2:\ \ \mathbf{x}\in\mathbb{R}^d\right\}.$$

We write

$$\frac{-\pi|\mathbf{t}-\mathbf{x}|^2}{\sigma}+\frac{\pi J(\sigma)|\mathbf{t}|^2}{3}-\pi|\mathbf{x}|^2$$

$$=\left(\frac{\pi J(\sigma)}{3}-\frac{\pi}{\sigma}\right)|\mathbf{t}|^2+\frac{2\pi}{\sigma}\mathbf{t}\cdot\mathbf{x}-\left(\frac{\pi}{\sigma}+\pi\right)|\mathbf{x}|^2.$$

This family of functions of $(\mathbf{t},\sigma)$ (parameterized by $\mathbf{x}$) is at most $d+2$ dimensional, as each function is a linear combination of the $d+2$ fixed functions

$$\left\{\frac{\mathbf{t}_1}{\sigma},...,\frac{\mathbf{t}_d}{\sigma},\left(\frac{\pi J(\sigma)}{3}-\frac{\pi}{\sigma}\right)|\mathbf{t}|^2,\ -\left(\frac{\pi}{\sigma}+\pi\right)\right\}$$

(with coefficients dependent on the parameter $\mathbf{x}$). By [27], this family therefore has VC dimension bounded by $d+2$ and the proposition is proved. ■

The result of the corollary now follows from $\|\Lambda(\mathbf{t},\sigma)\|_{L^p} = \|\lambda(\mathbf{t})B(\sigma)\|_{L^p} = \|\lambda(\mathbf{t})\|_{L^p}\,\|B(\sigma)\|_{L^p}$, $\|f\|_{\mathcal{L}^p_s} = \|\lambda(\mathbf{t})\|_{L^p}$, and noting that by Theorem 2, for every $\epsilon > 0$ there exists a sample set $\{\mathbf{t}_{ij}\}_{i=1}^n$ such that

$$\begin{aligned}
&\left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} c_{ij}\|\Lambda g\|_{L^1}\frac{\exp\left(\frac{-\pi}{\sigma_i}|\mathbf{x}-\mathbf{t}_{ij}|^2\right)}{g(\mathbf{t}_{ij})}\right\|_{L^\infty,\ \exp(-\pi|\mathbf{x}|^2)} \\
&= \left\| \exp(-\pi|\mathbf{x}|^2)\left\{ \int_{\mathbb{R}^{d+1}} \frac{\exp\left(\frac{-\pi}{\sigma}|\mathbf{x}-\mathbf{t}|^2\right)}{g(\mathbf{t})} g(\mathbf{t})\Lambda(\mathbf{t},\sigma)d\mathbf{t}d\sigma \right.\right. \\
&\qquad \left.\left. -\frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(\Lambda(\mathbf{t}_{ij},\sigma))\|\Lambda g\|_{L^1}\frac{\exp\left(\frac{-\pi}{\sigma_i}|\mathbf{x}-\mathbf{t}_{ij}|^2\right)}{g(\mathbf{t}_{ij})}\right\}\right\|_{L^\infty} \\
&\qquad \le 4\tau\|\Lambda(\mathbf{t},\sigma)\|_{L^p}\|g\|_{L^q}\sqrt{\frac{(d+2)\ln\frac{2en}{d+2}+\ln 4}{n}} + \epsilon
\end{aligned}$$

where, as above, L^p norms of functions of σ and $\mathbf{t}$ are joint in the two variables. We note that in this case $\tau = \sup K = 1$. This completes the proof of Corollary 4. ■

We now mention some applications for reproducing kernel Hilbert spaces. Let H be a Hilbert space with inner product $(\cdot,\cdot)$, whose elements are real or complex-valued functions defined on a set $\mathbb{S}$, such that for every $\mathbf{x} \in \mathbb{S}$, the point-evaluation functional $f \to f(\mathbf{x})$ on H is bounded. By the Riesz representation theorem, for $\mathbf{x} \in \mathbb{S}$, there is an element $K_{\mathbf{x}} \in H$ such that for every $f \in H$,

$$f(\mathbf{x}) = (f, K_{\mathbf{x}}).$$

The function K on $\mathbb{S}\times\mathbb{S}$, defined by

$$K(\mathbf{x},\mathbf{t}) = (K_{\mathbf{x}}, K_{\mathbf{t}}) = K_{\mathbf{x}}(\ \mathbf{t}),$$

is called the reproducing kernel of H.

