

Lower Bounds for Approximation of some classes of Lebesgue measurable functions by Sigmoidal Neural Networks

José L. Montaña and Cruz E. Borges

Departamento de Matemáticas, Estadística y Computación,
Universidad de Cantabria, 39005 Santander, Spain
{montanj1, borgesce}@unican.es

Abstract. We propose a general method for estimating the distance between a compact subspace K of the space $L^1([0, 1]^s)$ of Lebesgue measurable functions defined on the hypercube $[0, 1]^s$, and the class of functions computed by artificial neural networks using a single hidden layer, each unit evaluating a sigmoidal activation function. Our lower bounds are stated in terms of an invariant that measures the oscillations of functions of the space K around the origin. As an application we estimate the minimal number of neurons required to approximate bounded functions satisfying uniform Lipschitz conditions of order α with accuracy ϵ .

Key words: Mathematics of Neural Networks, Approximation Theory

1 Introduction

During the last decade a great deal of research in the field of approximation theory has been done to approximate real valued functions using artificial neural networks (ANN's) with one or more hidden layers, each neuron evaluating a sigmoidal or radial basis function (see [1–9]). A typical result in this context is a density result showing that an ANN can approximate a given function in a given class to any degree of accuracy provided that enough number of neurons can be used.

One of the most general results in the core of approximation by ANN's states that any continuous function can be approximated to any degree of accuracy by a single hidden layer ANN having a non-polynomial activation function provided that sufficiently many hidden units are available ([10]). This powerful statement extends the seminal theorems by Hornik, Stinchcombe and White([11]) where approximation by multilayer ANN's is proven for L^p functions under the L^p norm and for continuous functions with respect to the uniform norm. Regarding the single hidden layer case it is also known that ANN's with a hidden layer are capable of providing an optimal order of approximation for functions assumed to admit a given number of derivatives, if the activation functions evaluated by each principal element satisfy certain technical conditions (see [6]). Under these conditions it is also possible to construct networks that provide a geometric order of approximation for analytic target functions.

A related important problem is the *complexity problem*; i. e. to determine the number of neurons required to guarantee that all functions, assumed to belong to a certain function class, can be approximated within a prescribed accuracy ϵ . For this problem upper and lower bounds are of interest. An upper bound for the complexity problem is provided by the result of Barron ([1]). He proved that if the function is assumed to satisfy certain conditions in terms of its Fourier transform, and each of the neurons evaluates a sigmoidal activation function, at most $O(\epsilon^{-2})$ neurons are needed to achieve the order of approximation ϵ . A very common assumption about the function class is defined in terms of the number of derivatives that a function possesses. For example one is interested in approximating all real valued functions of s real variables having a continuous gradient. It is shown in [12] that any reasonable scheme to approximate up to order ϵ any function of this class must depend upon at least $\Omega(\epsilon^{-s})$ parameters. If one restricts to functions having r continuous derivatives then the general results by Devore, Howare and Michelli [12] indicate that the number of neurons is $\Omega(\epsilon^{-\frac{s}{r}})$.

While the complexity problem of approximation by ANN's has been studied in depth for classes of functions satisfying certain conditions on the derivatives, the problem of whether a certain class of integrable functions without assuming differentiable conditions can be optimally approximated is less known. However this last problem is of practical interest since very often target functions coming from real world problems cannot be assumed to possess differentiability properties.

Through out this paper we focus on the complexity problem for L^1 functions, i.e., given a class of non-polynomial activation functions (like for instance the sigmoidal function), determine lower bounds for the number of neurons needed to approximate compact subspaces of Lebesgue measurable functions defined on hypercubes of the kind $[0, 1]^s$ to achieve order of approximation ϵ . More concretely, we provide a general method for estimating the distance between a compact subspace K of $L^1([0, 1]^s)$ and the class of functions computed by ANNs using a single hidden layer, each unit evaluating a sigmoidal activation function. Our lower bounds are stated in terms of an invariant that measures the oscillations of functions of the space K around the origin. As an application we estimate the minimal number of neurons required to approximate bounded continuous functions satisfying an uniform Lipschitz condition of given order α with accuracy ϵ . Our lower bound for these functions is of the kind (using the *Big O* notation for complexity bounds)

$$\Omega \left(\left[\frac{s!}{(s + \alpha) \cdots (1 + \alpha)} \epsilon \right]^{-\frac{s}{\alpha}} \right).$$

The paper is organized as follows. Section 2 describes the mathematical formalization of our approximation problem and the statement of the main results. Section 3 is devoted to explain the main technical details of our method for computing lower bounds. Section 4 and Section 5 sketch the proofs of the main result and their applications.

