

# Neural networks for functional approximation and system identification

H. N. Mhaskar\*  
Department of Mathematics,  
California State University  
Los Angeles, CA 90032, U.S.A.

Nahnwoo Hahm  
Department of Mathematics  
University of Texas at Austin  
Austin, TX 78712, U.S.A.

December, 1995

## Abstract

We construct generalized translation networks to uniformly approximate a class of nonlinear, continuous functionals defined on  $L^p([-1, 1]^s)$  for integer  $s \geq 1$ ,  $1 \leq p < \infty$  or  $C([-1, 1]^s)$ . We obtain lower bounds on the possible order of approximation for such functionals in terms of *any* approximation process depending continuously upon a given number of parameters. Our networks almost achieve this order of approximation in terms of the number of parameters (neurons) involved in the network. The training is simple and noniterative; in particular, we avoid any optimization such as that involved in the usual back-propagation.

## 1 Introduction

One of the important applications of neural networks is to approximate functions defined on a compact subset of a Euclidean space in a highly parallel manner. (We give precise definitions in Section 2.) Following the well known initial results of Cybenko [5], Hornik, White and Stinchcombe [11], Park and Sandberg [23], and Barron [1] among others, there has been a great deal of activity in this area in recent years. Motivated by the work of Girosi, Jones, and Poggio [8], Mhaskar and Micchelli [18] have introduced the notion of generalized translation networks, unifying the theme of neural networks, radial basis function networks, and generalized regularization networks of [8]. They studied in detail how the choice of the activation function affects the degree of approximation of the target function. Subsequently, the first named author [17] has constructed generalized translation networks utilizing specific activation functions to obtain an optimal order of approximation when the only a priori information about the target function is that it belongs to some Sobolev class.

In this paper, we obtain analogous results when a generalized translation network is used to approximate functionals on compact subsets of an  $L^p$  space, rather than functions of finitely many real variables. The problem of approximating a functional arises naturally in the study of identification and control of nonlinear systems. In an identification problem, the hidden states of a nonlinear system are not known, and one wishes to develop a model merely by observing the input-output relationships of such a system (cf. Levin and Narendra [12] for a recent discussion). The output at any given time is a functional of the input signal, which is a function of time. Similarly, a control problem can be thought of as a functional, which, at any given time, maps the input control signals to an output signal. The actual system itself is often unknown, and a simple model is desirable (cf. Sontag [28] for a recent review).

Narendra and Parthasarathy [20] have proposed different models which can utilize the approximation capabilities of neural networks for system identification and control. Levin and Narendra [12] have studied neural networks for the identification of systems using only observations on the input/output relationships

---

\*This research was supported, in part, by National Science Foundation Grant DMS 9404513 and Air Force Office of Scientific Research Grant F49620-93-1-0150. **Published:** *Neural Computation*, **9** (1997), 143–159.

of the system. In both of these papers, the output signal at any time is considered as a function of finitely many samples of the input and output signals, typically taken prior to the moment in question.

A different approach has been suggested by Sandberg [26], where the system is thought of directly as a functional acting on the input signal, and an “ $NL$ ”-system is used as a model for this functional. Here,  $N$  is a static neural network,  $L$  has a simple representation in terms of bounded linear functionals, and the model is the composite map  $N \circ L$ . Sandberg has shown that any real continuous functional on any compact subset of a real normed linear space can be uniformly approximated in this way, and has given several related results concerning, for example, the choice of the linear functional  $L$  that can be used. He has also studied the case when radial (or elliptic) basis function networks are used instead of the conventional neural networks. Further research in this direction is also done by Chen and Chen [2, 3, 4], Dingankar [7], and Modha and Hecht-Nielsen [19].

In this paper, we investigate the question of obtaining bounds on the size of the networks involved in terms of the desired accuracy of approximation. As expected, the bound will depend upon the properties of the compact subset of the function space on which the functional acts, and also on the structural properties of the functional itself. Typically, both of these are likely to be unknown. Therefore, one is led to the notion of universal approximation; i.e., approximating the functional under minimal a priori assumptions on the functional and the input signals on which it acts. We prove that there are some inherent lower bounds for the accuracy of such universal approximation. We will also construct generalized translation networks which “almost” achieve this lower bound. The networks to be constructed are very general, and include as special cases, Gaussian networks, thin plate spline networks, neural networks with the hyperbolic tangent activation function, and a variety of other commonly used networks. Although our focus is to obtain a theoretical insight into the problem, our proof suggests an explicit formula for the networks. There is no training involved in the usual sense. In particular, we do not use any optimization technique, such as minimization of a back-propagation error surface. Thus, the large number of neurons required for our networks is offset by the extremely simple “training”, and we do not encounter any problems commonly associated with optimization, such as local minima.

In the next section, we state and discuss our results. The proofs are given in Section 3. We are grateful to Professor Dr. I. Sandberg and one of the referees for clarifying the history of this problem, as well as to both the referees for their valuable suggestions for the improvement of the presentation in this paper.

