Some Comparisons Between Linear Approximation and Approximation by Neural Networks

M. Sanguineti [1], K. Hlaváčková-Schindler [2]
[1]Department of Communications, Computer and System Sciences
DIST - University of Genova, Via Opera Pia 13, 16145, Genova, Italy
E-mail: marce@dist.unige.it
[2] Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07, Praha 8, Czech Republic
E-mail: katka@uivt.cas.cz

Abstract

We present some comparisons between the approximation rates relevant to linear approximators and the rates relevant to neural networks, i.e., nonlinear approximators represented by sets of parametrized functions corresponding to a type of computational unit. Our analysis uses the concept of variation of a function with respect to a set. The comparison is made in terms of Kolmogorov n-width for linear spaces and a proper nonlinear n-width for the nonlinear context represented by neural networks.

The results of this paper contribute to the theoretical understanding of the superiority of neural networks with respect to linear approximators in complex tasks, as is confirmed by a wide variety of applications (recognition of handwritten characters and spoken numerals, approximate solution of functional optimization problems from control theory, etc.).

1 Introduction

Artificial neural networks have greatly outperformed with respect to linear approximators in complex applications such as recognition of handwritten characters and spoken numerals [3], stabilization of high-order strongly nonlinear dynamic systems [12], vocalization of text [15], etc. This performance brings forward the need for theoretical comparison of the approximation capabilities of linear and nonlinear approximation schemes.

The *universal approximation* property has been proved by various authors (see, for example, [4], [7], [13]), in many function spaces and for different types of architectures and activation functions of hidden units (e.g., radial-basis-functions and perceptrons). *Rates of approximation* express the relationship between the accuracy of approximation and the complexity of the approximators required to achieve such an accuracy. The complexity is usually expressed as the size of a properly defined parameter vector: for example, the degree of a polynomial or the number of knots of a fixed-knots spline in the linear case, the degree of a rational function or the number of hidden units of a neural network in the nonlinear context.

To theoretically understand the superior experimental performance of neural networks with respect to linear approximators, it is important to study the comparison of rates by linear and nonlinear approximation schemes in the same functional spaces. What makes this comparison difficult is the fact that each approximator proposed in the literature has been developed to approximate functions from different spaces, i.e., has been obtained under different assumptions on the functions to be approximated. It is then expected that each approximator performs better than the others if such assumptions are satisfied [5]. A better convergence rate for nonlinear approximators with respect to linear ones in the same functional space has been proved by Barron in [1]. However, this comparison is again made only for functions belonging to a particular space and for a specific class of nonlinear approximators.

In [9] Kůrková has defined a norm, called variation of a function with respect to a set of functions, which extends a concept introduced by Barron in [2]. Such a norm is assigned to a given class of networks and allows the comparison of rates of convergence within a common framework (see [9], [10]). In [6] we have used this norm as a tool for comparison of the optimal bounds on the approximation error achievable by nonlinear approximators in certain spaces of functions of finite-dimensional Hilbert spaces and the rates obtained in the same spaces by linear approximators.

In this paper, we further develop the comparisons started in [6]. Using the variation norm, we show that, in some functional spaces, lower bounds on the rates of linear approximation are greater than upper bounds on the rates achievable by nonlinear approximators represented by neural networks (i.e., sets of parametrized functions corresponding to a type of computational unit).

The organization of the paper is the following. Section 2 contains preliminary notations and definitions. Section 3 presents the comparison of the optimal bounds and some final remarks are concluded in Section 4.

2 Preliminary Notations and Definitions

The following notations and definitions will be used (see also [8] and [10]). We assume to work in a normed linear space $(\mathcal{X}, \|.\|)$; $\|.\|_2$ denotes the norm induced by the inner product in case of a Hilbert space.

