MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1288 April 1995
C.B.I.P. Paper No. 73

Convergence Rates of Approximation by Translates

Federico Girosi and Gabriele Anzellotti

Abstract

In this paper we consider the problem of approximating a function belonging to some function space Φ by a linear combination of n translates of a given function G. Using a lemma by Jones (1990) and Barron (1991) we show that it is possible to define function spaces and functions G for which the rate of convergence to zero of the error is $O(\frac{1}{\sqrt{n}})$ *in any number of dimensions*. The apparent avoidance of the "curse of dimensionality" is due to the fact that these function spaces are more and more constrained as the dimension increases. Examples include spaces of the Sobolev type, in which the number of weak derivatives is required to be larger than the number of dimensions. We give results both for approximation in the L_2 norm and in the L_∞ norm. The interesting feature of these results is that, thanks to the *constructive* nature of Jones' and Barron's lemma, an iterative procedure is defined that can achieve this rate.

© Massachusetts Institute of Technology, 1995

This paper describes research done within the Center for Biological Information Processing, in the Department of Brain and Cognitive Sciences, at the Artificial Intelligence Laboratory and at the Department of Mathematics, University of Trento, Italy. Gabriele Anzellotti is with the Department of Mathematics, University of Trento, Italy. This research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Artificial Intelligence Center of Hughes Aircraft Corporation (S1-801534-2). Support for the A. I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010, and in part by ONR contract N00014-85-K-0124.

1 Introduction

Let Φ be a normed space of functions and let A be a subset of Φ. The prototypical problem in approximation theory consists in approximating an element f of Φ by an element of A, that is looking for an element in A that has minimum distance from f. It is also natural to consider the *distance of f from A* as

$$\delta(f, A) \equiv \inf_{a \in A} \|f - a\| \tag{1}$$

and to study this quantity for different choices of A and $f \in \Phi$. In the classical theory of approximation the set A is usually a linear k-dimensional subspace $A_k \subset \Phi$ (Lorentz, 1986) (the algebraic or the trigonometric polynomials of given degree and the splines with fixed knots are typical examples of such subspaces), while in nonlinear approximation theory the linear subspace A_k is replaced by a k-dimensional manifold M_k (DeVore, 1991). Usually one has a family of manifolds $\{M_k\}_{k=1}^{\infty}$ such that $\bigcup_k M_k$ is dense in Φ and

$$M_1 \subset M_2 \subset \ldots \subset M_n \subset \ldots$$

so that the quantity $\delta(f, M_k)$ is a monotone decreasing function of k converging to zero and the approximation in M_k gets arbitrarily close to f provided one takes k sufficiently large. However, since the computational time needed to find an approximation to f in M_k is going to increase with k, it is of great interest to know the rate of convergence to zero of $\delta(f, M_k)$ as a function of k. This rate of convergence can be taken as a measure of the complexity of f with respect to the manifolds M_k, in the sense that "simple" functions should have a fast rate of convergence.

As an example, let us consider as space Φ the space $\Lambda^d_{s\alpha}$ of the functions whose partial derivatives of order s are bounded in the uniform norm on the d-dimensional cube $I = [0,1]^d$ and satisfy a Lipschiz condition with exponent α (Lorentz, 1986, p. 50). On the space Φ we consider the uniform norm $\|f\| = \max_I |f(x)|$. Choosing as manifold M_k the set of polynomials of degree $n-1$ in each of the d variables, that is a linear space of dimension $k = n^d$, the following bound can be obtained (Lorentz, 1986):

$$\delta(f, M_k) \leq N d k^{-\frac{s+\alpha}{d}} \tag{2}$$

where N is a constant that depends on f and s.

From this example we see that the rate of convergence dramatically slows down when the dimension d increases, revealing the discouraging phenomenon known under the name of "curse of dimensionality" (Bellman, 1961). However, for every fixed number of dimensions, arbitrary inverse-power rates of convergence can be obtained if the smoothness index s is chosen big enough. This result is typical in linear approximation theory since the computation of the n-width of the space $\Lambda^d_{s\alpha}$ shows that the best linear technique cannot improve the rate of convergence $O(k^{-\frac{s+\alpha}{d}})$ (Lorentz 1986, p. 135).

Similar results, in both linear and nonlinear approximation theory (DeVore, 1991), hold for other spaces of functions in which smoothness is measured in a different way. We are therefore led to argue that in practical situations we can only approximate functions whose smoothness increases with the dimension. As an example we consider again the spaces $\Lambda^d_{s\alpha}$ for $s = d$. It is clear from eq. (2) that in this case the rate of convergence of polynomial approximation to an $f \in \Lambda^d_{d\alpha}$ is $O(k^{-1})$ and it is in this sense "independent on dimensionality".