2 Main Results

Let $s, n \geq 1$ be integers. An artificial network \mathcal{N} with a single hidden layer consisting of n neurons with sigmoidal activation function σ evaluates a function $\Phi : \mathbb{R}^s \rightarrow \mathbb{R}$ of the form

$$\Phi(x) = \sum_{k=1}^n a_k \sigma(w_k \cdot x + b_k), \tag{1}$$

where $x \in \mathbb{R}^s$, the weights w_k are vectors with s real components, the thresholds $b_k \in \mathbb{R}$, the coefficients $a_k \in \mathbb{R}$ for $1 \leq k \leq n$, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, $\sigma(y) = \frac{1}{1+e^{-y}}$, is the activation function and ' \cdot ' denotes the inner product. We shall denote by $\Pi_{\Phi, n, s}$ the set of functions:

$$\Pi_{\Phi, n, s} := \left\{ x \rightarrow \sum_{k=1}^n a_k \sigma(w_k \cdot x + b_k) : a_k, b_k \in \mathbb{R} \ w_k \in \mathbb{R}^s \right\} \tag{2}$$

Let $L^1([0, 1]^s)$ be the normed linear space of Lebesgue measurable functions on the hypercube $[0, 1]^s$. We measure the degree of approximation of $f \in L^1([0, 1]^s)$ by the quantity:

$$\varepsilon_{\Phi, n, s}(f) := \inf_{g \in \Pi_{\Phi, n, s}} \{\|f - g\|\}, \tag{3}$$

where $\|\cdot\|$ denotes the $L^1(\mu)$ -norm in $L^1([0, 1]^s)$, i.e., $\|f\| = \int_{[0, 1]^s} |f| d\mu$, and μ is the Lebesgue measure. Since the target function f is unknown the quantity of interest is the degree of approximation of subspaces K of L by functions of $\Pi_{\Phi, n, s}$, that is:

$$\varepsilon_{\Phi, n, s} := \sup_{f \in K} \varepsilon_{\Phi, n, s}(f) \tag{4}$$

Our search for lower bounds of $\varepsilon_{\Phi, n, s}$ leads to the following definition.

Definition 1. For any compact subset K of $L^1([0, 1]^s)$ and any non-negative integer m , the invariant $j(K, m)$ is defined as the supremum of all real numbers $\delta \geq 0$, such that there are points $x_1, \dots, x_m \in [0, 1]^s$ satisfying:

1. The morphism $eval(f) = (f(x_1), \dots, f(x_m))$, $f \in K$, has S -norm bounded by 1, where S -norm is defined as $\|(x_1, \dots, x_m)\|_S = \sum_{1 \leq i \leq m} |x_i|$.
2. For every sign sequence $s = (s_1, \dots, s_m) \in \{-1, +1\}^m$ there is $f \in K$ such that $s_i f(x_i) \geq \delta$ for $i = 1, \dots, m$.

Next we state our main result.

Theorem 1. There is an integer number $p > 0$, such that for m big enough depending only on n and s and for any compact subset K of the normed linear space $L^1([0, 1]^s)$ the following holds:

$$\varepsilon_{\Phi, n, s}(K) \geq \frac{m}{p} j(K, m) \tag{5}$$

As application of Theorem 1 we have the following result.

Corollary 1. *Let $\|\cdot\|$ denotes the Euclidean norm in $[0, 1]^s$. For $\alpha \in (0, 1]$ let $H_{\alpha, s}$ be the set of real valued functions f on $[0, 1]^s$ bounded by 1, satisfying an uniform Lipschitz condition of order α , i.e., a function $f \in H_{\alpha, s}$ satisfies:*

$$|f(x+t) - f(x)| \leq \|t\|^\alpha \quad (6)$$

for all x and $x+t$ in $[0, 1]^s$. Then in $L^1([0, 1]^s)$ the following holds:

$$\varepsilon_{\Phi, n, s}(H_{\alpha, s}) = \Omega\left(n^{-\frac{\alpha}{s}} \frac{s!}{(s+\alpha)\cdots(1+\alpha)}\right) \quad (7)$$

Remark 1. Note that from Corollary 1 we immediately obtain a lower bound for the minimal number of neurons required to approximate within accuracy ϵ functions in $H_{\alpha, s}$. This bound is of the kind:

$$\Omega\left(\left[\frac{s!}{(s+\alpha)\cdots(1+\alpha)}\epsilon\right]^{-\frac{s}{\alpha}}\right),$$

as it was claimed in the introduction.