The approximation is called *linear approximation*, when the approximating functions form a linear subspace of $(\mathcal{X}, \|.\|)$. On the contrary, the approximating functions can be members of unions of finite-dimensional subspaces generated by a given computational unit. In other words, $\mathcal{G} = \{g(.,\theta) : Y \to \mathcal{R}; \theta \in \Theta \subset \mathcal{R}^p\} \subset (\mathcal{X}, \|.\|)$, $Y \subset \mathcal{R}^d$, is a parametrized set of functions corresponding to the computational unit represented by the (activation) function g. The set of all linear combinations of n elements of $\mathcal{G}$ is considered. This set, denoted by $span_n\mathcal{G}$, is the union of all linear subspaces formed (spanned) by n-tuples of elements of $\mathcal{G}$, i.e., $span_n\mathcal{G} := \{f \in \mathcal{X}; f = \sum_{i=1}^n w_i g_i; w_i \in \mathcal{R}, g_i \in \mathcal{G}\} = \bigcup\{span\{g_1, \ldots, g_n\}; g_i \in \mathcal{G}, i = 1, \ldots, n\}$. In this case, the approximation is called *nonlinear approximation*. Note that $span\,\mathcal{G} = \bigcup_{n \in \mathcal{N}} span_n\mathcal{G}$. $\mathcal{G}^0$ denotes the set of normalized element of a given set $\mathcal{G}$, i.e. $\mathcal{G}^0 = \{g^0 = \frac{g}{\|g\|}, g \in \mathcal{G}\}$.

Let $\mathcal{G}(b) := \{wg; w \in \mathcal{R}, |w| \leq b, g \in \mathcal{G}\}$. For a subset $\mathcal{G}$ of a normed linear space $(\mathcal{X}, \|.\|)$, *$\mathcal{G}$-variation* of $f \in \mathcal{X}$ is

$$\|f\|_{\mathcal{G}} := \inf\{b > 0; f \in cl\,conv\,\mathcal{G}(b)\}$$

where the notation is motivated by the fact that $\mathcal{G}$-variation is a norm on $\{f \in \mathcal{X}; \|f\|_{\mathcal{G}} < \infty\}$ [9]. Although in general the concept of $\mathcal{G}$-variation depends on the choice of the norm, to simplify the notation we write $\|f\|_{\mathcal{G}}$ instead of $\|f\|_{(\mathcal{G},\|.\|)}$ (note that when $\mathcal{X}$ is finite-dimensional, all norms on it are equivalent, hence in such a case $\mathcal{G}$-variation does not depend on $\|.\|$).

For a subset $\mathcal{S}$ of $(\mathcal{X}, \|.\|)$, the $n-$width in the sense of Kolmogorov (or the *Kolmogorov n-width*) of $\mathcal{S}$ in $\mathcal{X}$ [14] is

$$d_n(\mathcal{S}, \mathcal{X}) := \inf_{\mathcal{X}_n} d(\mathcal{S}, \mathcal{X}_n) = \inf_{\mathcal{X}_n} \sup_{f \in \mathcal{S}} \inf_{h \in \mathcal{X}_n} \|f - h\|$$

where the left-most infimum is taken over all n-dimensional subspaces $\mathcal{X}_n$ of $\mathcal{X}$ and

$$d(\mathcal{S}, \mathcal{Y}) := \sup_{f \in \mathcal{S}} \|f - \mathcal{Y}\|$$

Nonlinear n-width of $\mathcal{S}$ in $\mathcal{X}$ is defined as

$$\delta_n(\mathcal{S}, \mathcal{X}) :=$$

$$\inf_{\mathcal{G}} d(\mathcal{S}, span_n\mathcal{G}) = \inf_{\mathcal{G}} \sup_{f \in \mathcal{S}} \inf_{h \in span_n\mathcal{G}} \|f - h\|$$

where $\mathcal{G}$ is a member of a family of parametrized subsets of $\mathcal{X}$. Nonlinear n-width as the alternative to the Kolmogorov linear $n-$width for nonlinear approximation was first suggested in [8].

We finally denote

$$d_n(f, \mathcal{X}) := \inf_{\mathcal{X}_n} \inf_{h \in \mathcal{X}_n} \|f - h\|$$

and

$$d(f, span_n\mathcal{G}) := \inf_{h \in span_n\mathcal{G}} \|f - h\|$$

that correspond to $d_n(\mathcal{S}, \mathcal{X})$ and $d(\mathcal{S}, span_n\mathcal{G})$, respectively, for $\mathcal{S} = \{f\}$.

