International Journal of Pure and Applied Mathematics
Volume 47 No. 4 2008, 459-473

STATISTICAL LEARNING METHODS FOR UNIFORM APPROXIMATION BOUNDS IN MULTIRESOLUTION SPACES

Mark A. Kon[1], Louise A. Raphael[2] §

[1]Department of Mathematics and Statistics
Boston University
Boston, MA 02215, USA
e-mail: mkon@math.bu.edu

[2]Department of Mathematics
Howard University
Washington, DC 20059, USA
e-mail: lraphael@howard.edu

Abstract: New constructive and non-constructive non-asymptotic uniform error bounds for approximating functions in $\mathcal{L}^2_s(\mathbb{R}^d)$, $d \geq 1$, by finite compactly supported multiresolution expansions are proved using approximation theoretic bounds derived from statistical learning theory.

AMS Subject Classification: 41A25, 41A65, 68T05
Key Words: statistical learning theory (SLT), VC dimension, multiresolution analysis (MRA), wavelets, reproducing kernel Hilbert space (RKHS)

1. Introduction

Given a function f_0 in an L^2 Sobolev space $\mathcal{L}^2_s(\mathbb{R}^d)$, $s > 0$, $d \geq 1$, we will examine here both constructive and non-constructive weighted L^∞ approximations of f_0 using finite combinations of compactly supported scaling functions or, equivalently, compactly supported wavelets.

We follow an approach which parallels statistical learning theory (SLT) methods first developed by Girosi [5]. Our non-constructive result is an extension of one appearing in [11]. The constructive results are probabilistic, and

Received: July 24, 2008

§Correspondence author

approximate f_0 in a multiresolution analysis (scaling) space V_n using a finite number of terms, whose cardinality is controlled by a desired approximation error, the probability η of such error, and the VC dimension of the scaling space V_n.

Let $\{V_n\}_{n=-\infty}^{\infty}$ be a multiresolution analysis (MRA) on $\mathbb{R}^d$ with a bounded compactly supported scaling function $\phi(\mathbf{x}) \in L^2(\mathbb{R}^d)$ and (consequently) with an associated family of bounded compactly supported wavelets $\{\psi^\lambda(\mathbf{x})\}_{\lambda\in\Lambda}$ (see [4]). The error of our L^∞ approximation $\widetilde{f}$ of f_0 has two components. The first is a bound on approximation error, i.e., the distance $d(f_0, V_n)$ between f_0 and its closest approximation f_n (assuming it exists) in the hypothesis space V_n, based on L^∞ wavelet approximation results (see [7], [8], [9]). The second is based on an analogue of statistical learning estimation error (the *estimation error*) i.e., the distance between f_n and its SLT-based approximation $\widetilde{f}$, given by a finite combination of translates of the reproducing kernel for (the projection onto) V_n (see Section 5).

In our main result, the approximation $\widetilde{f}$ is effectively obtained from sampling a probability distribution derived from f_n, which yields non-asymptotic weighted L^∞ error bounds. Specifically, in controlling the estimation error one has a weighted L^∞ approximation $\widetilde{f}$ of f_0, if f is in an L^p Sobolev space ($p > 1$), see [11], the approximation is in fact unweighted when $p = 1$. We also show existence of analogous non-constructive weighted L^∞ approximations which, with the same number of translations of the scaling function, give better approximations.

Girosi [5] first applied SLT to obtain approximation theoretic bounds in L^1. Generalizations of Girosi's result to L^p ($p \geq 1$) Sobolev spaces are found in [11]. These results are extended to a reproducing kernel Hilbert space (RKHS) L^2 setting in [10]. The bounds there depend only on the complexity of the reproducing kernel, and the number of data points. Specifically, for functions in RKHS the results give weighted sup norm non-asymptotic bounds.

The present multiresolution expansion results are obtained by considering reproducing kernels $K(\mathbf{x}, \mathbf{y})$ which are L^2 projections onto the closed MRA spaces V_n. Equivalently, the RKHS inner product is the L^2 inner product restricted to V_n.

We note that RKHS, [1] as hypothesis spaces are often spaces of choice in statistical learning theory (e.g., Cucker and Smale [3], Smale and Zhou [15], [16], Poggio and Smale [12], Vapnik [18], Zhou [20]), and applications such as support vector machines (e.g. Schölkopf and Smola [13], Shawe-Taylor and Chistianini [14]).

We remark that purely approximation theoretic results are important in SLT to the extent that they represent a portion of the full generalization error in approximating a function f from partial information. See for example Smale and Zhou [15], who obtained L^2 approximation error bounds for (quite different) hypothesis spaces of compact Sobolev balls (and more generally reproducing kernel Hilbert balls) in $L^2(\mathbb{R}^d)$.