It is clear that the conclusion of Theorem 2 holds in the special case $\lambda = f$, which we will consider in the examples below.

The results in this paper also remain valid when the space $\mathbb{R}^d$ is replaced with $\mathbb{C}^d$, $\mathbb{C}$ the complex plane. Here in (9) (in the case that both 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 because Theorem 1 above 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 $f(z)$.

By the above observations, Corollary 4 (with these possible modifications) is valid for function spaces associated with the following reproducing kernels.

- Projection kernels of multiresolution spaces for frames and wavelets.
- The sinc kernel

$$K(z,w) = \frac{\sin\pi(z-\overline{w})}{\pi(z-\overline{w})}$$

for the Paley-Weiner space, i.e., the set of all entire functions of exponential type at most π that are square integrable on the real axis. The integral representation in this case is

$$f(z) = \int_{-\infty}^{\infty} f(t) \frac{\sin \pi(t-z)}{\pi(t-z)} dt, \quad z \in \mathbb{C}.$$

- The Szegö kernel
$$K(z,w) = \frac{1}{1 - z\overline{w}}$$
for the Hardy space, i.e., all functions f in the open unit disk in the complex plane, whose Taylor coefficients are square-summable.
- The Bergman kernel
$$K(z,w) = \frac{1}{\pi(1 - z\overline{w})^2}$$
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. The integral representation in this case is (here $z = x + iy$)
$$f(w) = \frac{1}{\pi} \int \int_{|z|<1} f(z) \frac{1}{(1 - \overline{z}w)^2} dxdy, \quad z,\ w \in \mathbb{C}.$$

Note that the problem of unboundedness of the above kernels can in some cases be eliminated with proper choice of weights $k(\mathbf{x})$ and $g(\mathbf{t})$.

Our final application is for Haar scaling functions and wavelets. We recall that a multiresolution analysis is a decomposition of $L^2(\mathbb{R}^d)$ into an increasing nested sequence of closed subspaces V_n, such that a function $f(\mathbf{x}) \in V_n$ if and only if $f(2\mathbf{x}) \in V_{n+1}$; $\bigcap V_j = 0$; $\overline{\bigcup V_j} = L^2(\mathbb{R}^d)$, where overline denotes closure; and V_0 is closed under multi-integer translations.

Let

$$\phi_d(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in [0,1]^d \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

denote the Haar scaling function in $\mathbb{R}^d$. Then at the scale $n = 0$, the 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

$$\phi(x_i) = \begin{cases} 1 & x_i \in [0,1] \\ 0 & \text{otherwise} \end{cases} \quad \text{or} \quad \psi(x_i) = \begin{cases} 1 & x_i \in [0,1/2] \\ -1 & x_i \in (1/2,1] \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Thus the total number of basic wavelets is $2^d - 1$ [31]. We define the homogeneous Sobolev space for $s \in \mathbb{R}$ by

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

Note that the bound below consists of two parts - the first is a standard error bound [11] for the difference between a function and its best approximation f_n (which itself is generally an infinite sum) in the scaling space V_n. The second term allows the infinite sum defining f_n to be replaced by a finite one with m terms, with an additional cost of $4\tau\|g\|\ \|f\| \sqrt{\frac{2\ln em + \ln 4}{m}}$.