3 Comparison of Bounds on Approximation Rates

The following is a reformulation of Jones-Barron's theorem in terms of $\mathcal{G}^0$–variation:

Theorem 1 *[10] Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space, $\mathcal{G}$ be its subset. Then for every $f \in \mathcal{X}$ and for every positive integer n*

$$\|f - span_n\mathcal{G}\|_2^2 \leq \frac{\|f\|_{\mathcal{G}^0}^2 - \|f\|_2^2}{n}.$$

It is easy to see that $\|f\|_2 \leq \|f\|_{\mathcal{G}^0}$ holds for every $f \in (\mathcal{X}, \|.\|_2)$, i.e. the unit ball of $\mathcal{G}^0$-variation is contained in the unit ball of $\|.\|$. Then there exists a constant $c_{f,\mathcal{G}} \geq 1$ such that $c_{f,\mathcal{G}}\|f\|_2 = \|f\|_{\mathcal{G}^0}$. We then get the following corollary:

Corollary 1 *Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space and $\mathcal{G}$ its subset. Then for every $f \in \mathcal{X}$ there exists a constant $c_{f,\mathcal{G}} \geq 1$ such that for every positive integer n*

$$\|f - span_n\mathcal{G}\|_2^2 \leq \|f\|_2^2 \frac{c_{f,\mathcal{G}}^2 - 1}{n}.$$

It follows that Theorem 1 gives a "good" upper estimate for the class of functions $F_\epsilon := \{f : c_{f,\mathcal{G}}\|f\|_2 = \|f\|_{\mathcal{G}^0}, 1 \leq c_{f,\mathcal{G}} \leq 1 + \epsilon\}$, $\epsilon > 0$.

Let us now consider the following result from Pinkus:

Theorem 2 *[14] Let S_n be the unit ball of any $(n+1)$-dimensional subspace $\mathcal{X}_{n+1}$ of a normed linear space $(\mathcal{X}, \|.\|)$. Then*

$$d_k(S_n, \mathcal{X}) = 1, \; k = 0, 1, \ldots, n.$$

In the following, the unit ball of an $(n+1)$-dimensional subspace $\mathcal{X}_{n+1}$ of a normed linear space $(\mathcal{X}, \|.\|)$ will be denoted by S_n. Let $B_r(\|.\|)$ be the ball of radius r in the metric $\|.\|$, i.e.

$$B_r(\|.\|) := \{f \in \mathcal{X}; \|f\| \leq r\}.$$

Then, $B_r(\|.\|_{\mathcal{G}^0})$ is the ball of radius r in $\mathcal{G}^0$-variation. It follows from Theorem 1 that the approximation error ε in the case of nonlinear approximation by the parametrized family $\mathcal{G}$ is

$$\varepsilon^2 \leq \frac{\|f\|_{\mathcal{G}^0}^2 - \|f\|_2^2}{n}.$$

On the other hand if, for a given $r \in \Re^+$, there exists $n \in \mathcal{N}$ such that $B_r(\|.\|_{\mathcal{G}^0}) \supseteq S_n$, then from Theorem 2 we get

$$d_n(B_r(\|.\|_{\mathcal{G}^0}), \mathcal{X}) \geq 1$$

Since a sufficient condition for $\frac{\|f\|_{\mathcal{G}^0}^2 - \|f\|_2^2}{n} \leq 1$ is $r \leq \sqrt{n}$ in $B_r(\|.\|_{\mathcal{G}^0})$, we obtain the following proposition:

Proposition 1 *Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space, $\mathcal{G}$ its subset and $n \in \mathcal{N}$ such that $B_{\sqrt{n}}(\|.\|_{\mathcal{G}^0}) := \{f \in \mathcal{X}; \|f\|_{\mathcal{G}^0} \leq \sqrt{n}\} \supseteq S_n$. Then the upper bound on $d(B_{\sqrt{n}}(\|.\|_{\mathcal{G}^0}), span_n\mathcal{G})$ is less than the lower bound on $d_n(B_{\sqrt{n}}(\|.\|_{\mathcal{G}^0}), \mathcal{X})$.*

Now we focus on Kolmogorov n-width. We will use the following characterization of Kolmogorov n-width.