In a recent paper (1990) Jones showed how to construct a sequence of functions f_n that approximate certain functions in a Hilbert space with a rate of convergence $O(\frac{1}{\sqrt{n}})$. A statement of Jones' lemma is given in section 2. An application of this result to projection pursuit regression and neural networks has already been presented in (Jones 1990; Barron 1991), where appropriate approximation schemes and spaces Φ^d of functions in R^d are described in which the complexity of approximation increases mildly with d. It is worthwhile to observe that this is obtained at the expense that the functions contained in Φ^d are more and more "regular" when d increases. Moreover, it is not completely clear yet how computationally expensive the approximation f_n may be. A very short review of Jones' and Barron's results is given in section 5.

The aim of this paper is to present an application of Jones' lemma to the approximation by linear combination of translates of a given function G. In particular for appropriate choices of G we obtain estimates for the rate of convergence of certain Radial Basis Functions schemes (Micchelli, 1986; Powell, 1987; Dyn, 1991; Poggio and Girosi, 1990) on certain spaces of functions of Sobolev type. For the convenience of the reader we collect in the appendix a few known results about Sobolev spaces and integration of Banach valued functions.

2 The Maurey-Jones-Barron Lemma

Our result is based on a lemma by Jones (1990) on the convergence rate of an iterative approximation scheme in Hilbert spaces. A formally similar lemma, brought to our attention by R. Dudley (Dudley, 1991), is due to Maurey, and was published by Pisier in 1981. However Jones' lemma is constructive while Maurey's is not. Here we report a version of the lemma due to Barron (Barron 1991) that contains a slight refinement of Jones' result:

Lemma 2.1 (Maurey-Jones-Barron) *If f is in the closure of the convex hull of a set $\mathcal{G}$ in a Hilbert space H with $\|g\| \leq b$ for each $g \in \mathcal{G}$, then for every $n \geq 1$ and for $c > b^2 - \|f\|^2$ there is a f_n in the convex hull of n points in $\mathcal{G}$ such that*

$$\|f - f_n\|^2 \leq \frac{c}{n} \ .$$

The interesting feature of this lemma is that the sequence $\{f_n\}_{n=0}^{\infty}$ has the following structure:

$$f_{n+1} = \alpha_n f_n + (1 - \alpha_n) g_n \tag{3}$$

where α_n and g_n are chosen in order to "approximately solve" the following minimization problem:

$$\inf_{\alpha_n \in R, g_n \in \mathcal{G}} \|f - \alpha_n f_n - (1 - \alpha_n) g_n\|$$

where by "approximately solve" we mean that it is sufficient at each step to reach a distance from the infimum of order $O(\frac{1}{n^2})$. The lemma is therefore constructive, providing a procedure that can achieve the prescribed rate.

In order to exploit this result we need to define suitable classes of functions which are the closure of the convex hull of some subset $\mathcal{G}$ of a Hilbert space H. We are therefore naturally led to study functions that can be represented as "infinite" convex combinations of the type

$$f = \sum_{i=1}^{\infty} \alpha_i g_i \quad \alpha_i \geq 0 \ , g_i \in \mathcal{G} \ \ , \sum_{i=1}^{\infty} \alpha_i = 1 \ . \tag{4}$$

One way to approach the problem consists in utilizing the *integral representation* of functions. Suppose that the functions in a Hilbert space H can be represented by the integral

$$f(\mathbf{x}) = \int_{\mathcal{M}} G_{\mathbf{t}}(\mathbf{x}) d\alpha(\mathbf{t}) \tag{5}$$

where $d\alpha$ is some measure on the parameter set $\mathcal{M}$. If $d\alpha$ is a finite measure, the integral (5) can be seen as an infinite convex combination of the type of eq. (4), and therefore the function f belongs to the closure of the convex hull of some subset of H. In the next section we formalize this idea in the special case in which the functions $G_{\mathbf{t}}(\mathbf{x})$ are the translates $G(\mathbf{x}-\mathbf{t})$ of a fixed function G and we show how it leads to define approximation techniques whose rate of convergence in appropriate spaces of functions is $O(\frac{1}{\sqrt{n}})$.