Theorem 6. *Let the function $f(\mathbf{x}) \in \mathcal{L}_s^2(\mathbb{R}^d)$, $\frac{d}{2} < s < \frac{d}{2}+1$. Then there exists an approximation of the form $\sum_{i=1}^m c_i \phi_d(2^n\mathbf{x} - \mathbf{k}_i)$, $\mathbf{k}_i \in \mathbb{Z}^d$ so that for any fixed $r > d/4$,*

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

$$\leq 2^d \frac{2^{-(n+1)(s-d/2)}}{1-2^{(d/2-s)}} \|f\|_{\mathcal{L}_s^2} \sup_\lambda \|\psi_d^\lambda\|_{\mathcal{L}_{hom,-s}^2} + 4\tau \|g\|_{L^2} \|f\|_{L^2} \sqrt{\frac{2\ln em + \ln 4}{m}}.$$

Here,

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

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.

We remark that in the proof below we show that

$$\|g\|_{L^2}^2 = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \beta\left(\frac{d}{2}, 2r - \frac{d}{2}\right)$$

where β is the beta function.

The allowed range on s is due to the fact that $\psi_d^\lambda \in \mathcal{L}_{hom,-s}^2$ only for $\frac{d}{2} < s < \frac{d}{2}+1$ (see below - this essentially results from the limited number of vanishing moments of Haar wavelets); with other wavelets the range is larger. In addition, for $f \in \mathcal{L}_{s^*}^2$, $s^* \geq \frac{d}{2}+1$, the bound on the right side above obviously continues to hold for any choice of s satisfying $\frac{d}{2} < s < \frac{d}{2}+1$, since $\mathcal{L}_{s^*}^2 \subset \mathcal{L}_s^2$.

Proof: Let V_n denote the scaling space spanned by $\{\phi_d(2^d\mathbf{x} - \mathbf{k})\}_{\mathbf{k}\in\mathbb{Z}^d}$. First note that by [11], proof of Theorem 2.2.4 for $f \in \mathcal{L}_s^2$, $s > \frac{d}{2}$, there exists $f_n \in V_n$ such that

$$\|f - f_n\|_\infty \leq 2^d \frac{2^{-(n+1)(s-\frac{d}{2})}}{1-2^{(d/2-s)}} \|f\|_{\mathcal{L}_s^2} \sup_\lambda \|\psi_d^\lambda\|_{\mathcal{L}_{hom,-s}^2}. \tag{22}$$

Since ψ_d^λ is a Haar wavelet and so its translates all have disjoint supports. The constant A_λ appearing in the above-mentioned proof in [11] has value

$$A_\lambda \equiv \sup_{\mathbf{x}} \sum_{\mathbf{k}\in\mathbb{Z}^d} |\psi_d^\lambda(\mathbf{x} - \mathbf{k})| = 1$$

so that $\sum_\lambda A_\lambda = 2^d - 1 < 2^d$, since the cardinality of λ in the sum (the number of different wavelets) is $2^d - 1$. This justifies the constant 2^d on the right of (22).

The Fourier transform of a one dimensional Haar wavelet is (for $\omega \in \mathbb{R}$)

$$\widehat{\psi}(\omega) = -\frac{2i}{\pi\omega} e^{i\pi\omega} \sin^2\left(\frac{\pi\omega}{2}\right) = O(|\omega|)(|\omega| \to\ 0),$$

while the transform of the one dimensional scaling function is

$$\widehat{\phi}(\omega) = \frac{1}{\pi\omega} e^{i\pi\omega} \sin(\pi\omega) = O(1)(|\omega| \to 0).$$

Therefore, since the Fourier transform $\widehat{\psi}_d^\lambda(\omega)$ of a $d-$dimensional wavelet is a product of at least a single one dimensional copy of $\widehat{\psi}(\omega)$ and at most $d-1$ one dimensional copies of $\widehat{\phi}(\omega)$, it follows that for all λ (note now $\omega \in \mathbb{R}^d$)

$$\widehat{\psi}_d^\lambda(\omega) = O(|\omega|) \quad (|\omega| \to 0).$$

Thus $\psi_d^\lambda \in \mathcal{L}^2_{hom,-s}$ for $s < \frac{d}{2} + 1$, leading to the bound on s in the statement of the theorem.