Theorem 3 *[14] If K is a closed, convex, centrally symmetric proper subset of an $(n+1)$-dimensional subspace $\mathcal{X}_{n+1}$ of a normed linear space $(\mathcal{X}, \|.\|)$ and δK denotes the boundary of K, then*

$$d_n(K, \mathcal{X}) = inf\{\|f\| : f \in \delta K\}.$$

Note that, based on the properties of the Kolmogorov n-width [14], if $\mathcal{K} \subset (\mathcal{X}, \|.\|)$, and K is a centrally symmetric set created from the closure of the convex hull of $\mathcal{K}$, then

$$d_n(\mathcal{K}, \mathcal{X}) = d_n(K, \mathcal{X}).$$

Given $f \in \mathcal{X}$, it follows from the definition of $\mathcal{G}$-variation that $f \in clconv(\mathcal{G}(\|f\|_{\mathcal{G}})) = B_{\|f\|_{\mathcal{G}}}(\|.\|_{\mathcal{G}})$. If we denote $b := \sup_{g \in \mathcal{G}} \|g\|$, $B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}) := B_{\|f\|_{\mathcal{G}}}(\|.\|_{\mathcal{G}}) \bigcap \mathcal{X}_{n+1}$ and apply Theorem 3 with $K = B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}})$, we get

$$\|f\|_{\mathcal{G}^0} \leq b\|f\|_{\mathcal{G}} = b\, d_n(B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}), \mathcal{X})$$

and, using Theorem 1:

$$\|f - span_n\mathcal{G}\|_2 \leq \sqrt{\frac{b^2\, d_n(B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}), \mathcal{X})^2 - \|f\|_2^2}{n}}$$

$$\leq b \frac{d_n(B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}), \mathcal{X})}{\sqrt{n}}.$$

This is concluded in the following proposition:

Proposition 2 *Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space, $\mathcal{G}$ its subset, $b := \sup_{g \in \mathcal{G}} \|g\|$ and $f \in \mathcal{X}$. Moreover, let $B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}) := B_{\|f\|_{\mathcal{G}}}(\|.\|_{\mathcal{G}}) \bigcap \mathcal{X}_{n+1}$, where $\mathcal{X}_{n+1}$ is an $(n+1)$-dimensional subspace of $(\mathcal{X}, \|.\|_2)$. Then*

$$d(B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}), span_n\mathcal{G}) \leq b \frac{d_n(B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}}), \mathcal{X})}{\sqrt{n}}.$$

In other words, the upper bound on nonlinear approximation by a parametrized set $\mathcal{G}$ of functions corresponding to a type of computational unit is better at least for a multiplicative factor $\frac{1}{\sqrt{n}}$ than the upper bound on linear approximation in the set $B_{\|f\|_{\mathcal{G}}}^{n+1}(\|.\|_{\mathcal{G}})$. It follows that linear approximation of functions in the intersection of $\mathcal{G}$-balls whith $(n+1)$-dimensional subspaces, is a weak tool in comparison to nonlinear approximation for $n >> 1$.

Now, suppose that $(\mathcal{X}, \|.\|)$ is a normed functional space defined on $\mathcal{R}^d$ and that linear approximators in $B^{n+1}_{\|f\|_{\mathcal{G}}}(\|.\|_{\mathcal{G}})$ suffer of the *curse of dimensionality*, i.e. the number of parameters necessary to achieve a given accuracy increases exponentially with increasing dimension d. This is expressed by a factor of the form Cn^d in the approximation rate, where C is a constant with respect to n. Note that Proposition 2 does not a priori imply that neural networks corresponding to the parametrized set of functions $\mathcal{G}$ avoid this problem, since the multiplicative factor $\frac{1}{\sqrt{n}}$ can not cope with an exponential term.