## 2 Main Results

Following the approach in the papers [26], [25] of Sandberg, we think of an arbitrary nonlinear system as a continuous functional  $F$  acting on a compact subset of some  $L^p$  space. Let  $s \geq 1$  be an integer, which will be fixed throughout the rest of this paper. If  $A \subset \mathbb{R}^s$  is a (Lebesgue-) measurable set, and  $f : A \rightarrow \mathbb{R}$  is a measurable function, we write

$$\|f\|_{p,A} := \begin{cases} \left\{ \int_A |f(\mathbf{t})|^p d\mathbf{t} \right\}^{1/p}, & \text{if } 1 \leq p < \infty, \\ \text{ess sup}_{\mathbf{t} \in A} |f(\mathbf{t})|, & \text{if } p = \infty. \end{cases} \quad (2.1)$$

The space  $L^p(A)$  is defined to be the class of all functions  $f : A \rightarrow \mathbb{R}$  for which  $\|f\|_{p,A} < \infty$ . As usual, we consider two functions to be equal if they are equal almost everywhere in the measure theoretic sense. In this paper, if  $A = [-1, 1]^s$ , then we will omit the mention of this set from the notation; thus, we write  $\|f\|_p$  to denote  $\|f\|_{p,[-1,1]^s}$  etc. Further, we will have no occasion to consider functions in  $L^\infty(A)$  which are not continuous. Therefore, we will simplify our notations by using the symbol  $L^\infty(A)$  to denote also the class  $C(A)$ , consisting of bounded, uniformly continuous functions on  $A$ .

We also recall a few facts about compact subsets of  $L^p$ . There are several known characterizations of compact sets in  $L^p(A)$  ([27]); for illustrating our theorems, we find the following characterization useful, where we restrict our attention to the case when  $A = [-1, 1]^s$ . For any real number  $\lambda > 0$ , we denote the class of all algebraic polynomials in  $s$  variables of coordinatewise degree not exceeding  $\lambda$  by  $\Pi_{\lambda,s}$ . If

$f \in L^p$  and  $\lambda \geq 0$  we write

$$E_{\lambda,p,s}(f) := \inf_{P \in \Pi_{\lambda,s}} \|f - P\|_p. \quad (2.2)$$

It is not too difficult to verify (cf. [14], p. 33) that a closed set  $K \subset L^p$  is compact if and only if there is a constant  $M_K > 0$  and a sequence of positive numbers  $\{\epsilon_{m,K}\}$ , converging to 0 as  $m \rightarrow \infty$ , such that for every  $f \in K$ ,

$$\|f\|_p \leq M_K, \quad E_{m,p,s}(f) \leq \epsilon_{m,K}, \quad m = 0, 1, \dots \quad (2.3)$$

For example, for the set of functions satisfying a Hölder condition of order  $\alpha$ , one may choose  $\epsilon_{m,K} = cm^{-\alpha}$  for a suitable constant  $c$ .

Given a system defined by a functional  $F$  on a compact set  $K \subset L^p$ , the objective of system identification is to build another functional  $G$  on this compact set which will approximate  $F$  well, and also serve as a model with known parameters. Although the whole problem arises because  $F$  is typically unknown, it is reasonable to make some a priori assumptions on a class of functionals to which  $F$  may belong; as is done commonly in the theory of approximation of functions. Moreover, if we wish to consider the system not as a functional, but as an operator  $T$  producing an output signal  $Tf$  upon receiving an input signal  $f$ , this operator can be thought of as a set of functionals: each  $t$  in the domain of the output signal defines a functional  $Tf(t)$ . A model for this operator will achieve uniform approximation of all these functionals. Thus, in this paper, we are interested in approximating every functional belonging to a class  $\mathcal{F}$ , satisfying some minimal conditions, which we now describe.

We recall that a function  $\Omega : (0, \infty) \rightarrow (0, \infty)$  is said to be a *modulus of continuity* ([13]) if each of the following properties holds. (a)  $\Omega(h) \rightarrow 0$  as  $h \rightarrow 0$ , (b)  $\Omega$  is a positive and increasing function, and (c)  $\Omega$  is subadditive; i.e.,

$$\Omega(h_1 + h_2) \leq \Omega(h_1) + \Omega(h_2), \quad h_1, h_2 > 0. \quad (2.4)$$