Note that (22) clearly still holds when the $\|\cdot\|_\infty$ on the left is replaced by the weighted $\|\cdot\|_{L^\infty,(1+|\mathbf{x}|^2)^{-r/2}}$ norm. Thus it suffices to show that for $f_n \in V_n$, there exist $\{c_i\}_{i=1}^m \subset \mathbb{R}$ of the form above, $\{\mathbf{t}_i\}_{i=1}^m \subset \mathbb{R}^d$, and multi-integers $\{\mathbf{k}_i\}_{i=1}^m \subset \mathbb{Z}^d$, such that

$$\left\| f_n - \sum_{i=1}^m c_i \phi_d(2^n\mathbf{x} - \mathbf{k}_i) \right\|_{L_\infty,(1+|\mathbf{x}|^2)^{-r/2}} \leq 2\tau \|f_n\|_{L^2} \|(1+|\mathbf{t}|^2)^{-r}\|_{L^2} \sqrt{\frac{2\ln em + \ln 4}{m}}. \tag{23}$$

This estimate gives an L_∞ approximation bound using a finite number of scaling function translates (note also that $\|f_n\|_{L^2} \leq \|f\|_{L^2}$).

Define the kernel

$$G(\mathbf{x},\mathbf{t}) = 2^{nd} \sum_{\mathbf{k}\in\mathbb{Z}^d} \phi_d(2^n\mathbf{x} - \mathbf{k})\phi_d(2^n\mathbf{t} - \mathbf{k}),$$

which is a reproducing kernel for V_n since $\{2^{nd/2}\phi(2^n\mathbf{x}-\mathbf{k})\}_{\mathbf{k}\in\mathbb{Z}^d}$ is an orthonormal basis for V_n. Thus for $f_n \in V_n$

$$f_n(\mathbf{x}) = \int_{R^d} G(\mathbf{x},\mathbf{t}) f_n(\mathbf{x}) d\mathbf{t}.$$

Note that for fixed $\mathbf{x}$,

$$G(\mathbf{x},\mathbf{t}) = 2^{nd}\phi_d(2^n\mathbf{t} - \mathbf{k}), \tag{24}$$

where $\mathbf{k} \in \mathbb{Z}^d$ is uniquely determined by $\phi_d(2^n\mathbf{x} - \mathbf{k}) \neq 0$.

The VC dimension of the family

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

(in the parameter $\mathbf{k}$ and the variable $\mathbf{t}$) is $h = 2$. Indeed, if $\{\mathbf{t}_i\}_{i=1}^3$ are three points in $\mathbb{R}^d$, then for them to be shattered by $\mathcal{F}$ requires that they be in different dyadic cubes of order 2^{-n}. In this case, however, there is no element Φ of the family

$$\{\ 2^{nd}\phi_d(2^n\mathbf{t} - \mathbf{k}) - \alpha\}\ _{\mathbf{k}\in\mathbb{Z}^d,\alpha\in\mathbb{R}}$$

such that for two of the points say $\mathbf{t}_1, \mathbf{t}_2$,

$$\Phi(\mathbf{t}_1), \Phi(\mathbf{t}_2) > 0,$$

while for the third point $\Phi(\mathbf{t}_3) \leq 0$. This shows the VC dimension of $\mathcal{F}$ is 2.