For an orthonormal basis $\mathcal{A}$ of a finite-dimensional Hilbert space $(\mathcal{X}, \|.\|_2)$ we denote the *l_1-norm with respect to $\mathcal{A}$* by $\|.\|_{1,\mathcal{A}}$, i.e. $\|f\|_{1,\mathcal{A}} = \sum_{i=1}^m |w_i|$, where $f = \sum_{i=1}^m w_i g_i$. It easy to verify that, for every $f \in \mathcal{X}$, $\|f\|_{\mathcal{A}} = \|f\|_{1,\mathcal{A}}$, i.e. $\mathcal{A}$-variation is the l_1-norm with respect to $\mathcal{A}$ [10]. The following theorem holds.

Theorem 4 *[10] Let $(\mathcal{X}, \|.\|_2)$ be a finite-dimensional Hilbert space and $\mathcal{A}$ its orthonormal basis. Then for every $f \in \mathcal{X}$ and for every positive integer n there exists $f_n \in span_n\mathcal{A}$ such that*

$$\|f - f_n\|_2 \leq \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n}}.$$

This implies that $\forall \mathcal{S} \subset \mathcal{X}$, $d(\mathcal{S}, span_n\mathcal{A}) \leq \sup_{f\in\mathcal{S}} \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n}}$. If the only information available about f is the value of its $\mathcal{A}$-variation, then this upper bound can not be improved [10]. However, the upper bound in Theorem 4 can be improved if in addition to $\|f\|_{1,\mathcal{A}}$ also $\|f\|_2$ is known [10].

We are now interested in approximating functions from the unit ball S_n of an $(n+1)$-dimensional subspace $\mathcal{X}_{n+1}$ of a Hilbert space $(\mathcal{X}, \|.\|_2)$. It follows from Theorem 4 that

$$\sup_{\|f\|_2=1} \{\|f - span_n\mathcal{A}\|_2\} \leq \sup_{\|f\|_2=1} \left\{\frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n}}\right\}.$$

If $\mathcal{A} = \mathcal{E}_l$ (the Euclidean basis of $\mathcal{X}_l$), where $l \geq n$, we get

$$\sup\{\|f\|_{1,\mathcal{E}_l} : \|f\|_2 = 1\} =$$
$$\max\{\|f\|_{1,\mathcal{E}_l} : \|f\|_2 = 1\} =$$
$$\max\left\{\sum_{i=1}^{l} |f_i| : \|f\|_2 = 1\right\} = \sqrt{l}.$$

Then

$$d(S_{l-1}, span_n\mathcal{E}_l) = \sup_{f\in S_n} \inf_{g\in span_n\mathcal{E}_l} \|f-g\|_2 \leq \frac{1}{2}\sqrt{\frac{l}{n}}.$$

If we use this for $l = n+1$, we get

$$d(S_n, span_n\mathcal{E}_{n+1}) \leq \frac{1}{2}\sqrt{1+\frac{1}{n}}.$$

On the other hand, we know from Theorem 2 that

$$d_n(S_n, \mathcal{X}) = \inf_{X_n} \sup_{f\in S_n} \inf_{g\in X_n} \|f-g\|_2 = 1.$$

The above results can be summarized in the following proposition.

Proposition 3 *Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space and $\mathcal{E}_{n+1}$ the Euclidean basis of its $(n+1)$-dimensional subspace $\mathcal{X}_{n+1}$. Then the upper bound on $d(S_n, span_n\mathcal{E}_{n+1})$ is less than $d_n(S_n, \mathcal{X})$.*

In other words, the upper bound on nonlinear approximation by $span_n\mathcal{E}_{n+1}$ of the unit ball S_n of $\mathcal{X}_{n+1}$ is better than the upper bound on linear approximation of the same ball for $n \geq 1$.

We finally make some remarks on the approximation of a single function. The above results on approximation are too general for this case and they need not provide the optimal rates for approximation of a single function. We compare them with the rates of approximation in Hilbert spaces achieved in [11].