In contrast to the approximation error results of [15], we bound our full error (known in SLT as *generalization error*) by a sum of approximation and estimation errors, in the context of a set of nested hypothesis spaces consisting of increasing multiresolution spaces V_n. The L^∞ bounds in the approximation error component of this analysis are shown in Kon and Raphael [8], [9] to be optimal, in the sense that the exponent characterizing the convergence rate cannot be improved under the given hypotheses. Similarly, the L^∞ bounds on the estimation error component are *close* to optimal (within a logarithm) when they are stated for L^1 Sobolev functions (see [5]) and hence by adaptation [11] for L^2 Sobolev functions, given a choice of weight functions (see [5]). Therefore, at least for this method of obtaining bounds through a partition into estimation and approximation error, these bounds are "logarithmically close" to the best possible of their type for a choice of weight functions.

In Section 2 we review some basic facts on VC dimension and give references to SLT background. In Section 3 we cite our generalization of Girosi's result for our reproducing kernel Hilbert space. We define some basic wavelet concepts in Section 4. Section 5 contains a new proposition on VC dimension bounds of a family of scaled projection kernels associated with an MRA, as well as our main result. The latter involves constructive uniform non-asymptotic error bounds for approximation by finite combinations of bounded compactly supported scaling functions.

2. VC Dimension

Girosi was the first to establish a probabilistic connection between SLT and approximation theory, showing that SLT methods can be used to prove results of a purely approximation theoretic nature. In [11] we extended this probabilistically-based bound for L^1 Sobolev functions to an analogous result for L^p Sobolev spaces for $1 < p \leq \infty$ with smoothness $s > 0$. In [10] we showed these results true in a reproducing kernel Hilbert space setting.

Here we use the main results of [11], [10] to derive a probabilistic VC

dimension-based bound on the error of approximating $f \in L^2$ by a finite sum of translates of a bounded compact scaling function or equivalently, the wavelets associated with this scaling function. This uses the VC dimension (below) of reproducing (projection) kernels of the scaling spaces V_n of our multiresolution expansion. We start by giving a general definition of VC dimension.

Definition 1. The VC dimension of a family F of functions on a space X is the maximum number h of points $\{\mathbf{t}_i\}_{i=1}^h$ which can be shattered, i.e., separated into two classes in all possible ways, using classes of the form $\{t : f(t) - \alpha \geq 0\}_{f \in F, \alpha \in \mathbb{R}}$. More specifically we require that there exist a subset $S \subset X$ of cardinality $|S| = h$, and an $\alpha \in \mathbb{R}$ such that $f(\mathbf{t}) - \alpha \geq 0$ iff $\mathbf{t} \in S$. We also require that h be the largest cardinality of such a set S.

Given a kernel $K(\mathbf{x}, \mathbf{t})$ its VC dimension is the VC dimension of the family $\{K(\mathbf{x}, \mathbf{t})\}_{\mathbf{x} \in \mathbb{R}^d}$ in the variable $\mathbf{t}$ and parameter $\mathbf{x}$.

Examples of function classes of VC dimension $d + 1$ include [5]:

— characteristic functions of half-spaces on R^d;

— characteristic functions of balls on R^d;

— equivariant Gaussian kernels on R^d, see [5].

We will use the next theorem in our main results of Section 5 to bound the VC dimension of the projection kernel onto a multiresolution space.

Theorem 2. (Sontag, see [17]) *If X is a set and F is a finite dimensional vector space of functions $f : X \to \mathbb{R}$ (which does not contain the constant function 1), then the VC dimension of F equals dim $F + 1$, with dim F the dimension of F.*

We note that this statement differs from that of the theorem in [17], since the definition of VC dimension there does not include the parameter α. If the class F were to contain 1, then inclusion of α would not change the VC dimension, since multiples of 1 would already play this role. In the latter case the VC dimension according to the theorem would be $\dim F$.

We refer the reader to the seminal theorem of Vapnik and Chervonenkis known as the VC Bound Theorem [18], which gives probabilistic estimates of integrals by finite sums. This is a main tool in [5], [10], [11], and is used in the following section.

3. Generalization of Girosi's Result and RKHS

An important consequence of the VC bound theorem [18] is that the uniform deviation between an integral and finite sum approximation can be bounded in terms of VC dimension. Girosi [5] used this to estimate integrals of the form

$$\int J(\mathbf{x},\mathbf{t})\rho(\mathbf{t})d\mathbf{t} \tag{1}$$

in L^∞ using quadrature-based sums

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(\rho(\mathbf{t}_i))J(\mathbf{x},\mathbf{t}_i)\|\rho\|_{L^1} \tag{2}$$

with $\rho(\mathbf{t}) \in L^1(\mathbb{R}^d)$.

For completeness we cite our version [11] of Girosi's [5] original result.