3 Partitions of the space of parameters

Let $x_1, \dots, x_m \in [0, 1]^s$ be a set of m points. Let us consider the set of m functions given by:

$$\Phi_i(a, b, w) := \sum_{k=1}^n a_k \sigma(w_k \cdot x_i + b_k), \quad 1 \leq i \leq m, \quad (8)$$

where $a := (a_1, \dots, a_n) \in \mathbb{R}^n$, $b := (b_1, \dots, b_n) \in \mathbb{R}^n$, $w := (w_1, \dots, w_n)$ with $w_k \in \mathbb{R}^s$ for all k , $1 \leq k \leq n$ and σ being the sigmoid function. In what follows for each k , $1 \leq k \leq n$ and for each j , $1 \leq j \leq s$, the j -th component of vector w_k will be denoted by w_{kj} . Similarly the j -th component of point x_i will be denoted by x_{ij} . Our analysis of functions Φ_i relies in the properties of Pfaffian functions. We briefly recall this notion (see [13] for a complete exposition of the theory of Pfaffian functions).

Definition 2. *Let $U \subset \mathbb{R}^n$ be an open domain. A Pfaffian chain of length $q \geq 0$ and degree $D \geq 1$ in U is a sequence of real analytic functions f_1, \dots, f_q in U satisfying differential equations*

$$\frac{\partial f_i}{\partial y_j} = P_{i,j}(y, f_1(y), \dots, f_i(y)) \quad (9)$$

for $i = 1, \dots, q$ where $P_{i,j} \in [y_1, \dots, y_n, z_1, \dots, z_i]$ are polynomials of degree $\leq D$.

Definition 3. A function f on U is called a Pfaffian function of order q and degree (D, d) if

$$f(y) = P(y, f_1(y), \dots, f_q(y)) \quad (10)$$

Proposition 1. Functions $\Phi_i : \mathbb{R}^{n(s+2)} \rightarrow \mathbb{R}$ are Pfaffian of order $q = nm$ and degree $(D, d) = (2, 2)$

Proof. For each k , $1 \leq k \leq n$, for each i , $1 \leq i \leq m$ define the functions $f_{ki} : \mathbb{R}^{n(s+2)} \rightarrow \mathbb{R}$ as follows.

$$f_{ki}(a, b, w) := \sigma(w_k \cdot x_i + b_k), \quad (11)$$

where σ is the sigmoid function. Since the derivative of σ , σ' , satisfies $\sigma' = \sigma(1 - \sigma)$, functions f_{ki} satisfies the following partial differential equations for all k, i with $1 \leq k \leq n$ $1 \leq i \leq m$.

$$\frac{\partial f_{ki}}{\partial a_j} = 0, \quad 1 \leq j \leq n \quad (12)$$

$$\frac{\partial f_{ki}}{\partial b_j} = f_{ki}(1 - f_{ki}), \quad \text{if } k = j \quad (13)$$

and

$$\frac{\partial f_{ki}}{\partial b_j} = 0, \quad \text{if } k \neq j \quad (14)$$

$$\frac{\partial f_{ki}}{\partial w_{kj}} = x_{ij} f_{ki}(1 - f_{ki}), \quad \text{if } 1 \leq j \leq s \quad (15)$$

$$\frac{\partial f_{ki}}{\partial w_{lj}} = 0, \quad \text{if } l \neq k \text{ and } 1 \leq j \leq s \quad (16)$$

It follows from Equations 12, 13, 14, 15 and 16 that the family of nm functions $\{f_{ki}\}_{1 \leq k \leq n, 1 \leq i \leq m}$ is a Pfaffian chain of length $q = nm$ of degree bounded by $D = 2$. Let y_1, \dots, y_n be new variables. For $1 \leq i \leq n$ define de degree two polynomials:

$$P_i(a, b, w, y_1, \dots, y_n) := \sum_{1 \leq k \leq n} a_k y_k \quad (17)$$

It is clear from the definition of functions Φ_i that for each i , with $1 \leq i \leq m$, the following holds.

$$\Phi_i(a, b, w) = P_i(a, b, w, f_{1i}, \dots, f_{ni}) \quad (18)$$

Next, from Equation 18 we conclude that functions Φ_i 's are Pfaffian of length nm and degree $(D, d) = (2, 2)$ as wanted.