3 Approximation by Translates of a Function G

Let G be a fixed function belonging to $L_2(R^d) \equiv L_2$. We define the space L_G as the set of the functions of the form

$$f = G * \lambda \tag{6}$$

where λ is any signed Radon measure whose total variation $|\lambda|_{R^d} \equiv \|\lambda\|$ is finite and the symbol $*$ stands for the convolution operation. Assuming from now on that $\|G\|_{L_2} = 1$, the following inequality holds (Stein and Weiss, 1971)

$$\|f\|_{L_2} \leq \|\lambda\|$$

showing the inclusion $L_G \subset L_2$. It is natural to approximate elements of L_G by elements of the set

$$G_n = \{f \in L_2 \mid f = \sum_{i=1}^{n} \lambda_i G_{\mathbf{t}_i} \ , \ \lambda_i \in R \ , \ \mathbf{t}_i \ \in R^d\} \ , \tag{7}$$

where we indicate by $G_{\mathbf{t}}$ the function G translated by the vector $\mathbf{t}$, that is $G_{\mathbf{t}}(\mathbf{x}) = G(\mathbf{x}-\mathbf{t})$. Using lemma 2.1 we can now prove the following

Theorem 3.1 *Let f be a function in L_G, so that $f = G * \lambda$, where $G \in L_2$, $\|G\|_{L_2} = 1$, and λ is a Radon signed measure of bounded total variation $\|\lambda\|$. Then f belongs to the L_2-closure of the convex hull of the set*

$$A = \{sG_{\mathbf{t}} \mid \mathbf{t} \in R^d, \ |s| \leq \|\lambda\|\}$$

and there exist n coefficients c_α and n vectors $\mathbf{t}_\alpha$ such that:

$$\|f - \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x}-\mathbf{t}_\alpha)\|_{L_2}^2 \leq \frac{c}{n}$$

for all $c > \|\lambda\|^2 - \|f\|_{L_2}^2$.

Proof: We consider the vector-valued function

$$T : R^d \rightarrow L_2(R^d)$$

such that

$$T(\mathbf{t}) = G_{\mathbf{t}} \ .$$

The function T is continuous, hence λ-measurable, moreover one has

$$\int_{R^d} \|T(\mathbf{t})\|_{L_2} d|\lambda|(\mathbf{t}) = \|G\|_{L_2} \int_{R^d} d|\lambda|(\mathbf{t}) = \|\lambda\| < +\infty \ .$$

Therefore it exists the Bochner integral of T with respect to λ (see appendix A):

$$\eta = \int_{R^d}^{\mathcal{B}} T(\mathbf{t}) d\lambda(\mathbf{t}) \ ,$$

and by lemma (A.2) we have

$$\eta \in \overline{co\ A} \tag{8}$$

where $A = \{sG_{\mathbf{t}} \mid \mathbf{t} \in R^d, \ |s| \leq \|\lambda\|\}$, $co\ A$ stands for the convex hull of the set A and the bar stands for the closure in L_2. Now we shall prove that $\eta = f$. This can be done by proving that

$$F^* f = F^* \eta \quad , \ \forall F^* \in (L_2)^* \tag{9}$$

where $(L_2)^*$ is the dual space of L_2, that is L_2 itself. From the properties of the Bochner integral we have:

$$F^* \eta = F^* \int_{R^d}^{\mathcal{B}} T(\mathbf{t}) d\lambda(\mathbf{t}) = \int_{R^d} (F^* G_{\mathbf{t}})) d\lambda(\mathbf{t}) \ .$$

Taking this into account, the identity (9) can be written as:

$$\int_{R^d} d\mathbf{x} \ \phi(\mathbf{x}) \int_{R^d} G(\mathbf{x}-\mathbf{t}) d\lambda(\mathbf{t}) = \int_{R^d} d\lambda(\mathbf{t}) \int_{R^d} d\mathbf{x} \ \phi(\mathbf{x}) G(\mathbf{x}-\mathbf{t}) \ , \ \forall \phi \in L_2 \ .$$

Now by Fubini's theorem the two sides of this last equation are equal, and therefore $\eta = f$.
By eq. (8) $f = \eta$ belongs to the L_2 closure of the convex hull of the set A, which is contained in the ball of radius $\|\lambda\|$. By the Maurey-Jones-Barron lemma we can find a set of n coefficients c_α and n vectors $\mathbf{t}_\alpha$ such that:

$$\|f - \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x} - \mathbf{t}_\alpha)\|_{L_2}^2 \leq \frac{c}{n}$$

for all $c > C(f) = \|\lambda\|^2 - \|f\|_{L_2}^2$. $\square$

In theorem (3.1) the approximation error is measured in the L_2 norm. Imposing some restrictions on the function G a similar estimate can be obtained for other norms, and in particular for the L_∞ norm. In fact, suppose that $G \in H^{s,2}$, where $H^{s,2}(R^d) \equiv H^{s,2}$ is the Sobolev space of the functions whose weak derivatives up to order s are in L_2 (see Appendix B). Then one can easily see that theorem (3.1) can be formulated in the Hilbert space $H^{s,2}$ instead of L_2:

Theorem 3.2 *Let f be a function such that $f = G * \lambda$, where $G \in H^{s,2}$, $\|G\|_{H^{s,2}} = 1$, and λ is a Radon signed measure of bounded total variation $\|\lambda\|$. Then f belongs to the $H^{s,2}$-closure of the convex hull of the set*

$$A = \{sG_{\mathbf{t}} \mid \mathbf{t} \in R^d, \ |s| \leq \|\lambda\|\}$$

and there exist n coefficients c_α and n vectors $\mathbf{t}_\alpha$ such that:

$$\|f - \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x} - \mathbf{t}_\alpha)\|_{H^{s,2}}^2 \leq \frac{c}{n}$$

for all $c > \|\lambda\|^2 - \|f\|_{H^{s,2}}^2$.

We notice that if the condition $s > \frac{d}{2}$ holds, then the Sobolev embedding theorem (see Appendix B) guarantees that $H^{s,2} \subset C^0$ and that it exists $c_1 > 0$ such that

$$\|\cdot\|_\infty \leq c_1 \|\cdot\|_{H^{s,2}} \ .$$

Therefore the approximating sequence $\{f_n\}$ converges uniformly, and the following corollary holds:

Corollary 3.1 *Under the conditions of theorem (3.2), if $s > \frac{d}{2}$ there exists n coefficients c_α, n vectors $\mathbf{t}_\alpha$ and a constant c_1 such that:*

$$\|f - \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x} - \mathbf{t}_\alpha)\|_{L_\infty}^2 \leq c_1 \frac{c}{n}$$

for all $c > \|\lambda\|^2 - \|f\|_{H^{s,2}}^2$.

From a practical point of view, in many cases, what it is really interesting is an estimate of the error in the sup norm, instead of the L_2 or $H^{s,2}$ norm. Think for example of the problem of approximating the trajectory of a robot arm: it is clear that what is really needed in this case is a small L_∞ norm of the difference between the desired and the approximated trajectory, while a small L_2 norm is of little interest.

Remark: we notice that the elements of the set G_n defined by eq. (7) can also be seen as points of a manifold M_k whose dimension is $k = n(d+1)$. Therefore theorem (3.1) can also be formulated in terms of the number of parameters k that are needed to achieve a certain error, saying that if $f \in L_G$ then

$$\delta(f, M_k) \leq C(f)\sqrt{\frac{d+1}{k}} \ .$$

If we compare this result with the typical estimates (DeVore, 1991), we notice that in this case the way the dimension affects the convergence curve is much less dramatic, corresponding to a simple scale dilation. This means that in some sense the complexity of the space L_G does not increase very much when the dimension increases. It is interesting to characterize, for several specific choices of G, the structure of L_G and to understand whether it contains a "sufficiently large" set of functions, where by "sufficiently large" we mean large enough to contain functions that are encountered in practical cases. This will be done in the next section for two particular choices of G.

4 Examples of functions G

In this section we consider two choices for the function G and study the corresponding functions spaces L_G. We remind that for any given $G \in L_2(R^d)$ the space L_G is defined as

$$L_G = \{f \in L_2(R^d) \mid f = G * \lambda \ , \ \lambda \in \mathcal{M}(R^d)\}$$

where $\mathcal{M}(R^d) \equiv \mathcal{M}$ is the space of Radon signed measures of bounded total variation on R^d.

4.1 The Gaussian

We consider the Gaussian function $G(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$, since approximation with Gaussian basis functions is often used in practical applications (Moody and