Proposition 3. *Let $f(\mathbf{x})$ be represented as an integral in the form (1), with $\rho \in L^1(\mathbb{R}^d)$. If the kernel J satisfies $A \le J(\mathbf{x},\mathbf{t}) \le B$ for $\mathbf{x},\mathbf{t} \in \mathbf{R}^d$, 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 $\frac{|\rho(\mathbf{x})|d\mathbf{x}}{\|\rho\|_{L^1}}$:*

$$\left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(\rho(\mathbf{t}_i))J(\mathbf{x},\mathbf{t}_i)\|\rho\|_{L^1}\right\|_{L^\infty} \le 4\tau\|\rho\|_{L^1}\sqrt{\frac{h\ln\frac{2en}{h} - \ln\frac{\eta}{4}}{n}}, \tag{3}$$

where $\tau = B - A$ and h is the VC dimension of J.

Letting $\eta \to 1$, Proposition 3 leads to the following non-constructive corollary, see [11].

Corollary 4. *Under the assumptions of Proposition 3, for every $\epsilon > 0$ there exists a sample $\{\mathbf{t}_i\}_{i=1}^n \in \mathbb{R}^d$ such that*

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

We can improve the result in this corollary with some additional assumptions to let $\epsilon = 0$.

Corollary 5. *If $J(\mathbf{x},\mathbf{t}) \to 0$ as $x \to \infty$, the inequality in the above corollary holds when $\epsilon = 0$, with the possible replacement of the set $\{\mathbf{t}_i\}_{i=1}^n$ by a set of smaller cardinality (with no change in n on the right side).*

Proof. Since ϵ can be made arbitrarily small, we let $\{\mathbf{t}_{ki}\}_{i=1}^n$ be a sequence of samples for $k = 1, 2, \ldots$ for which the error $\epsilon = \epsilon_k \to 0$ as $k \to \infty$. The

vector $\mathbf{T}_k = (\mathbf{t}_{k1}, \ldots, \mathbf{t}_{kn})$ (i.e., the concatenation of $\{\mathbf{t}_{ki}\}_{i=1}^n$) must have a convergent subsequence in the one point compactification $\overline{\mathbb{R}}$ of $\mathbb{R}$ (thus allowing some components to converge to ∞).

We pass to this subsequence, and denote the limit for each i by

$$\mathbf{t}_i = \lim_{k\to\infty} \mathbf{t}_{ki};$$

note that some $\mathbf{t}_i$ may equal infinity. Without loss of generality, we may by taking a further subsequence assume that $\operatorname{sgn}(\rho(\mathbf{t}_{ki}))$ converges for each i. Finally, we redefine the density function ρ on (at most a finite number of points) $\mathbf{t}_i$ so that

$$\operatorname{sgn}(\rho(\mathbf{t}_i)) = \lim_{k\to\infty} \operatorname{sgn}(\rho(\mathbf{t}_{ki})).$$

Note for any convergent sequence of functions $a_i(x)$ on a measure space,

$$\limsup_{i\to\infty} \|a_i\|_\infty \geq \|\lim_{i\to\infty} a_i\|_\infty = L.$$

Indeed, for any $\beta > 0$ there are arbitrarily large i for which $|a_i(x)| \geq L - \beta$ on a set of positive measure.

Thus defining $J(\mathbf{x}, \infty) = 0$, it follows easily that

$$\left\| f(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\rho(\mathbf{t}_i)) J(\mathbf{x}, \mathbf{t}_i) \|\rho\|_{L^1} \right\|_{L^\infty}$$
$$\leq \lim_{k\to\infty} \left[4\tau\|\rho\|_{L^1} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}} + \epsilon_k \right] = 4\tau\|\rho\|_{L^1} \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}}.$$

Note in particular that the sum $\sum_{i=1}^n \operatorname{sgn}(\rho(\mathbf{t}_{ki})) J(\mathbf{x}, \mathbf{t}_{ki}) \|\rho\|_{L^1}$ converges almost everywhere in $\mathbf{x}$ as $k \to \infty$. In addition, if some $\mathbf{t}_i = \infty$ (and hence $J(\mathbf{x}, \mathbf{t}_i) = 0$)), then we are left with a lower order sum, i.e., the corresponding terms in the sum vanish. The effective sample $\{\mathbf{t}_i\}_{i=1}^n$ is now smaller, with its size n replaced by $m < n$ in the left sum only, proving the result. □

In [11] we have generalized Girosi's result in two different ways to $\mathcal{L}^p(\mathbb{R}^d)$ for $p > 0$. In [10] we have also extended it to reproducing kernel Hilbert spaces. Here we use the RKHS setting and apply it to compact scaling function multiresolution expansions.

Definition 6. On a measure space X we define a Hilbert space $\mathcal{H}$ of real functions to be a reproducing kernel Hilbert space if for each $\mathbf{x} \in X$, the map $f \to f(\mathbf{x})$ is a continuous linear functional on $\mathcal{H}$.