Therefore by Theorem 2 with $k(\mathbf{x}) = (1+|\mathbf{x}|^2)^r$ and $g(\mathbf{t}) = (1+|\mathbf{t}|^2)^{-r}$, for each $j > 1$ there exist $\{d_{ij}\}_{i=1}^m$ (with $d_{ij} = \pm 1$), $\{\mathbf{t}_{ij}\}_{i=1}^m$ with $\mathbf{t}_{ij} \in \mathbb{Z}^d$ such that

$$\left\| f_n(x) - \frac{1}{m}\sum_{i=1}^m d_{ij} \left\| f_n(1+|\mathbf{t}|^2)^{-r} \right\|_{L^2} \frac{G(\mathbf{x},\mathbf{t}_{ij})}{(1+|\mathbf{t}_{ij}|^2)^{-r}} \right\|_{L^\infty,(1+|\mathbf{x}|^2)^{-r}}$$

$$\leq 4\tau' \|f_n\|_{L^2} \|(1+|\mathbf{t}|^2)^{-r}\|_{L^2} \sqrt{\frac{2\ln em + \ln 4}{m}} + \frac{1}{j} \tag{25}$$

where

$$\tau' = \sup_{\mathbf{x},\mathbf{t}} \left[G(\mathbf{x},\mathbf{t}) \frac{(1+|\mathbf{t}|^2)^r}{(1+|\mathbf{x}|^2)^r} \right],$$

and f_n the projection of f onto V_n.

We can replace $1/j$ on the right side above using an argument similar to that at the end of the proof of Corollary 3. Namely, if the sequence $\{T_j\} = \{\mathbf{t}_{ij}\}_{i=1}^m$ is bounded, we take a convergent subsequence (converging to some $\mathbf{t}_i$), and by taking a sub-subsequence, we may assume that $\{d_{ij}\}$ all have limits (of ± 1) as $j \to \infty$. In this case, it is easy to show that by taking a further subsequence, the functions $G(\mathbf{x}, \mathbf{t}_{ij})$ converge in $L^\infty[\mathbf{x}]$ (in fact from their definition they can be chosen to be unchanging as functions of $\mathbf{x}$ for sufficiently large j).

On the other hand if $\{T_j\}$ is unbounded, we can as before eliminate the terms i in the sum for which $\{\mathbf{t}_{ij}\}_{i=1}^\infty$ converges to ∞. Thus as at the end of the proof of Corollary 3, there exists a $k \leq m$, $\{d_i\}_{i=1}^k$ (with $d_i = \pm 1$) and $\{\mathbf{t}_i\}_{i=1}^k$ such that

$$\left\| f_n(x) - \frac{1}{k} \sum_{i=1}^k d_i \left\| f_n (1+|\mathbf{t}|^2)^{-r} \right\|_{L^2} \frac{G(\mathbf{x},\mathbf{t}_i)}{(1+|\mathbf{t}_i|^2)^{-r}} \right\|_{L^\infty, (1+|\mathbf{x}|^2)^{-r}}$$

$$\leq 4\tau' \|f_n\|_{L^2} \|(1+|\mathbf{t}|^2)^{-r}\|_{L^2} \sqrt{\frac{2\ln em + \ln 4}{m}} \tag{26}$$

where τ' is defined above. Note that $k = m$ if no terms have been eliminated, and otherwise $k < m$.

Define

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

with $+$ or $-$ in front according to the sign of $f_n(\mathbf{t}_i)$. To verify (23) above, we will now show $\tau' \leq \tau = 2^{nd}(1+2^{-2n}d)^r$.

In any dyadic cube C of side 2^{-n}, we have that in $C \times C$, $G(\mathbf{x},\mathbf{t}) = 2^{nd}$, while

$$\frac{(1+|\mathbf{t}|^2)}{(1+|\mathbf{x}|^2)} \leq 1 + d2^{-2n}.$$

The latter follows from the fact that in $C \times C$, the ratio above is maximized if this cube has one corner at the origin. In that case the numerator is largest when $|\mathbf{t}|$ is largest, i.e., equal to the length of the diagonal of this cube, which gives $1 + |\mathbf{t}|^2 = 1 + d2^{-2n}$, and when $\mathbf{x} = \mathbf{0}$.

Thus since $G(\mathbf{x},\mathbf{t}) \leq 2^{nd}$

$$\tau' \leq \tau \equiv 2^{nd}(1+d2^{-2n})^r.$$

Choosing $\mathbf{k}_i$ so that for all $\mathbf{x}$, $G(\mathbf{x},\mathbf{t}_i) = 2^{nd}\phi(2^n\mathbf{x} - \mathbf{k}_i)$, we see that (26) above implies (23), completing the proof.