Darken, 1989; Poggio and Girosi, 1990; Poggio and Edelman, 1990; Sanner and Slotine, 1992). Clearly $G \in L_2(R^d)$, so that the space L_G is well defined in any dimension. Due to the smoothness of the Gaussian and to its fast decay property this space of functions is rather small. However it contains an interesting subset of the space of band limited functions, the functions whose Fourier transform has compact support. In particular, let us define the space of functions $B_k(R^d)$:

$$B_k(R^d) \equiv \{f \mid \tilde{f} \in C_0^k(R^d)\} \ , \tag{10}$$

that is the set of functions whose Fourier transform has compact support and k continuous derivatives. Then the following inclusion holds:

$$B_k(R^d) \subset L_G \ , \ \forall k > \frac{d}{2} \ . \tag{11}$$

In fact if $f \in B_k(R^d)$ then we have

$$\frac{\tilde{f}(\mathbf{s})}{\tilde{G}(\mathbf{s})} = \alpha e^{\|\mathbf{s}\|^2} \tilde{f}(\mathbf{s}) \equiv \tilde{\lambda} \in C_0^k(R^d) \ ,$$

where α is a constant depending only on the dimension d. Therefore $f = G * \lambda$ where λ is the Fourier transform of the function $\tilde{\lambda} = \frac{\tilde{f}}{\tilde{G}}$. Since the following inclusion holds (see appendix B):

$$C_0^k(R^d) \subset A(R^d) \ , \ \forall k > \frac{d}{2} \ ,$$

where $A(R^d)$ is the space of the functions whose Fourier transform belongs to $L_1(R^d)$, then $\lambda \in L_1$ and $f \in L_G$.

We notice that the Gaussian function and its derivatives of any order belongs to L_2, and therefore $G \in H^{s,2}$ for any $s > 0$. Hence we can apply corollary (3.1) to conclude that the convergence rate $O(\frac{1}{\sqrt{n}})$ also holds for approximation in the sup norm.

4.2 Bessel-Macdonald Kernels

We now consider the Bessel-Macdonald kernels, a family of functions $G_m(\mathbf{x})$ defined in terms of their Fourier transforms:

$$\tilde{G}_m(\mathbf{s}) = \frac{1}{(1 + 4\pi^2\|\mathbf{s}\|^2)^{\frac{m}{2}}} \qquad m > 0 \ .$$

The functions $G_m(\mathbf{x})$ are integrable functions that decay exponentially at infinity and may have a singularity at the origin (Stein, 1970, p. 132). However if $m > d$ they are continuous and actually differentiable of any order $q < m - d$. We want to work with continuous funtions and in what follows we will always make the assumption $m > d$. Since $\tilde{G}_m(\mathbf{s})$ is positive and radial, we also have that, by Bochner's theorem, $G_m(\mathbf{x})$ is positive definite (Micchelli, 1986), and therefore approximation by translates of $G_m(\mathbf{x})$ is a Radial Basis Functions approximation scheme. The following observations can be done regarding the functions G_m and the space L_{G_m}:

1. One has

$$G_m \in H^{s,2} \quad \text{for} \quad 0 < s < m - \frac{d}{2} \ .$$

   Since we have made the assumption $m > d$ one can take s such that $\frac{d}{2} < s < m - \frac{d}{2}$. Then we can apply corollary (3.1) to conclude that the rate of convergence $O(\frac{1}{\sqrt{n}})$ also holds for approximation in the sup norm.

2. Since $L_1 \subset \mathcal{M}$, the space L_{G_m} contains the space $\mathcal{L}^1_m(R^d) \equiv \mathcal{L}^1_m$ of those functions that can be written as $f = G_m * \lambda$ with $\lambda \in L_1$. For more information about the space $\mathcal{L}^1_m$, which is a special instance of the so called *potential spaces*, the reader is referred to (Stein, 1970). The space $\mathcal{L}^1_m$ is related to the Sobolev space $H^{m,1}(R^d) \equiv H^{m,1}$ of the functions whose weak derivatives up to order m are in L_1 (see Appendix B). More precisely one has (Stein 1970, p. 160):

$$H^{m,1} \subset \mathcal{L}^1_m \quad \subset L_{G_m} \qquad \text{for all } m \text{ even} \ .$$

   Therefore we conclude that if $m > d$ and m is even, by superposition of translates of G_m we can approximate with a rate of convergence $O(\frac{1}{\sqrt{n}})$ all the functions of $H^{m,1}$, and hence all C^m functions which rapidly decrease to infinity.

3. Again for $s < m - \frac{d}{2}$ and $m > d$, m even, we have the following characterization of the space L_{G_m}:

$$L_{G_m} = \{f \in H^{s,2} \ \mid (I - \Delta)^{\frac{m}{2}} f \in \mathcal{M}\} \ .$$

In fact, if $f \in L_{G_m}$ that is $f = G_m * \lambda$ with $\lambda \in \mathcal{M}$, then $(I-\Delta)^{\frac{m}{2}} f = \lambda$ since G_m is the fundamental solution of the operator $(I-\Delta)^{\frac{m}{2}}$. On the other hand, if $f \in H^{s,2}$ and $(I-\Delta)^{\frac{m}{2}} f = \lambda \in \mathcal{M}$, then by taking the convolution of both sides with G_m we have $f = G_m * \lambda$.