By the Riesz Representation Theorem, for any real reproducing kernel

Hilbert space there exists a symmetric real function $K(\mathbf{x}, \mathbf{y})$ such that for $f \in \mathcal{H}$, $\mathbf{x} \in X$,

$$f(\mathbf{x}) = \langle (K(\mathbf{x}, \cdot), f(\cdot) \rangle .$$

Definition 7. The weighted L^∞ norm of a function f with respect to a weight function $a(\mathbf{x})$ is

$$\|f\|_{L^\infty, a(\mathbf{x})} = \text{ess sup}_{\mathbf{x}} |f(\mathbf{x}) a(\mathbf{x})|. \tag{5}$$

The following proposition follows from results in [10].

Proposition 8. *Assume that on $\mathbb{R}^d$ a reproducing kernel Hilbert space $\mathcal{H}_K$ is a subspace of $L^2(\mathbb{R}^d)$ and inherits the L^2 inner product, with a bounded, symmetric, and positive semi-definite reproducing kernel $K(\mathbf{x}, \mathbf{y})$. 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*

$$\text{ess sup}_{\mathbf{x}, \mathbf{t}} \left| \frac{K(\mathbf{x}, \mathbf{t})}{g(\mathbf{t}) k(\mathbf{x})} \right| \leq \tau. \tag{6}$$

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 $\eta > 0$ the weighted L^∞ norm

$$\left\| f(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^{n} \text{sgn} f(\mathbf{t}_i) \|fg\|_{L^1} \frac{K(\mathbf{x}, \mathbf{t}_i)}{g(\mathbf{t}_i)} \right\|_{L^\infty, 1/k(\mathbf{x})}$$
$$\leq 4\tau \|f\|_{L^2} \|g\|_{L^2} \sqrt{\frac{h \ln \frac{2en}{h} - \ln \frac{\eta}{4}}{n}}, \tag{7}$$

with probability greater than $1 - \eta$, if $\mathbf{t}_i$ are chosen independently with respect to the density $\frac{|fg|}{\|fg\|_1}$.

4. MRA and Wavelet Background

A multiresolution analysis (MRA) is defined as an increasing nested sequence of closed subspaces $\{V_n\}_{n=-\infty}^{\infty}$ of $L^2(\mathbb{R}^d)$ $(d \geq 1)$ such that

$$f(\mathbf{x}) \in V_n \text{ iff } f(2\mathbf{x}) \in V_{n+1},$$

with an intersection $\bigcap_{n=-\infty}^{\infty} V_n = \{0\}$, a union $\bigcup_{n=-\infty}^{\infty} V_n$ which is dense in $L^2(R^d)$, and with V_0 invariant under integer translations. It is also generally assumed that there exists a *scaling function* $\phi(\mathbf{x})$ whose integer translates form an orthonormal basis for V_0. For detailed definitions and theory of an MRA we

refer to Chapter 10 in [4] and [2], [6], [19].

Let W_n be the orthogonal complement of V_n in V_{n+1}, i.e., $W_n = V_{n+1} \ominus V_n$, so that $V_{n+1} = V_n \oplus W_n$. From existence of ϕ it follows [4] that there is a set of basic wavelets $\{\psi^\lambda(\mathbf{x})\}_{\lambda\in\Lambda}$ (with Λ a finite index set) such that $\psi^\lambda_{j\mathbf{k}}(\mathbf{x}) \equiv 2^{jd/2}\psi^\lambda(2^j\mathbf{x} - \mathbf{k})$ $(j \in \mathbb{Z}, \mathbf{k} \in \mathbb{Z}^d, \lambda \in \Lambda)$ form an orthonormal basis for W_j for fixed j, and form an orthonormal basis for $L^2(\mathbb{R}^d)$ as $\lambda, j, \mathbf{k}$ vary. Our results will hold for any wavelet set $\{\psi^\lambda\}_\lambda$ on $\mathbb{R}^d$ associated with a bounded, compactly supported scaling function, regardless of how they are constructed, see [4].

For $f(\mathbf{x})$ a function on $\mathbb{R}^d$, let $\hat{f}(\xi)$ denote its Fourier transform.

Definition 9. Define the *Sobolev space* $\mathcal{L}^2_s(\mathbb{R}^d)$, $s \in \mathbb{R}$ by

$$\mathcal{L}^2_s(\mathbb{R}^d) \equiv \left\{ f \in L^2(\mathbb{R}^d) : \|f\|_{\mathcal{L}^2_s} \equiv \sqrt{\int |\widehat{f}(\xi)|^2(1+|\xi|^2)^s d\xi} < \infty \right\}.$$

The *homogeneous Sobolev space* is

$$\mathcal{L}^2_{\mathrm{hom},s}(\mathbb{R}^d) \equiv \left\{ f \in L^2(\mathbb{R}^d) : \|f\|_{\mathcal{L}^2_{\mathrm{hom},s}} \equiv \sqrt{\int |\widehat{f}(\xi)|^2|\xi|^{2s} d\xi} < \infty \right\}.$$

Note the spaces in fact contain the same functions (by virtue of the fact that $\mathcal{L}^2_{\mathrm{hom},s}$ is restricted to L^2). However, the norms on the two spaces differ, and the second space is incomplete as defined (its completion contains non-L^2 functions which grow at ∞).