3.1 Counting the number of consistent sign assignments to functions Φ_i

A non zero sign assignment to a real valued function f is one of the (in)equalities: $f > 0$, $f < 0$. A non zero sign assignment to a set of m real valued functions is consistent if all m inequalities can be satisfied simultaneously by some assignment of real numbers to the variables. We state here a bound on the number of consistent non zero sign assignments to a finite set of m functions $\{\Phi_i\}_{1 \leq i \leq m}$ as defined in Equation 8.

Lemma 1. *The number of consistent non-zero sign assignments to any subset of m functions $\{\Phi_i\}_{1 \leq i \leq m}$ is bounded above by the quantity:*

$$(4\epsilon m)^{8n^2(s+2)^2} \quad (19)$$

Lemma 1 is now a consequence of Proposition 1 and the following technical statement. The technical details of the proof are omitted.

Theorem 2. *([14]) Suppose $k \geq r$. Let Q_i ($1 \leq i \leq r$) be elements of the polynomial ring $\mathbf{R}[y_1, \dots, y_k, f_1, \dots, f_q]$ of degree at most d where the $f_i(y_1, \dots, y_q)$ form a Pfaffian chain of length q and degree $D \geq 1$. Suppose $(\epsilon_1, \dots, \epsilon_r)$ is a regular value of $Q = (Q_1, \dots, Q_r) : \mathbf{R}^k \rightarrow \mathbf{R}^r$ (i.e. $Q^{-1}(\epsilon_1, \dots, \epsilon_r)$ is a $k - r$ dimensional C^∞ -submanifold of \mathbf{R}^k). Then, the number B of connected components of $Q^{-1}(\epsilon_1, \dots, \epsilon_r)$ satisfies*

$$B \leq 2^{q(q-1)/2} d^k (k(d+D))^k (k^2(d+D))^q \quad (20)$$

Corollary 2. *Let $\Phi_1(a, b, w), \dots, \Phi_m(a, b, w)$ as in Equation 8. There is a polynomial function $g(n, s)$ depending only on n and s and a constant $p > 0$ such that for each $m \geq g(n, s)$ there exists a sign sequence $(s_1, \dots, s_m) \in \{-1, 1\}^m$ which differs in more than m/p places from any sequence of terms $+1, -1$ taken on by sign $\Phi(a, b, w) := (\text{sign } \Phi_1(a, b, w), \dots, \text{sign } \Phi_m(a, b, w))$ when $(a, b, w) \in \mathbb{R}^{(n(s+2))}$.*

Remark 2. *The proof of Corollary 2 is a consequence of Lemma 1, Stirling approximation of $k!$ and some easy combinatorics.*

4 Lower bounds for $\epsilon_{\Phi, n, s}$

This section sketches the proof of Theorem 1. Let m be an integer greater than or equal to $g(n, s)$, where $g(n, s)$ is as in Corollary 2. The case $j(K, m) = 0$ is trivial. Otherwise, let δ be any positive number $< j(K, m)$. Choose real numbers $x_1, \dots, x_m \in [0, 1]^s$ satisfying the conditions of Definition 1, that is:

1. The morphism $eval(f) := (f(x_1), \dots, f(x_m))$ has S -norm bounded by 1.
2. For every sign sequence $s = (s_1, \dots, s_m) \in \{-1, +1\}^m$ there is $f \in K$ such that $s_i f(x_i) \geq \delta$ $i = 1, \dots, m$.

Let Φ as in Equation 1. Consider the functions $\Phi_i(a, b, w)$ defined in Equation 8. By Corollary 2 there exists a sign sequence $s = (s_1, \dots, s_m) \in \{0, 1\}^m$ which differs in more than m/p places from any sequence of terms $+1, -1$ taken on by $\text{sign } \Phi(a, b, w) := (\text{sign } \Phi_1(a, b, w), \dots, \text{sign } \Phi_m(a, b, w))$. Let us take $f \in K$ satisfying condition 2 above. Note that for every $(a, b, w) \in \mathbb{R}^{s(n+2)}$, there is $i \in \{1, \dots, m\}$ such that $|f(x_i) - \Phi_i(a, b, w)| \geq \delta$. Then define $S(f) := (f(x_1), \dots, f(x_m))$. It is straightforward to check that

$$\|S(f) - (\Phi_1(a, b, w), \dots, \Phi_m(a, b, w))\|_S \geq m\delta. \quad (21)$$

Finally, we have:

$$\begin{aligned} \varepsilon_{\Phi, n, s} &\geq \varepsilon_{\Phi, n, s}(f) \geq \varepsilon_{(S(\Phi), n, s)}(S(f)) \geq \\ &\geq \inf_{(a, b, w) \in \mathbb{R}^{s(n+2)}} \|S(f) - (\Phi_1(a, b, w), \dots, \Phi_m(a, b, w))\| \geq m\delta \end{aligned}$$

and Theorem 1 follows.