5 Other Approximation Schemes

Other choices of integral representation lead to different approximation schemes and different spaces of functions that can be approximated with a similar convergence rate. For example, using the Fourier representation of a function (if it exists) we have:

$$f(\mathbf{x}) = \int_{R^d} d\mathbf{s} \ \cos(\mathbf{s} \cdot \mathbf{x} + \theta(\mathbf{s}))|\tilde{f}(\mathbf{s})| \tag{12}$$

where $\theta(\mathbf{s})$ is the phase of the Fourier transform $\tilde{f}(\mathbf{s})$ of f. Jones (1990) considers the space $A(R^d)$ (appendix B) of the functions such that their Fourier transform is in $L_1(R^d)$ and shows that they can be approximated by functions of the form

$$f_n(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \cos(\mathbf{t}_i \cdot \mathbf{x} + \theta_i) \tag{13}$$

with the rate of convergence $O(\frac{1}{\sqrt{n}})$.

Another result of this type has been proved by Barron (1991). He considers the set of the functions such that

$$\int_{R^d} d\mathbf{s} \ \|s\||\tilde{f}(\mathbf{s})| < +\infty \tag{14}$$

that is the functions whose gradient is in $A(R^d)$, and approximates elements of this set by functions of the form

$$f_n(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \sigma(\mathbf{t}_i \cdot \mathbf{x} + \theta_i) \ ,$$

where $\sigma(\cdot)$ is any sigmoidal function. Condition eq. (14) can be rewritten as

$$\|s\||\tilde{f}(\mathbf{s})| \in L_1(R^d). \tag{15}$$

Denoting by I_d the function

$$I_d(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^{d-1}}$$

and noticing that its Fourier transform is $\tilde{I}_d(\mathbf{s}) = \|\mathbf{s}\|^{-1}$ we can also say that the space of function that satisfy condition eq. (14) is the space of function that can be written as

$$f = I_d * \lambda \ , \ \lambda \in A(R^d). \tag{16}$$

There is a remarkable analogy between this set of function and the function space $\mathcal{L}^1_m$ considered in section (4.2), that is the set of functions such that:

$$f = G_m * \lambda \ , \ \lambda \in L_1(R^d) \ , \ m > d \ . \tag{17}$$

In eq. (16), the function I_d goes to zero faster and faster as d increases, while its Fourier transform remains unchanged. In eq. (17), because of the constraint $m > d$, it is *the Fourier transform* of G_m that goes to zero faster and faster as d increases, while the asymptotic decay of G_m is always exponential. Moreover, in eq. (17) λ has to belong to L_1, while in eq. (16) it is *the Fourier transform* of λ that belongs to L_1.

6 Conclusions

We briefly summarize the main results presented in this paper.

- Let f be a function on R^d and assume that f can be written as $f = G*\lambda$, where G is square integrable on R^d and λ is a signed Radon measure of bounded total variation. Then there is a linear superposition of n translates of G that approximates f in the L_2 norm with a rate of convergence $O(\frac{1}{\sqrt{n}})$.

- Let f be a function on R^d whose Fourier transform has compact support and k continuous derivatives, with $k > \frac{d}{2}$. Then there exists a Gaussian Radial Basis Functions expansion with n basis functions that approximates f in the L_2 norm with a rate of convergence $O(\frac{1}{\sqrt{n}})$. The same result holds for approximation in the sup norm.

- Let f be any function of the Sobolev space $H^{m,1}(R^d)$, with $m > d$, m even. Then there exists a Radial Basis Functions expansion, whose basis function is the Bessel-Macdonald kernel $G_m(\mathbf{x})$, that approximates f with a rate of convergence $O(\frac{1}{\sqrt{n}})$ in the norm of $H^{s,2}$, with $\frac{d}{2} < s < m - \frac{d}{2}$. A similar rate of convergence can also be obtained for the approximation in the sup norm.

All these examples involve spaces of functions with a number of derivatives that increases with the dimension, and are consistent with the intuitive idea that spaces of function in a high number of dimensions are very difficult to approximate, unless some constraints are imposed to prevent their "size" to grow exponentially fast.