5. Main Result

We now combine approximation and estimation errors to obtain sup norm error bounds for approximations of Sobolev functions using multiresolution expansions with finite numbers of terms.

Our next proposition, which is of independent interest, bounds the VC dimension of a family of projection operators associated with an MRA for a compactly supported scaling function. It can be used to bound the VC dimension h in Definition 7. In the proposition, the scaling function ϕ is compactly supported, and the projection P_n onto V_n has kernel

$$P_n(\mathbf{x}, \mathbf{t}) \equiv 2^{nd} \sum_{\mathbf{k}\in\mathbb{Z}^d} \phi(2^n\mathbf{x} - \mathbf{k})\phi(2^n\mathbf{t} - \mathbf{k}).$$

Proposition 10. *Let ϕ be a compact scaling function associated with a real multiresolution analysis V_n, with P_n the projection onto V_n with kernel*

$P_n(\mathbf{x},\mathbf{t})$. Let $k(\mathbf{x})$, $g(\mathbf{t})$ be positive functions. Then the VC dimension h of the family

$$K_n(\mathbf{x},\mathbf{t}) = \frac{P_n(\mathbf{x},\mathbf{t})}{k(\mathbf{x})g(\mathbf{t})}$$

in the variable $\mathbf{t}$ and parameter $\mathbf{x}$ satisfies

$$h \leq l^2(l+1)+1\,,$$

where l is the maximum number of translates $\phi(\mathbf{x}-\mathbf{k})$ (with $\mathbf{k}$ a multi-integer) whose support in $\mathbf{x}$ intersects the open unit cube $C=[0,1]^d$.

Proof. The VC dimension h of the kernel $K_n(\mathbf{x},\mathbf{t})$ is the maximum number of points $\mathbf{t}_i$ that can be separated by the family of functions $\{K_n(\mathbf{x},\mathbf{t})-\alpha\}_{\alpha\in\mathbb{R}}$, with $\mathbf{x}$ and α as function family parameters. We assume henceforth that $n=0$ without loss of generality and let $K_0=K$. Given a set H of h points in $\mathbb{R}^d$, we note that in order to shatter H in the variable $\mathbf{t}$ with functions in the class $\mathbf{\Phi}=\{K(\mathbf{x},\mathbf{t})-\alpha : \mathbf{x}\in\mathbb{R}^d, \alpha\in\mathbb{R}\}$, there may be at most $l+1$ points in H contained in any given unit cube $C-\mathbf{k}$. This follows from the fact that the dimensionality of $\mathbf{\Phi}$ (in the variable $\mathbf{t}$) in such a cube is at most $l+1$ (counting the degree of freedom from the parameter α), bounding the VC dimension of $\mathbf{\Phi}$ restricted to such a cube (see Theorem 2 above). Note we are again incorporating the constant α in the dimensionality of Φ in using Theorem 2.

On the other hand, notice that for fixed $\mathbf{x}$, the above sum in $\mathbf{k}$ has at most l non-zero terms (since for such $\mathbf{x}$, $\phi(\mathbf{x}-\mathbf{k})$ is non-zero for at most l values of $\mathbf{k}$). Note that the support in $\mathbf{t}$ of $\sum_k \phi(\mathbf{x}-\mathbf{k})\phi(\mathbf{t}-\mathbf{k})$ is also bounded by the fact that each term is supported in $\mathbf{t}$ on at most l integer translates of C. Thus for fixed $\mathbf{x}$ the full sum $P_0(\mathbf{x},\mathbf{t})=\sum_k \phi(\mathbf{x}-\mathbf{k})\phi(\mathbf{t}-\mathbf{k})$ is supported in $\mathbf{t}$ on at most l^2 translates of C by integer vectors.

Finally note if H is shattered by Φ, then the maximum number of different translates $C-\mathbf{k}$ of the cube C which intersect with H can be bounded by the following fact. First the support (in $\mathbf{t}$) of $\sum_k \phi(\mathbf{x}-\mathbf{k})\phi(\mathbf{t}-\mathbf{k})$ contains at least $h-1$ elements of H for at least one $\mathbf{x}$. Indeed, otherwise there would be no member of Φ which would separate a single member of H from the others. These $h-1$ points of H must thus be contained within l^2 integer translates of C. As shown above, restricted to any fixed translate of C, the dimensionality of Φ is $l+1$. Thus the total dimensionality of Φ restricted to the above l^2 integer translates of C is $l^2(l+1)$. Based on Theorem 2, the most points which can be shattered within this set is $l^2(l+1)$. Thus the bound

$$h = \mathrm{VC}(\Phi) \leq l^2(l+1)+1$$

on the VC dimension of the family Φ is based on Theorem 2. □

We now give our main result allowing a constructive probabilistic sup-norm approximation of a function in V_n with a finite number of terms. We note that the number of terms will be controlled by desired approximation error, its probability η and the VC dimension of the family of kernels.