We note that the above norm

$$\|(1+|\mathbf{t}|^2)^{-r}\|_{L^2}^2 = \int_{\mathbb{R}^d} (1+|\mathbf{t}|^2)^{-2r} d\mathbf{t} = 2\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty (1+\rho^2)^{-2r} \rho^{d-1} d\rho$$

$$= \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty (1+u)^{-2r} u^{d/2-1} du = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \beta\left(\frac{d}{2}, 2r - \frac{d}{2}\right)$$

where $\beta(\mu,\nu) = \int_0^1 x^{\nu-1}(1-x)^{\mu-1}dx$ denotes the Beta function [10], section 3.194. ■

5. Conclusion

There are technical difficulties in establishing analogues of Theorem 2 for probabilistic error bounds using V_γ dimension (a generalization of VC dimension) and covering numbers. Of the three function space capacity measures consisting of VC dimension, V_γ dimension and covering numbers, the VC dimension is most difficult to calculate. Moreover there is an example of an infinite dimensional reproducing kernel Hilbert space (RKHS) with infinite VC dimension, but finite bounds for its V_γ dimension [6].

When the loss function is the least squares error there are sharp bounds for V_γ dimension [1, 6] and for covering numbers [4, 25, 26, 32, 33]. We remark that bounds for covering numbers [4] are useful in the stability property approach to SLT [7, 12, 25].

We also note that the approach that Girosi has developed is derived from probabilistic methods developed by Dudley [5] and others, which are effectively formalizations of Monte Carlo methods for estimating integrals such as those for expected risk. These probabilistic theorems involve estimates which are uniform in a parameter (the function f) in the integral. When this parameter is translated into $\mathbf{x}$, we obtain uniformity in $\mathbf{x}$ for estimates of functions in reproducing kernel Hilbert spaces. This gives the advantage of Monte Carlo, with its low dimensional dependence, together with estimation of full functions and not just single parameters. This is an important benefit of the translation of the above-mentioned probabilistic results into approximation theoretic ones.

Acknowledgments: The authors thank the referee for a most helpful and careful analysis of our paper. The second author thanks F. Girosi and D. Watson for sharing insights on SLT, and R. Strichartz for providing an opportunity to spend a sabbatical at Cornell University.

References

[1] Alon, N., S. Ben-David, N. Cesa-Bianchi and D. Haussler. Scale Sensitive Dimension, Uniform Convergence, and Learnability. *JACM* 44 (4), 615-631, 1997.

[2] Barron, A.R. Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning* 14, 115 - 133, 1994.

[3] Barron, A.R. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transaction on Information Theory* 39, No. 3, 930-945, 1993.

[4] Cuker, F. and S. Smale. *On the Mathematical Foundations of Learning.* AMS Bulletin 39, No. 1, 1-49, Jan. 2002.

[5] Dudley, R.M. Balls in $\mathbb{R}^k$ Do Not Cut All Subsets of $k+2$ Points. *Advances in Mathematics* 31, 306-398, 1979.

[6] Evgeniou, T. and M. Pontil. On the V_γ Dimension for Regression in Reproducing Kernel Hilbert Spaces. *Lecture Notes in Computer Science, Algorithmic Learning Theory.* Tokyo, Japan, 1999.

[7] Evgeniou, T. M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics,* 13, 1-50, 2000.

[8] Girosi, F., Approximation Error Bounds that Use VC Bounds. In *Proc. International Conference on Artificial Neural Networks,* F. Fogelman-Soulied and P. Gallinari, Eds., 1, 295-302, Paris, Oct. 1995.