One interesting feature of these results is that, thanks to the *constructive* nature of Jones' and Barron's lemma, an iterative procedure is provided that can achieve that rate. Clearly, these results concern the approximation of a function f which is known everywhere, while in many practical situations one would like to construct an approximation of a function f knowing only the values of f on some (finite) set of points. For this last problem, in the case of approximation by sigmoidal ridge functions, some results by Barron (1992) are already available, and show that also with this further source of error one can obtain results "independent on the dimension", for suitable spaces of functions. It should be possible to obtain similar results for the approximation scheme we considered here, using the same technique.

Acknowledgements We thank Tomaso Poggio for useful discussions and for a critical reading of the manuscript.

A The Bochner Integral

Let $\Omega \subset R^d$ and let λ be a positive measure on Ω. For functions $f : \Omega \to X$ with X a Banach space there are several available notions of measurability and integration (Dunford and Schwartz, 1958; Diestel and Uhl, 1977). In particular for all (strongly) λ-measurable functions f such that $\int_\Omega \|f\|_X \, d\lambda < +\infty$ we can define the Bochner integral

$$\int_\Omega^{\mathcal{B}} f \, d\lambda \quad . \tag{18}$$

Clearly if λ is a Borel measure the continuous functions $f : \Omega \to X$ are (strongly) measurable. One has lemma A.1 below (Diestel and Uhl 1977, page 48).

Lemma A.1 *Let λ be a positive Borel measure on $\Omega \subset R^d$ and $f(\mathbf{t}) : \Omega \to X$ with X a Banach space. If f is Bochner integrable with respect to λ then*

$$\frac{1}{\lambda(\Omega)} \int_\Omega^{\mathcal{B}} f(\mathbf{t}) d\lambda(\mathbf{t}) \in \overline{co \ f(E)} \ .$$

If one considers a signed Radon measure λ on Ω one can still define the integral of a measurable function $f : \Omega \to X$ with respect to λ as

$$\int_{\Omega}^{\mathcal{B}} f(\mathbf{t}) d\lambda(\mathbf{t}) \equiv \int_{\Omega}^{\mathcal{B}} f(\mathbf{t}) \frac{d\lambda}{d|\lambda|}(\mathbf{t}) d|\lambda|(\mathbf{t}) \tag{19}$$

where $|\lambda|$ is the total variation of λ and $\frac{d\lambda}{d|\lambda|}$ denotes the Radon-Nikodym derivative of λ with respect to $|\lambda|$. From lemma (A.1) one can easily obtain:

Lemma A.2 *Let λ be a signed Radon measure on $\Omega \subset R^d$ and $f(\mathbf{t}) : \Omega \to X$ with X a Banach space. If f is λ-measurable and is such that*

$$\int_{\Omega} \|f\| \ d|\lambda| < +\infty$$

then the Bochner integral of f with respect to λ is well defined and

$$\frac{1}{|\lambda|(\Omega)} \int_{\Omega}^{\mathcal{B}} f(\mathbf{t}) \ d\lambda(\mathbf{t}) \in \overline{co\ S} \ . \tag{20}$$

where

$$S = \{ s f(\Omega) \mid s \in R \ , |s| \leq 1 \} \ .$$

In fact the scalar function $\frac{d\lambda}{d|\lambda|}(\mathbf{t})$ is measurable, the function $f(\mathbf{t})\frac{d\lambda}{d|\lambda|}(\mathbf{t})$ is measurable, and moreover

$$\int_{\Omega} \|f \frac{d\lambda}{d|\lambda|}\| \int_{\Omega} \|f\| \ d|\lambda| < +\infty \ .$$

Hence the integral $\int_{\Omega}^{\mathcal{B}} f \ d\lambda$ is well defined as the right member of (14). Then by lemma (A.1) applied to the function $h(\mathbf{t}) = f(\mathbf{t})\frac{d\lambda}{d|\lambda|}(\mathbf{t})$ one has:

$$\frac{1}{|\lambda|(\Omega)} \int_{\Omega}^{\mathcal{B}} f(\mathbf{t}) \frac{d\lambda}{d|\lambda|}(\mathbf{t}) \ d|\lambda|(\mathbf{t}) \in \overline{co\ h(\Omega)} \ .$$

On the other hand since $|\frac{d\lambda}{d|\lambda|}| = 1$ one has

$$co\ h(\Omega) = co\ S$$

and (20) follows.

B Sobolev Spaces and the Space A

Here we collect a few facts about certain spaces of functions frequently used in the paper.