As noted earlier the error of our approximation of f_0 has two components. The first involves the classical approximation error in the MRA, i.e., the distance $\mathbf{AE} = d(f_0, V_n)$ between f_0 and its closest (projected) approximation f_n in V_n, whose estimates will be based on L^∞ wavelet error bounds (see [7], [8], [9]). The second is based on estimation error **EE**, i.e., the distance between f_n and its SLT-based approximation as a finite combination of translates of the reproducing kernel for V_n. We define

$$A_\lambda = \sup_{\mathbf{t}} \sum_{\mathbf{k}\in\mathbb{Z}^\mathbf{d}} |\psi^\lambda(\mathbf{t}-\mathbf{k})|;$$

note below τ is defined as in (6). Recall that as above, l is defined as the maximum number of multi-integer translates $\phi(\mathbf{x}-\mathbf{k})$ whose support in the variable $\mathbf{x}$ intersects the open unit cube $C = (0,1)^d$.

Theorem 11. *Let $\{V_n\}_n$ be a multiresolution analysis on $\mathbb{R}^d$ generated by a bounded compactly supported scaling function ϕ with a corresponding family of (bounded and compactly supported) wavelets $\{\psi^\lambda\}_\lambda \in \mathcal{L}^2_{hom,-s}$, with P_n the orthogonal projection onto V_n.*

Assume that $f \in \mathcal{L}^2_s(\mathbb{R}^d)$ with $s > d/2$, and let $g(\mathbf{t})$ be any bounded continuous positive function in L^2. Then there exists a positive function $k(\mathbf{x})$ bounded away from 0 such that for all n, $\frac{P_n(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \le 2^n\tau$ for some constant $\tau > 0$.

For every integer $m > 0$ and every $0 < \eta < 1$, a scaling function (equivalently, wavelet) approximation of f

$$\widetilde{f}(\mathbf{x}) = \sum_{j=1}^{p} d_{\mathbf{k}_j}\phi(2^n\mathbf{x}-\mathbf{k}_j)$$

can be constructed, with $\{\mathbf{k}_j\}_{j=1}^p$ a finite indexed subset of $\mathbb{Z}^d$, and

$$\|f-\widetilde{f}\|_{\infty,1/k(\mathbf{x})} \le \mathbf{AE} + \mathbf{EE}$$

with probability at least $1-\eta$. Here

$$\mathbf{AE} \le \|1/k\|_\infty \|f\|_{\mathcal{L}^2_s} \|\psi\|_{\mathcal{L}^2_{\mathrm{hom},-s}} C(s,d)$$

$$\mathbf{EE} \le 4\cdot 2^{nd}\tau\|g\|_{L^2}\|f\|_{L^2}\sqrt{\frac{h\ln\left(\frac{2em}{h}\right)-\ln\left(\frac{\eta}{4}\right)}{m}}$$

and $C(s,d) = \frac{2^{-(n+1)(s-d/2)}}{1-2^{d/2-s}} \sum_\lambda A_\lambda$. *Above* h *is the (bounded) VC dimension of the kernel* $\frac{P_n(\mathbf{x},\mathbf{t})}{k(\mathbf{x})g(\mathbf{t})}$, *and* $p \leq ml$.

Remark. From Proposition 6 the VC dimension h satisfies

$$h \leq l^2(l+1)+1,$$

even though it may depend on n above.

Proof. We first construct the approximation error (**AE**) bound. Note that ψ^λ are bounded and compactly supported, which follows from the same property for ϕ, and

$$A_\lambda \leq \sup_{\mathbf{t}} |\psi^\lambda(\mathbf{t})| L_\lambda,$$

where L_λ denotes the number of integer translates $\psi_\lambda(\mathbf{x}-\mathbf{k})$ supported in the open unit cube C. Let $f_n = P_n f$. By the boundedness and compact support assumptions on ϕ, it follows that the hypotheses of (see [8], Theorem 2.2.4) are satisfied, and in particular the wavelets $\psi^\lambda(\mathbf{x})$ are bounded by an $L^1(\mathbb{R}^d)$ radial function. Thus (see [8])

$$\|f-f_n\|_\infty \leq \sum_\lambda \sum_{j=n+1}^{\infty} 2^{-j(s-d/2)} A_\lambda \|f\|_{L^2_{\mathrm{hom},s}} \|\psi^\lambda\|_{\mathcal{L}^2_{\mathrm{hom},-s}}$$
$$= \frac{2^{-(n+1)(s-d/2)}}{1-2^{d/2-s}} \sum_\lambda A_\lambda \|f\|_{\mathcal{L}^2_{hom,s}} \|\psi^\lambda\|_{\mathcal{L}^2_{\mathrm{hom},-s}}.$$