For example, for any  $\alpha \in (0, 1)$ , the function  $\Omega(h) = h^\alpha$  is a modulus of continuity. For any real function  $F$  uniformly continuous on  $L^p$ , the quantity  $\sup\{|F(f) - F(g)| : \|f - g\|_p \leq \delta\}$ , considered as a function of  $\delta$ , is a modulus of continuity (of  $F$ ). More generally, if  $X$  is a topological space, and  $T : L^p \rightarrow C(X)$  is a bounded, continuous function, then the quantity

$$\sup\{|Tf(x) - Tg(x)| : x \in X, \|f - g\|_p \leq \delta\},$$

considered as a function of  $\delta$ , is modulus of continuity (of  $T$ ). For each  $x \in X$ , the modulus of continuity of  $T$  is a *majorant* for the modulus of continuity of the functional  $f \rightarrow Tf(x)$ . Motivated by this observation, and the Jackson theorem in the theory of trigonometric approximation, the only assumption we wish to make about the functional is that there be a known majorant for its modulus of continuity. Accordingly, we will be interested in this paper in the universal approximation of the class

$$\mathcal{F} := \mathcal{F}_{\Omega,p,s} := \{F : L^p \rightarrow \mathbb{R} : |F(f) - F(g)| \leq \Omega(\|f - g\|_p), f, g \in L^p\}, \quad (2.5)$$

where  $\Omega$  is a given modulus of continuity.

Our main objective in this paper is to construct generalized translation networks to model every functional in  $\mathcal{F}$ . We now proceed to describe these networks. Let  $1 \leq d \leq D$ ,  $N \geq 1$  be integers, and  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ . A *generalized translation network* with  $N$  neurons evaluates a function of the form  $\sum_{k=1}^N a_k \phi(A_k(\cdot) + \mathbf{b}_k)$  where the *weights*  $A_k$ 's are  $d \times D$  real matrices, the *thresholds*  $\mathbf{b}_k \in \mathbb{R}^d$  and the *coefficients*  $a_k \in \mathbb{R}$  ( $1 \leq k \leq N$ ). The set of all such functions (with a fixed  $N$ ) will be denoted by  $\Pi_{\phi;N,D}$ . When  $d = 1$ ,  $\Pi_{\phi;N,D}$  is the set of all outputs of a neural network with  $N$  neurons, each evaluating the activation function  $\phi$ , and receiving  $D$  inputs. When  $d = D$ , and  $\phi$  is a radially symmetric function, then  $\Pi_{\phi;N,D}$  denotes the set of all outputs of a radial basis function network. In [8], Girosi, Jones, and Poggio have pointed out the importance of the study of the more general case considered here. They have demonstrated how such general networks arise naturally in such applications as image processing and graphics, as solutions of certain extremal problems. Our first theorem in this section is the following Theorem 2.1. For multi-integers  $\mathbf{k} = (k_1, \dots, k_d)$ ,  $\mathbf{m} = (m_1, \dots, m_d) \in \mathbb{Z}^d$ , we write  $\mathbf{k} \geq \mathbf{m}$  if  $k_j \geq m_j$ ,  $1 \leq j \leq d$ . If  $t \in \mathbb{R}$ , we write  $\mathbf{k} \geq t$  if  $k_j \geq t$ ,  $1 \leq j \leq d$ . The set of all  $\mathbf{k} \in \mathbb{Z}^d$  with  $\mathbf{k} \geq 0$  will be denoted by  $\mathbb{Z}_+^d$ . Throughout the rest of this paper, we adopt the following convention regarding constants. The letters  $c, c_1, c_2 \dots$  will denote positive constants depending only on  $p, d, s, \phi, K$  and  $\mathcal{F}$  (equivalently  $\Omega$ ), but independent of any other variables not explicitly indicated. Their values may be different at different occurrences, even within the same formula.

**Theorem 2.1** *Let  $d \geq 1$  be an integer, and  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  be infinitely many times continuously differentiable in some open sphere in  $\mathbb{R}^d$ . We further assume that there exists  $\mathbf{b}$  in this sphere such that*

$$D^{\mathbf{k}}\phi(\mathbf{b}) \neq 0, \quad \mathbf{k} \in \mathbb{Z}_+^d. \quad (2.6)$$

*Let  $s \geq 1$ ,  $N \geq 1$ ,  $m \geq (\log d)/s$  be integers,  $1 \leq p \leq \infty$ ,  $K$  be a compact subset of  $L^p$ ,  $\Omega$  be a modulus of continuity and  $\mathcal{F}$  be the set as in (2.5). We write  $D := (2m + 1)^s$ . There exist continuous linear functionals  $\gamma_k : C(K) \rightarrow \mathbb{R}$ ,  $k = 1, \dots, N$ , and a continuous linear operator  $L_m : L^p \rightarrow \mathbb{R}^D$ , with the following property. For every  $F \in \mathcal{F}$ , there exist  $d \times D$  matrices  $A_k := A_k(F)$  with rank  $d$ ,  $k = 1, \dots, N$ , such that*

$$\sup_{f \in K} |F(f) - \sum_{k=1}^N \gamma_k(F) \phi(A_k(L_m(f)) + \mathbf{b})| \leq c \{ \Omega(\epsilon_{m,K}) + \Omega(m^\beta N^{-1/D}) \}, \quad (2.7)$$

where  $\beta := 2s \max(1/p, 1 - 1/p)$ .