Let $(\mathcal{X}, \|.\|_2)$ be a separable Hilbert space with $\mathcal{G}$ an orthogonal basis of $\mathcal{X}$. Then every $f \in \mathcal{X}$ can be written in the form $f = \sum_{k=1}^{\infty} a_k(f) g_k$, where the series converges in the norm of $\mathcal{X}$. Define $\mathcal{G} := \{g_k; k = 1, 2, \ldots\}$ and $S_{\mathcal{G}} = \{f \in \mathcal{X}; \sum_{k=1}^{\infty} |a_k(f)| \leq 1\}$. Let $\Lambda \subset Z_0^+$ (Z_0^+ represents the set of non-negative integers) and $\mathcal{U}_\Lambda := span\{g_k, k \in \Lambda\}$. Let $\mathcal{T}_\Lambda$ denote the projection operator on $\mathcal{U}_\Lambda$ and $\mathcal{C}_\Lambda(f) := \inf_{g\in\mathcal{U}_\Lambda} \|f - g\|, g \in \mathcal{G}$. It holds that $\mathcal{C}_\Lambda(f) = \|f - \mathcal{T}_\Lambda(f)\|$. Denote

$$\Delta_n(S_{\mathcal{G}}, \mathcal{X}) = \sup_{f\in S_{\mathcal{G}}} \inf_{\Lambda\subset Z_0^+, |\Lambda|\leq n} \mathcal{C}_\Lambda(f),\ n = 1, 2, \ldots.$$

We deal only with infinite dimensional Hilbert spaces here as for $\mathcal{X}$ having a finite dimension we get the rate $\Delta_n(S_{\mathcal{G}}, \mathcal{X}) = 0$. The following result holds [11]:

Theorem 5

$$\Delta_n(S_{\mathcal{G}}, \mathcal{X}) \leq \frac{1}{\sqrt{n+1}},\ n = 1, 2, \ldots$$

Moreover, if $f \in \mathcal{S}_{\mathcal{G}}$ *then there is a sequence*$\{\rho_n\}$ *of numbers such that* $\rho_n \in (0,2]$, $n = 1,2,\ldots$, $\lim_{n\to\infty} \rho_n = 0$ *and*

$$\inf_{\Lambda \subset Z_0^+, |\Lambda| \le n} \mathcal{C}_\Lambda(f) \le \frac{\rho_n}{\sqrt{n}},\ n = 1,2,\ldots$$

Then we get for $f \in \mathcal{S}_{\mathcal{G}}$:

- $\Delta_n(f,\mathcal{X}) = \inf_{\Lambda \subset Z_0^+, |\Lambda| \le n} \mathcal{C}_\Lambda(f) = \inf_{\mathcal{X}_n} \inf_{g\in\mathcal{X}} \|f-g\| = d_n(f,\mathcal{X})$
- Linear approximation: $d_n(f,\mathcal{X}) = \inf_{\mathcal{X}_n} \inf_{g\in\mathcal{X}_n} \|f-g\|$, where $\mathcal{X}_n$ is an n-dimensional subspace of $\mathcal{X}$
- Nonlinear approximation: $d(f, span_n\mathcal{G}) = \inf_{g\in span_n\mathcal{G}} \|f-g\|$

As for every $\mathcal{X}_n \subset \mathcal{X}$, also $\mathcal{X}_n \subset span_n\mathcal{G}$, $d(f, span_n\mathcal{G})$ is infimum over a 'bigger' set than in the case of $d_n(f,\mathcal{X})$. So it follows that, for an orthonormal basis $\mathcal{G}$ of $(\mathcal{X}, \|.\|_2)$, we have

$$d(f, span_n\mathcal{G}) \le d_n(f,\mathcal{X}) = \Delta_n(f,\mathcal{X}) \le \frac{\rho_n}{\sqrt{n}}$$

and $\lim_{n\to\infty} \rho_n = 0$.