Now note that for $s>0$

$$\|f\|^2_{\mathcal{L}^2_{\mathrm{hom},s}} = \int |\,\widehat{f}(\xi)|^2 |\xi|^{2s} d\xi \leq \int |\,\widehat{f}(\xi)|^2 (1+|\xi|^2)^s d\xi = \|f\|^2_{\mathcal{L}^2_s}.$$

Thus for any $k(x)>0$

$$\mathbf{AE} \equiv \|f-f_n\|_{\infty,1/k(\mathbf{x})}$$
$$\leq \|1/k\|_\infty \frac{2^{-(n+1)(s-d/2)}}{1-2^{d/2-s}} \|f\|_{\mathcal{L}^2_s} \sup_\lambda \|\psi^\lambda\|_{\mathcal{L}^2_{\mathrm{hom},-s}} \sum_\lambda A_\lambda. \quad (8)$$

To construct the estimation error **EE**, we recall that an MRA space V_n has a reproducing kernel $P_n(\mathbf{x},\mathbf{t})$ which is the kernel of the projection operator P_n. We note that technically V_n is defined to be an RHKS only if its kernel $P_n(\mathbf{x},\mathbf{t})$ is continuous, which occurs if ϕ is assumed continuous. We do not require this continuity assumption, but we will still use the term RHKS to describe V_n, since even without continuity, V_n has the RKHS properties required by us. Note that the reproducing kernel inner product in this case is inherited from L^2. The reproducing kernel $P_n(\mathbf{x},\mathbf{t})$ is bounded, symmetric, and positive semi-definite.

Our result will follow from Proposition 5 as a bound on h, the VC dimension, is determined by Proposition 6.

Note for any non-zero continuous function $g(\mathbf{t}) \in L^2(\mathbb{R}^d)$ we can find a positive $k(\mathbf{x})$, bounded away from 0, such that

$$\text{ess sup}_{\mathbf{x},\mathbf{t}} \left| \frac{P_n(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})} \right| \leq 2^{nd}\tau, \tag{9}$$

for some constant $\tau > 0$ if $n \geq 0$. This follows from the fact that $P_n(\mathbf{x},\mathbf{t})$ is non-zero only on a restricted band of the form $|\mathbf{x}-\mathbf{t}| \leq \beta 2^{-n}$ for some $\beta > 0$, given ϕ is compactly supported. We could for example let

$$k(\mathbf{x}) = \sup_{|\mathbf{x}-\mathbf{x}'|\leq\beta} \frac{1}{g(\mathbf{x}')}\,.$$

This quantity is positive since g is positive and continuous. Note that we have chosen g, k, τ which do not depend on n and since $g(\mathbf{x}) \in L^2$ we have $gf \in L^1$.

Let h be the VC dimension of $\dfrac{P_n(\mathbf{x},\mathbf{t})}{g(\mathbf{t})k(\mathbf{x})}$ in the parameter $\mathbf{x}$ and the variable $\mathbf{t}$. Then $f_n = P_n f \in L^2$ and by Proposition 5, for every $\eta > 0$ there exist $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\} \subset \mathbb{R}^d$, and m coefficients $c_i = \text{sgn}(f_n(\mathbf{t}_i)) = \pm 1$, such that the weighted L^∞ norm

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

$$\leq 4 \cdot 2^{nd}\tau \|f\|_{L^2}\|g\|_{L^2} \sqrt{\frac{h\ln\left(\frac{2em}{h}\right) - \ln\left(\frac{\eta}{4}\right)}{m}}\,, \tag{10}$$

with probability at least $1-\eta$ (note that $\|f\| \geq \|f_n\|$; above 2^{nd} represents the standard scaling factor in front of $P_n(\mathbf{x},\mathbf{t})$). Putting together (8) and (10), $f \in \mathcal{L}^2_s$ in the weighted L^∞ norm is approximated by

$$\widetilde{f}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} c_i \|f_n g\|_{L^1} \frac{P_n(\mathbf{x},\mathbf{t}_i)}{g(\mathbf{t}_i)} \tag{11}$$

$$= \frac{1}{m}\sum_{i=1}^{m} c_i \|f_n g\|_{L^1} \frac{2^{nd}\sum_{\mathbf{k}\in\mathbb{Z}^d} \phi(2^n\mathbf{x}-\mathbf{k})\phi(2^n\mathbf{t}_i-\mathbf{k})}{g(\mathbf{t}_i)} \tag{12}$$

with probability at least $1-\eta$. Interchanging the finite sums, this is equivalent to:

$$\widetilde{f}(\mathbf{x}) = \sum_{j=1}^{p} d_{k_j}\phi(2^n\mathbf{x}-\mathbf{k}_j)$$

where

$$d_{k_j} = \frac{1}{m}\|f_n g\|_{L^1} \sum_{i=1}^{m} \frac{2^{nd} c_i \phi(2^n \mathbf{t}_i - \mathbf{k}_j)}{g(\mathbf{t}_i)}, \tag{13}$$

and $\{\mathbf{k}_j\}_{j=1}^{p}$ is an indexing of the multiintegers $\mathbf{k}$ for which $2^n \mathbf{t}_i - \mathbf{k}_j$ is in the support of ϕ for some $\mathbf{t}_i$ (hence $p \leq ml$). □