It may seem a bit awkward to define the matrices  $A_k$  and operator  $L_m$  as we did in the above theorem, rather than just defining operators from  $L^p$  into  $\mathbb{R}^d$ . The reason for this will be explained during our discussion of the lower bounds. It is proved in [17] that the condition (2.6) is satisfied for a large class of activation functions, including each of the following, where for  $\mathbf{x} \in \mathbb{R}^d$ , we write  $|\mathbf{x}| := \left( \sum_{j=1}^d x_j^2 \right)^{1/2}$ :

$$d = 1, \quad \phi(x) := (1 + e^{-x})^{-1}, \quad (\text{The squashing function})$$

$$d \geq 1, \quad \phi(\mathbf{x}) := (1 + |\mathbf{x}|^2)^\alpha, \quad \alpha \notin \mathbb{Z}, \quad (\text{Generalized multiquadrics})$$

$$d \geq 1, \quad q \in \mathbb{Z}, \quad q > d/2, \quad \phi(\mathbf{x}) := \begin{cases} |\mathbf{x}|^{2q-d} \log |\mathbf{x}|, & d \text{ even,} \\ |\mathbf{x}|^{2q-d}, & d \text{ odd,} \end{cases} \quad (\text{Thin plate splines})$$

and

$$d \geq 1, \quad \phi(\mathbf{x}) := \exp(-|\mathbf{x}|^2). \quad (\text{The Gaussian function})$$

An important special case of the compact set  $K$  in Theorem 2.1 is the case when  $\epsilon_{m,K} = \mathcal{O}(m^{-\alpha})$  for some  $\alpha > 0$ . In this case, the order of magnitude of the quantity on the right hand side of (2.7) is minimized if we choose  $m$  to be a solution of the equation

$$(2m + 1)^s \log m = \frac{\log N}{\alpha + \beta}; \quad (2.8)$$

i.e., (cf. [24], Ex. 5.7, p. 14)

$$m \sim \left( \frac{\log N}{\log \log N} \right)^{1/s}. \quad (2.9)$$

(Our notation here is different from that in [24]. By  $A \sim B$ , we mean  $c_1 B \leq A \leq c_2 B$ .) The right hand side of (2.7) is then estimated by

$$c\Omega \left( \left( \frac{\log \log N}{\log N} \right)^{\alpha/s} \right).$$

We observe that if  $N_P$  denotes the number of real parameters in the network of (2.7), then this order of magnitude is also

$$c\Omega \left( \left( \frac{\log \log N_P}{\log N_P} \right)^{\alpha/s} \right).$$

This looks like a very weak rate of convergence indeed, but we will prove below that under the weak a priori assumptions which we have made, this result cannot be improved in general, except possibly for the factor of  $\log \log N_P$  in the numerator above.

Before describing these lower bounds, we take this opportunity to make a few remarks about the construction of the networks. We observe that the operator  $L_m$  is independent of the functionals  $F$  or the input signals  $f$ . During the proof, we will give explicit expressions for the quantities  $L_m$ ,  $A_k$ , and the coefficients  $\gamma_k$ . It will then be seen that the matrices  $A_k$  (and clearly the threshold  $\mathbf{b}$ ) are all uniformly

bounded, independent of the accuracy desired. It will also be seen that the matrices  $A_k$  depend not so much on  $F$  itself, but on the entire class  $\mathcal{F}$  and on

$$\sup\{|F(P)| : P \in \Pi_{2m,s}, \|P\|_p \leq c(N, m, p, s)\}$$

for some constant  $c$  depending upon the indicated parameters only. In particular, if we restrict the set  $\mathcal{F}$  to consist of functionals satisfying an a priori bound, then the matrices  $A_k$  will also be independent of  $F$ . In function approximation, a bound on the Sobolev norm already implies such a bound on the norms of the target functions. There are technical reasons for not imposing such a bound at the outset in this context. Nevertheless, there is no training involved in finding these and other parameters of the network, and we have avoided all the shortcomings of the usual, optimization based techniques, such as back-propagation.

Necessarily, the coefficients  $\gamma_k$  will be unbounded as the accuracy tends to 0; this is inevitable even in the case of ordinary function approximation [16]. This situation is akin to the case of the familiar polynomial approximation on the interval. If one writes the approximating polynomials in terms of the monomials, the coefficients are far from being well behaved. A change of basis to some system of orthogonal polynomials cures this problem. It will be seen from our proof that this is exactly the situation here as well. Thus, to implement the networks in practice, one should first implement certain auxiliary networks. These are independent of the actual systems being implemented, and the extra effort will pay off in terms of more stable coefficients of the resulting networks in actual approximation of the systems.

In the case of radial basis function networks, our matrices are not all equal to the identity matrix; i.e., our networks are not pure translation networks. The ideas in [15] can be used to construct Gaussian networks to achieve the same rate of approximation, where only pure translations are required. We will not pursue these ideas here. An examination of our proofs here and those in [15] will show clearly how to do this. No new ideas are involved in this construction.