For functions $f \in \mathcal{S}_n \bigcap \mathcal{S}_{\mathcal{G}}$, this is a better bound than the one given in Theorem 2, i.e. $d_n(f,\mathcal{X}) = 1$, especially for $n >> 1$.

4 Concluding Remarks

A theoretical framework for the comparison of linear approximators and nonlinear approximation schemes is necessary for understanding of the experimental outperformance of neural networks with respect to traditional linear approximators. In this paper, we have shown the superiority of neural network approximation in some functional spaces. The variation of a function with respect to a set and a proper nonlinear n-width, analogous to the Kolmogorov n-width for the linear case, play a key-role in this analysis.

The proposed results provide further theoretical understanding of the surprising success of neural networks in complex approximation tasks (e.g, vocalization of text, optimization problems from control theory, etc.). The analysis in more general functional spaces is of high importance and is still an open problem.

Acknowledgements

We are grateful to Věra Kůrková for valuable discussions and remarks.

Marcello Sanguineti was partially supported by grants CNR 96.02472.CT07 and CNR 97.00048.PF42. Kateřina Hlaváčková–Schindler was partially supported by grants CR 201/96/0917 and 201/99/0092.

References

[1] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory 39, pp. 930-945, 1993.

[2] Barron, A. R.: Neural net approximation. Proc. 7th Yale Workshop on Adaptive and Learning Systems. K. Narendra Ed., Yale University Press, 1992.

[3] Burr, D.J.: Experiments on neural net recognition of spoken and written text. IEEE Trans. Acoust. Speech and Signal Processing 36, pp. 1162-1168, 1988.

[4] Cybenko, G.: Approximation by superposition of a sigmoidal function, Math. Control Signal Systems 2, pp. 303-314, 1989.

[5] Girosi, F., Jones, M. and Poggio, T.: Regularization theory and neural networks architectures. Neural Computation 7, pp. 219-269, 1995.

[6] Hlaváčková, K., Sanguineti, M.: On the rates of linear and nonlinear approximations. Proc. 3rd IEEE European Workshop on Computer-Intensive Methods in Control and Signal Processing (CMP), pp. 211-216, 1998.

[7] Hornik, K., Stinchcombe, M., White H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, pp. 359-366, 1989.

[8] Kainen, P.C., Kůrková, V., Vogt, A.: Approximation by neural networks is not continuous. Submitted to Neurocomputing.

[9] Kůrková, V.: Dimension-independent rates of approximation by neural networks. Computer-intensive methods in Control and Signal Processing: Curse of Dimensionality (Eds. K. War-

wick, M. Kárný). Birkhäuser, Boston, pp. 261-270, 1997.

[10] Kůrková, V., Savický, P., Hlaváčková, K.: Representations and rates of approximation of real–valued Boolean functions by neural networks. Neural Networks 11, pp. 651-659, 1998.

[11] Mhaskar, H.N., Micchelli, C.A.: Dimension-independent bounds on the degree of approximation by neural networks. IBM Journal of Research and Development 38, pp. 277-284, 1994.

[12] Parisini, T., Sanguineti, M., Zoppoli, R.: Nonlinear stabilization by receding-horizon neural regulators. International Journal of Control 70, no.3, pp. 341-362, 1998.

[13] Park J., Sandberg, I. W.: Approximation and radial-basis-function networks. Neural Computation 5, pp. 305–316, 1993.

[14] Pinkus, A.: $N-$Widths in Approximation Theory. Springer-Verlag, New York, 1986.

[15] Sejnowski, T.J., Rosenberg, C.: Parallel networks that learn to pronounce English text. Complex Systems 1, pp. 145-168, 1987.

Andrej Dobnikar
Nigel C. Steele
David W. Pearson
Rudolf F. Albrecht

Artificial Neural Nets
and Genetic Algorithms

Proceedings of the International Conference
in Portorož, Slovenia, 1999

SpringerWienNewYork

Andrej Dobnikar
Nigel C. Steele
David W. Pearson
Rudolf F. Albrecht (eds.)

Artificial Neural Nets and Genetic Algorithms

SpringerComputerScience

SpringerWienNewYork