Sobolev Spaces. For each positive integer s and $1 \leq p \leq \infty$ one defines the Sobolev Space $H^{s,p}(R^d) \equiv H^{s,p}$ as the space of those L_p functions in R^d whose derivatives up to the order s are L_p functions.The space $H^{s,p}$ is a Banach space with the norm

$$\sum_{|\alpha| \leq s} \|D^\alpha f\|_{L_p}$$

where α is a multi-index and D^α is the derivative of order α. The space $H^{s,2}$ is a Hilbert space with respect to the scalar product

$$(u,v) = \sum_{|\alpha| \leq s} \int_{R^d} D^\alpha u \ D^\alpha v \ .$$

One has also the characterization

$$H^{s,2} = \{u \in L_2 \quad | \quad (1 + |\boldsymbol{\omega}|^2)^{\frac{s}{2}} \tilde{u}(\boldsymbol{\omega}) \in L_2\}$$

which can be used also to define the Sobolev spaces $H^{s,2}$ for non integer s. One has the following result, which is a special case of the Sobolev embedding theorem (Stein, 1970):

Theorem B.1 *If k is a positive integer and $s > k + \frac{d}{2}$ then*

$$H^{s,2} \subset C^k$$

and there is a constant c_1 such that

$$\max_{|\alpha| \leq k} \sup_{x \in R^d} |D^\alpha f(x)| \leq c_1 \|f\|_{H^{s,2}}.$$

The Fourier algebra A. The space A of the tempered distributions whose Fourier transform is a summable function is in current use in Fourier analysis (Herz, 1968; Katznelson, 1968). One has

$$H^{k,2} \subset A \qquad \text{for} \quad k > \frac{d}{2}$$

In fact (Barron, 1991; footnote) one may write

$$\tilde{f} = \frac{1}{(1+|\boldsymbol{\omega}|^2)^{\frac{k}{2}}}[\tilde{f}(1+|\boldsymbol{\omega}|^2)^{\frac{k}{2}}]$$

where both factors on the right side belong to L_2 if $k > \frac{d}{2}$. In particular it follows that $C_0^k \subset H^{k,2} \subset A$ for $k > \frac{d}{2}$.

It is also clear that $A \subset C_0$ where C_0 is the completion in the L_∞ norm of C_0^0 i.e. the space of continuous bounded functions that converge to zero for $\|\mathbf{x}\| \to \infty$.

References

[1] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. Technical Report 58, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, March 1991.

[2] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 1992. (to appear).

[3] R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[4] R.A. DeVore. Degree of nonlinear approximation. In C.K. Chui, L.L. Schumaker, and D.J. Ward, editors, *Approximation Theory, VI*, pages 175–201. Academic Press, New York, 1991.

[5] J. Diestel and J.J. Uhl. *Vector Measures*, volume 15 of *Mathematical Surveys*. American Mathematical Society, Providence, Rhode Island, 1977.

[6] R.M. Dudley. Comments on two preprints: Barron (1991), Jones (1991). Personal communication, March 1991.

[7] N. Dunford and J. Schwartz. *Linear operators*. Pure and applied mathematics, v. 7. Interscience Publishers, New York, 1958.

[8] N. Dyn. Interpolation and approximation by radial and related functions. In C.K. Chui, L.L. Schumaker, and D.J. Ward, editors, *Approximation Theory, VI*, pages 211–234. Academic Press, New York, 1991.

[9] C.S. Herz. Lipschitz spaces and Bernstein's theorem on absolutely convergent Fourier transforms. *Indiana Journal of Mathematics*, pages 283–323, 1968.

[10] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistic*, 1990. (to appear).

[11] Y. Katznelson. *An introduction to harmonic analysis.* John Wiley and Sons, New York, 1968.

[12] G. G. Lorentz. *Approximation of Functions.* Chelsea Publishing Co., New York, 1986.

[13] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[14] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[15] G. Pisier. Remarques sur un resultat non publiè de B. Maurey. In Centre de Mathematique, editor, *Seminarie d'analyse fonctionelle 1980–1981*, Palaiseau, 1981.

[16] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1990.

[17] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990.

[18] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[19] M. J. D. Powell. Radial basis functions for multivariable interpolation: a review. In J. C. Mason and M. G. Cox, editors, *Algorithms for Approximation.* Clarendon Press, Oxford, 1987.

[20] R.M. Sanner and J.-J.E. Slotine. Gaussian networks for direct adaptive control. *IEEE Transactions on Neural Nets*, 3(6):837–863, November 1992.

[21] E.M. Stein. *Singular integrals and differentiability properties of functions.* Princeton, N.J., Princeton University Press, 1970.

[22] E.M. Stein and G. Weiss. *Introduction to Fourier analysis on Euclidean spaces.* Princeton mathematical series, 32. Princeton University Press, Princeton, NJ, 1971.