Now we turn our attention to the lower bounds. These will be obtained for *any* models, not just for generalized translation networks. A model can be described mathematically in terms of two functions, a function  $\pi : \mathcal{F} \rightarrow \mathbb{R}^N$  which selects the parameters of the model, and a function  $\mathcal{M}_N : \mathbb{R}^N \rightarrow C(K)$  which reconstructs the model using these parameters. For any system  $F \in \mathcal{F}$ ,  $\mathcal{M}_N(\pi(F))$  is the model simulating  $F$ . For example, in the case of a neural network model,  $N$  denotes the total number of its real parameters, including the coefficients, components of weights and thresholds. The training of the network consists of defining the function  $\pi$  which defines these parameters given a target function. The mapping  $\mathcal{M}$  then defines the output of the network itself, given these parameters. In our formulation of Theorem 2.1, the operator  $L_m$ , being independent of  $F$  and  $f$ , is part of the definition of  $\mathcal{M}$ . The lower bounds in the discussion below will apply also even if the weights  $A_k$  were to depend upon  $F$  in a much more complicated manner. The split formulation helps us to allow this dependence without introducing new definitions of “widths” (see below), and also underlines the possible interdependence of  $\pi$  and  $\mathcal{M}$ . It is naturally desirable that the function  $\pi$  should be continuous on  $\mathcal{F}$ , so as to achieve a robust model. As pointed out by Helmicki, et. al. [10], it is desirable to measure the error in this simulation in the “worst-case” scenario. Mathematically, this amounts to estimating the quantity

$$\sup_{f \in K} |F(f) - \mathcal{M}_N(\pi(F), f)|.$$

Since  $F$  itself is unknown as well, we are led to consider

$$\sup_{F \in \mathcal{F}, f \in K} |F(f) - \mathcal{M}_N(\pi(F), f)|.$$

The inherent error in modelling is then measured by the *nonlinear  $N$ -width* ([6]), defined by

$$\Delta_N(\mathcal{F}) := \inf_{\mathcal{M}_N, \pi} \sup_{F \in \mathcal{F}, f \in K} |F(f) - \mathcal{M}_N(\pi(F), f)|, \quad (2.10)$$

where the infimum is taken over all continuous functions  $\pi : \mathcal{F} \rightarrow \mathbb{R}^N$  and operators  $\mathcal{M}_N : \mathbb{R}^N \rightarrow C(K)$ .

The following theorem gives a lower bound for the nonlinear  $N$ -width for  $\mathcal{F}$  in case of certain compact sets  $K \subset L^2$ . To describe this compact set, we write  $|\mathbf{k}|_\infty = \max_{1 \leq j \leq s} |k_j|$ ,  $\mathbf{k} = (k_1, \dots, k_s) \in \mathbb{Z}^s$ . For  $f \in L^2$ , let  $\sum a_{\mathbf{k}}(f) P_{\mathbf{k}}$  denote the Fourier-Legendre expansion of  $f$  (cf. (3.2), (3.5) below).

**Theorem 2.2** *Let  $s \geq 1$  be an integer,  $\alpha > 0$ ,*

$$K := \{f \in L^2 : \sum_{\mathbf{k}} |\mathbf{k}|^{2\alpha} |a_{\mathbf{k}}(f)|^2 \leq 1\}. \quad (2.11)$$

*Let  $\Omega$  be a modulus of continuity. Then for integer  $N \geq 1$ ,*

$$\Delta_N(\mathcal{F}_{\Omega,2,s}) \geq c\Omega((\log N)^{-\alpha/s}). \quad (2.12)$$

### 3 Proofs.

In this section, we prove Theorems 2.1 and 2.2. The idea behind the proof of Theorem 2.1 is the following. From  $\mathcal{F}$  and  $K$ , we will construct a family of functions on a ball in  $\mathbb{R}^D$ . Each member of this family will be approximated by a polynomial of a certain degree  $n$ , where the rate of approximation is given by Newman and Shapiro in [22]. Neural networks to approximate such polynomials are developed in [17].

We start this program by recalling the definition and a few facts about Legendre polynomials. A standard way to define the univariate, orthonormalized Legendre polynomial is by the Rodrigues' formula ([29]) :

$$P_n(x) := \frac{(-1)^n \sqrt{n+1/2}}{2^n n!} \left(\frac{d}{dx}\right)^n \{(1-x^2)^n\}, \quad x \in \mathbb{R}, n = 0, 1, \dots \quad (3.1)$$

If  $\ell \geq 2$ ,  $\mathbf{k} = (k_1, \dots, k_\ell) \in \mathbb{Z}_+^\ell$ , and  $\mathbf{x} = (x_1, \dots, x_\ell) \in \mathbb{R}^\ell$ , then we write

$$P_{\mathbf{k}}(\mathbf{x}) := \prod_{j=1}^{\ell} P_{k_j}(x_j). \quad (3.2)$$

It is well known that

$$\int_{[-1,1]^\ell} P_{\mathbf{k}}(\mathbf{x}) P_{\mathbf{m}}(\mathbf{x}) d\mathbf{x} = \begin{cases} 1, & \text{if } \mathbf{k} = \mathbf{m}, \\ 0, & \text{otherwise.} \end{cases} \quad (3.3)$$