[9] Girosi, F., M. Jones and T. Poggio. Regularization Theory and Neural Networks Architectures. *Neural Computation* 7, 219-269, 1995.
[10] Gradshteyn, I. and I. Ryzhik. *Tables of Integrals, Series and Products*. 3rd ed., Academic Press, 1980.
[11] 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.
[12] Kurkova, V and M. Sanguineti. Learning with Generalization Capability by Kernel Methods of Bounded Complexity. *Journal of Complexity* to appear.
[13] Massachusetts Institute of Technology - Artificial Intelligence: http://www.ai.mit.edu/projects/cbcl/projects/NoticesAMS /PoggioSmale.html
[14] Max Planck Institute - Biological Cybernetics: http://www.kyb.tuebingen.mpg.de/bs/projects,html?pg=23
[15] Mhaskar, H. N. Neural Networks for Optimal Approximation of Smooth and Analytic Functions. *Neural Computation* 8, 164-177, 1996.
[16] Mhaskar, H. N. and C. Micchelli. Approximation by Superposition of Sigmoidal and Radial Basis Functions. *Advances in Applied Mathematics* 13, 350- 373, 1992.
[17] Mhaskar, H. N. and C. Micchelli. Dimension Independent Bounds on the Degree of Approximation by Neural Networks. *IBM, J. Res. Devel.* 38, 277-284, 1994.
[18] Micchelli, C. A. and M. Buhmann. On Radial Basis Approximation on Periodic Grids. Math. Proc. Camb. Phil. Soc. 112, 317-334, 1992.
[19] Mukherjee, S., P. Niyogi, T. Poggio, and R. Ryfkin. Statistical Learning: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization. MIT AI Memo 2002-024, MIT CBCL Memo 223, 1-54.
[20] Mukherjee, S., R. Ryfkin, T. Poggio. Regression and Classification with Regularization, MIT CBCL papers.
[21] Niyogi, P. and F. Girosi. On the Realization between Generalization Errors, Hypothesis Complexity, and Sample Complexity for Radial Basis Functions. *Neural Computation* 8, 819-842, 1996.
[22] Niyogi, P. and F. Girosi. Generalization Bounds for Function Approximation from Scattered Noisy Data. *Advances in Computational Mathematics* 10, 51-80, 1999.
[23] Park, J. and I Sandberg. Approximation and Radial-Basis Function Networks. *Neural Computation* 5, 305-316, 1993.
[24] Poggio, T and F. Girosi. Regularization Algorithms for Learning that are Equivalent to Multilayer Networks. *Science* 247, 978-982, 1990.
[25] Poggio, T. and S. Smale. The Mathematics of Learning: Dealing with Data. *Notices AMS*, 50, No. 5, 537-544, May 2002.
[26] Smale, S. and D-X. Zhou. Estimating the Approximation Error in Learning Theory. *Analysis and Applications* 1, No. 1, 1-25, 2003.
[27] Sontag, E. VC Dimension of Neural Networks. In *Neural Networks and Machine Learning*, C.M. Bishop, Ed., Springer-Verlag, Berlin, 1998, 69-95.
[28] Stein, E.M. *Singular Integrals and Differentiability Properties of Functions.* Princeton University, 1970.
[29] Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed., Springer, 2000.
[30] Vapnik, V.N. and A. Ja. Chervonenkis. The Necessary and Sufficient Conditions for Consistency in the Empirical Risk Minimization Method. *Pattern Recognition and Image Analysis* 3, 283-305, 1991.
[31] Walter, G, *Wavelets and Other Orthogonal Systems with Applications*, CRC Press, 1994.
[32] Zhou, D-X. The Covering Number in Learning Theory. *J. Complexity* 18, 738-767, 2002.
[33] Zhou, D-X. Capacity of Reproducing Kernel Spaces in Learning Theory, *IEEE Trans. Inform. Theory* 49, No. 7, 1743-1782, 2003.

(Kon) Department of Mathematics, Boston University, Boston, MA, 02215 USA
E-mail address: mkon@math.bu.edu

(Raphael and Williams) Department of Mathematics, Howard University, Washington, DC 20059 USA
E-mail address: lraphael@howard.edu dawilliams@howard.edu