5 Approximation in the L^1 norm

This section illustrates the application of Theorem 1. We sketch here the proof of Corollary 1. Let r be the least integer such that $r^s \geq g(n, s)$, where $g(n, s)$ is as in Corollary 2. Put $m = r^s$ and divide $I = [0, 1]^s$ into m disjoint cubes K_i each of side $1/r$. Define linear functionals $a_i(f) := \int_{K_i} f d\mu$ where μ is the Lebesgue measure on I . Note that the mapping $a(f) := (a_1(f), \dots, a_m(f))$ has norm one with respect to the S -norm on \mathbb{R}^m . Given any sign sequence (s_1, \dots, s_m) , $s_i \in \{+1, -1\}$, define h on I by $h(x) = s_i(\text{distance}(x, \partial K_i))^\alpha$, for x in K_i , where ∂K_i denotes the boundary of K_i and distance is Euclidean distance. It is easy to check that $h \in H_{\alpha, s}$. For $i = 1, \dots, m$,

$$s_i a_i(h) = \int_{K_i} s_i (\text{distance}(x, \partial K_i))^\alpha = d_\mu \quad (22)$$

$$2^s s! \int_0^{1/2r} \int_0^{y_i} \dots \int_0^{y_{s-1}} y_s^\alpha d_{y_s} \dots d_{y_1} \quad (23)$$

$$m^{-1} r^{-\alpha} 2^{-\alpha} \frac{s!}{(s+\alpha) \dots (1+\alpha)} \quad (24)$$

Thus the invariant $j(K, m)$ introduced in Definition 1 is bounded below by 24. Assuming that $r \geq 2$ Equation 7 of Corollary 1 follows from Theorem 1.

6 Acknowledgments

This work was partially supported by Spanish Grants TIN2007-67466-C02-02, MTM2007-62799 and Spanish FPU program.

References

1. Barron, R.: Universal approximation bounds for superposition of a sigmoidal function. *IEEE Trans. Information Theory* **39** (1993) 930–945
2. Barron, A., Barron, A.L.: Statistical learning networks: a unified view. In: *Symp. on the Interface: Statistics and Computing Science*, Reston, Virginia (1988)
3. Broomhead, D.S., Lowe, D.: Multivariable functional interpolation and adaptative networks. *Complex Systems* **2** (1988) 321–355
4. Cybenko, G.: Approximation by superposition of sigmoidal functions. *Mathematics of Control, signal and Systems* **2** (1989) 303–314
5. Girosi, F., Jones, M., Poggio, T.: Regularization theory and neuronal networks architectures. *Neural Computation* **7** (1995) 219–269
6. Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation* **8** (1996) 164–167
7. Moody, J., Darken, C.: Fast learning in networks of locally tuned processing units. *Neural Computation* **1**(2) (1989) 282–294
8. Park, J., w. Sandberg, I.: Universal approximation using radial basis function networks. *Neural Computation* **3**(2) (1991) 246–257
9. Poggio, T., Girosi, F., Jones, M.: From regularization to radial tensor and additive splines. In: *Proc. Neural Networks for signal Processing, III*, New York, IEEE (1993) 3–10
10. Leshno, M., Lin, V., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neuronal Networks* **2** (1989) 359–366
11. Hornink, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. *Neuronal Networks* **2** (1989) 359–366
12. Devore, R., Howard, R., Micchelli, C.A.: Optimal non-linear approximation. *Manuscripta Mathematica* **63** (1989) 469–478
13. Gabrielov, A.N., Vorobjov, N.: Complexity of computations with Pfaffian and Noetherian functions. *Normal Forms, Bifurcations and Finiteness Problems in Differential Equations*, Kluwer (2004)
14. Khovanskii, A.: On a class of systems of transcendental equations. *Soviet Math. Dokl* **22** (1980) 762–765