If  $g \in L^1([-1,1]^\ell)$ , its Fourier-Legendre coefficients are defined by

$$a_{\mathbf{k}}(g) := \int_{[-1,1]^\ell} g(\mathbf{t}) P_{\mathbf{k}}(\mathbf{t}) d\mathbf{t}, \quad (3.4)$$

and the expression

$$\sum_{\mathbf{k} \in \mathbb{Z}^\ell, \mathbf{k} \geq 0} a_{\mathbf{k}}(g) P_{\mathbf{k}}, \quad (3.5)$$

convergent or not, is called its Fourier-Legendre expansion. Even if the expansion might not converge, the multi-sequence  $\{a_{\mathbf{k}}(g)\}$  uniquely determines  $g$ .

We develop some further notation. If  $D$  is as in Theorem 2.1,  $\mathbf{a} \in \mathbb{R}^D$ , then we may index the coordinates of  $\mathbf{a}$  by elements of  $\{0, 1, \dots, 2m\}^s$  instead of the usual indexing by elements of  $\{1, 2, \dots, (2m+1)^s\}$ . Then the operator  $\mathbb{I} : \mathbb{R}^D \rightarrow \Pi_{2m,s}$  is defined by

$$\mathbb{I}(\mathbf{a}) := \sum_{0 \leq \mathbf{k} \leq 2m} \mathbf{a}_{\mathbf{k}} P_{\mathbf{k}}. \quad (3.6)$$

The quantity  $|\mathbf{a}|$  will denote the Euclidean norm of  $\mathbf{a}$ . The following lemma establishes an important connection between  $|\mathbf{a}|$  and  $\|\mathbb{I}(\mathbf{a})\|_p$ .

**Lemma 3.1** *Let  $m \geq 0$  be an integer,  $D = (2m+1)^s$ . Then, with*

$$A_{m,p,s} := m^{2s \max(1/p-1/2,0)}, \quad B_{m,p,s} := m^{2s \max(1/2-1/p,0)}, \quad (3.7)$$

*we have*

$$|\mathbf{a}| \leq c A_{m,p,s} \|\mathbb{I}(\mathbf{a})\|_p \leq c_1 A_{m,p,s} B_{m,p,s} |\mathbf{a}| = c_1 m^{\beta-s} |\mathbf{a}|, \quad (3.8)$$

*where  $\beta := 2s \max(1/p, 1-1/p)$ .*

PROOF. If  $0 < q, r \leq \infty$ ,  $\ell \geq 1$  be an integer, and  $P \in \Pi_{\ell m, 1}$  is a univariate polynomial, then it is known ([21], p. 114) that

$$\|P\|_{q, [-1, 1]} \leq \begin{cases} cm^{2(1/q-1/r)}\|P\|_{r, [-1, 1]}, & \text{if } r < q \\ c\|P\|_{r, [-1, 1]}, & \text{if } r \geq q. \end{cases} \quad (3.9)$$

The case  $q = \infty$  is not included on p. 114 of [21]. However, it forms the basis for the proof, and is proved separately in the form of Theorem 13 on p. 113, and Lemma 5 on p. 108 of [21]. The second inequality above is, of course, just a consequence of Hölder's inequality. If  $P \in \Pi_{\ell m, s}$ , we apply the above estimates one by one to each of the coordinates of its argument to obtain

$$\|P\|_q \leq \begin{cases} cm^{2s(1/q-1/r)}\|P\|_r, & \text{if } r < q \\ c\|P\|_r, & \text{if } r \geq q. \end{cases} \quad (3.10)$$

The estimates (3.8) are now easy to deduce using the Parseval identity.  $\square$

The following lemma gives the first step of our proof: the construction of a family of functions on a ball of  $\mathbb{R}^D$ . We assume the notation and conditions of Theorem 2.1.

**Lemma 3.2** *There exists a continuous linear operator  $L_m : L^p \rightarrow \mathbb{R}^D$  such that*

$$|F(f) - F(\mathbb{I}(L_m(f)))| \leq c\Omega(\epsilon_{m, K}), \quad F \in \mathcal{F}, f \in K. \quad (3.11)$$

Moreover,

$$|L_m(f)| \leq CA_{m, p, s}\|f\|_p, \quad f \in L^p, \quad (3.12)$$

where  $C > 0$  is a constant depending only on  $p$  and  $s$ , but not on  $m$ .