As an immediate consequence of letting $\eta \to 1$ in the above theorem, we have

Corollary 12. *Given an MRA as above, $f \in \mathcal{L}_s^2(\mathbb{R}^d)$ and the assumptions of Theorem 7 and Corollary 4, there exists a finite subset $\{\mathbf{t}_i\}_{i=1}^m \subset \mathbb{R}^d$ (with $m \leq n$) such that if $\tilde{f}$ is as above, then*

$$\|\widetilde{f} - f\|_{\infty, 1/k(x)} \leq \mathbf{AE} + \mathbf{EE}*,$$

where

$$\mathbf{EE}* = 4 \cdot 2^{nd} \tau \|g\|_{L^2} \|f\|_{L^2} \sqrt{\frac{h \ln\left(\frac{2en}{h}\right) + \ln(4)}{n}}.$$

Proof. By letting $\eta \to 1$, we see

$$\mathbf{EE}* = \lim_{\eta \to 1} \mathbf{EE}.$$

Note we can use the method of Corollary 4 to eliminate the additive epsilon on the right of the estimate for $\mathbf{EE}^*$. □

We remark that the above results can be rewritten using wavelets ψ_λ instead of ϕ by rewriting the projections P_n in terms of ψ_λ instead of ϕ. However, our estimate on VC dimension of the wavelet/scaling approximation (Proposition 6) is much more easily stated in terms of scaling function rather than wavelet properties.

We note that in practical terms the constructive statements of the above results (in which $\eta \neq 1$) are more helpful, to the extent that they make a claim about a random choice of points $\mathbf{t}_i$, which can be made, given that we know $\rho = gf$, which gives us the correct probability density function. The $\eta \to 1$ limit gives the sharpest bound, but for a set of $\{\mathbf{t}_i\}$ which are unknown.

6. Conclusion

By obtaining bounds through a partition into estimation and approximation errors and using approximation results from wavelet theory (see [8], [9]) and statistical learning theory [18], [5], we have obtained finite scaling function

approximations of Sobolev functions and shown that the weighted L^∞ errors are "logarithmically close" (based on analogous results in [5]) to the best possible of their type.

Acknowledgements

The second author thanks Dan Williams of Howard University for helpful conversations.

References

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

[2] C.K. Chui, *Wavelets: A Tutorial in Theory and Applications*, Academic Press, N.Y. (1992).

[3] F. Cucker, S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.*, **39**, No. 1 (2002), 1-49.

[4] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Series in Applied Mathematics, SIAM (1992).

[5] F. Girosi, Approximation error bounds that use VC bounds, In: *Proc. International Conference on Artificial Neural Networks* (Ed-s: F. Fogelman-Soulied, P. Gallinari), Paris (1995), 295-302.

[6] E. Hernandez, G. Weiss, *First Course in Wavelets*, CRC Press (1996).

[7] S. Kelly, M. Kon, L. Raphael, Pointwise convergence of wavelet expansions, *Bull. Amer. Math. Soc.*, **30** (1994), 87-94.

[8] M. Kon, L. Raphael, Convergence rates of multiscale and wavelet expansions, *Wavelet Transforms and Time-Frequency Signal Analysis*, CBMS Conference Proceedings (Ed. L. Debnath), Chapter 2, Birkhäuser (2001), 37-65.

[9] M. Kon, L. Raphael, A Charcterization of Wavelet Convergence in Sobolev Spaces, *Applicable Analysis*, **78** (2001), 271-324.

[10] M. Kon, L. Raphael, Approximating functions in reproducing kernel hilbert spaces via statisitcal learning theory, *Splines and Wavelets* (Ed-s: G. Chen, M-J Lai) 2005, 271-266.

[11] M. Kon, L. Raphael, 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.

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

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

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

[15] S. Smale, D-X. Zhou, Estimating the approximation error in learning theory, *Analysis and Applications*, **1**, No. 1 (2003), 17-41.

[16] S. Smale, D-X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.*, **41** (2004), 279-305.

[17] E. Sontag, VC dimension of neural networks, In: *Neural Networks and Machine Learning* (Ed. C.M. Bishop), Springer-Verlag, Berlin (1998), 69-95.

[18] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Second Edition, Springer (2000).

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

[20] D-X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory*, **49**, No. 7 (2003), 1743-1752.