PROOF. We take any continuous linear operator  $V_m : L^p \rightarrow \Pi_{2m, s}$  such that

$$\|f - V_m(f)\|_p \leq cE_{m, p, s}(f) \quad (3.13)$$

for all  $f \in L^p$ , where  $c$  is a positive constant depending only on  $p$  and  $s$ , and not on  $f$  or  $m$ . There are many such operators known in the literature (cf. [30], [13], [17]), and the actual choice is immaterial for this proof. We write

$$L_m(f) := (a_{\mathbf{k}}(V_m(f)))_{0 \leq \mathbf{k} \leq 2m, \mathbf{k} \in \mathbb{Z}^s}. \quad (3.14)$$

Then  $L_m$  is a continuous linear operator on  $L^p$ , and

$$\|f - \mathbb{I}(L_m(f))\|_p = \|f - V_m(f)\|_p \leq c\epsilon_{m, K}, \quad f \in K. \quad (3.15)$$

The estimate (3.11) is now clear. From Lemma 3.1, we obtain for  $f \in L^p$  that

$$|L_m(f)| \leq cA_{m, p, s}\|V_m(f)\|_p \leq cA_{m, p, s}\|f\|_p. \quad (3.16)$$

This proves (3.12).  $\square$

Until the end of the proof of Theorem 2.1, we retain the value of the constant  $C$  in (3.12).

In order to describe the next step of our proof, we define for any  $F \in \mathcal{F}$ , a function  $\mu_m(F)$  of  $D$  variables by the formula

$$\mu_m(F, \mathbf{a}) := F(\mathbb{I}(\mathbf{a})), \quad \mathbf{a} \in \mathbb{R}^D. \quad (3.17)$$

The next lemma gives a bound on the degree of polynomial approximation of  $\mu_m(F)$ , thus completing the second step in our proof.

**Lemma 3.3** *With the notation as in Theorem 2.1, for any  $F \in \mathcal{F}$ , and integer  $n \geq 1$ , there exists a polynomial  $Q(F) \in \Pi_{n, D}$  such that*

$$|\mu_m(F, \mathbf{a}) - Q(F, \mathbf{a})| \leq c\Omega(m^\beta/n), \quad |\mathbf{a}| \leq CA_{m, p, s}M_K. \quad (3.18)$$

PROOF. Using Lemma 3.1, and the definitions of the functions  $\mu_m$ ,  $\mathbb{I}$ , we obtain for any  $\mathbf{a}, \mathbf{d} \in \mathbb{R}^D$ ,

$$\begin{aligned} |\mu_m(F, \mathbf{a}) - \mu_m(F, \mathbf{d})| &= |F(\mathbb{I}(\mathbf{a})) - F(\mathbb{I}(\mathbf{d}))| \\ &\leq \Omega(\|\mathbb{I}(\mathbf{a}) - \mathbb{I}(\mathbf{d})\|_p) \\ &\leq c\Omega(B_{m,p,s}|\mathbf{a} - \mathbf{d}|). \end{aligned} \quad (3.19)$$

The assertions of Lemma 3.3 now follow from Theorem 3 in the paper [22] of Newman and Shapiro.  $\square$

Finally, we prove another lemma, which will help us to replace the polynomials  $Q(F)$  by generalized translation networks.

**Lemma 3.4** *Let  $\phi$  satisfy the conditions of Theorem 2.1,  $n \geq 1$  be an integer and  $\mathbf{k} \in \mathbb{Z}_+^D$  and  $\mathbf{k} \leq n$ . Then for every  $\epsilon > 0$ , there exists  $G_{\mathbf{k},n,\epsilon} \in \Pi_{\phi;(6n+1)^D,D}$  such that*

$$(3.16) \quad \|\mathbb{P}_{\mathbf{k}} - G_{\mathbf{k},n,\epsilon}\|_{\infty,[-1,1]^D} \leq \epsilon.$$

*The matrices involved in  $G_{\mathbf{k},n,\epsilon}$  may be chosen to be matrices of rank  $d$ . Further, the matrices in each  $G_{\mathbf{k},n,\epsilon}$  may be chosen from a fixed set with cardinality not exceeding  $(6n+1)^D$ . The threshold of each neuron in each  $G_{\mathbf{k},n,\epsilon}$  is  $\mathbf{b}$  as defined in (2.6).*

PROOF. The proof of this lemma is almost verbatim the same as that of Lemma 3.2 of [17]. We omit the details, but point out only the necessary changes. For  $\mathbf{w} \in \mathbb{R}^D$ , we define the  $d \times D$  matrices  $A_{\mathbf{w}}$  by

$$A_{\mathbf{w}} := \begin{pmatrix} & \vdots & w_{d+1}, \dots, w_D \\ \text{diag}(w_1, \dots, w_d) & \vdots & \dots \dots \dots \\ & \vdots & \mathbf{0} \end{pmatrix}, \quad (3.20)$$

and write

$$\Phi(\mathbf{w}; \mathbf{x}) := \phi(A_{\mathbf{w}}\mathbf{x} + \mathbf{b}), \quad \mathbf{x} \in \mathbb{R}^D.$$

Then, as in Lemma 3.2 of [17], we may express every monomial  $\mathbf{x}^{\mathbf{p}}$  ( $\mathbf{p} \in \mathbb{Z}_+^D$ ) as a multiple of a derivative of  $\Phi$  with respect to  $\mathbf{w}$ . A divided difference then gives a network  $\Phi_{\mathbf{p},h} \in \Pi_{\phi;(p_1+1)\dots(p_D+1),D}$  that approximates  $\mathbf{x}^{\mathbf{p}}$  uniformly on  $[-1, 1]^D$  to any desired degree of accuracy. The network  $G_{\mathbf{k},n,\epsilon}$  is obtained by expressing  $\mathbb{P}_{\mathbf{k}}$  in terms of monomials, and then replacing each monomial by the approximating network. The matrices involved are of the form (3.20). By slight perturbations, one may also ensure the matrices to be of rank  $d$ , and yet ensure that the set of matrices is fixed, independent of  $\mathbf{k}$ .  $\square$

PROOF OF THEOREM 2.1. If  $f \in K$ ,  $F \in \mathcal{F}$ , (3.12) and Lemma 3.3 gives a polynomial  $Q(F) \in \Pi_{n,D}$  such that

$$|F(\mathbb{I}(L_m(f))) - Q(F, L_m(f))| = |\mu_m(F, L_m(f)) - Q(F, L_m(f))| \leq c\Omega(m^\beta/n). \quad (3.21)$$

From (3.11), we now obtain for all  $f \in K$ ,  $F \in \mathcal{F}$ ,

$$|F(f) - Q(F, L_m(f))| \leq c \left\{ \Omega(\epsilon_{m,K}) + \Omega(m^\beta/n) \right\}. \quad (3.22)$$

We write

$$Q(F) =: \sum_{\mathbf{k} \in \mathbb{Z}_+^D, \mathbf{k} \leq n} d_{\mathbf{k}}(F) \mathbb{P}_{\mathbf{k}}.$$

Then

$$\begin{aligned} \sum_{\mathbf{k} \in \mathbb{Z}_+^D, \mathbf{k} \leq n} |d_{\mathbf{k}}(F)| &\leq c(n, m, p, s) \max_{|\mathbf{a}| \leq CA_{m,p,s}M_K} |Q(F, \mathbf{a})| \\ &\leq c(n, m, p, s) \max_{|\mathbf{a}| \leq CA_{m,p,s}M_K} |\mu_m(F, \mathbf{a})| \\ &\leq c(n, m, p, s) \max_{P \in \Pi_{2m,s}, \|P\|_p \leq c_2(n,m,p,s)} |F(P)| \end{aligned}$$

We may then use the networks constructed in Lemma 3.4 to arrive (after an elementary rescaling) at a network approximating  $Q(F)$  to an arbitrary degree of approximation on the ball in  $\mathbb{R}^D$  of radius  $CA_{m,p,s}M_K$ . Moreover, these networks have the form required in Theorem 2.1. In view of (3.22), these networks evaluated at  $L_m(f)$  give the required degree of approximation to  $F(f)$ .  $\square$

Next, we turn to the proof of Theorem 2.2. As expected, the proof of this theorem relies upon the notion of the Bernstein widths. If  $X$  is a normed linear space,  $\|\cdot\|$  is its norm,  $A \subseteq X$ , and  $N \geq 1$  is an integer, the Bernstein  $N$ -width of  $A$  (with respect to  $X$ ) is defined as follows. Let  $\mathcal{X}_{N+1}$  be the set of all linear subspaces of  $X$  having dimension not exceeding  $N+1$ . The Bernstein  $N$ -width is given by the expression

$$b_N(A) := \sup_{Y \in \mathcal{X}_{N+1}} \{\rho : \rho f / \|f\| \in A \text{ for all } f \in Y\}. \quad (3.23)$$

In the context of this paper, we take  $X$  to be the space  $C(K)$  with the natural supremum norm. According to a result of DeVore, Howard and Micchelli ([6]),

$$\Delta_N(\mathcal{F}) \geq b_N(\mathcal{F}), \quad N = 1, 2, \dots \quad (3.24)$$

Therefore, we will obtain a lower bound for  $b_N(\mathcal{F})$ . The idea is the same as in Theorem C in [22]. We provide some details since we are unable to find a precise reference in the context of Bernstein  $N$ -widths. PROOF OF THEOREM 2.2. In this proof, we write  $\mathcal{F} := \mathcal{F}_{\Omega,2,s}$ . Let  $m$  be an integer such that  $2m+1 \sim (\log N)^{1/s}$ , and  $D := (2m+1)^s$ . Following Newman and Shapiro ([22], Lemma 1), we obtain  $N+1$  points  $\mathbf{a}_1, \dots, \mathbf{a}_{N+1}$  in  $\mathbb{R}^D$ , such that

$$|\mathbf{a}_i| \leq 1/D^{\alpha/s}, \quad i = 1, \dots, N+1, \quad (3.25)$$

and

$$|\mathbf{a}_i - \mathbf{a}_j| \geq (2D^{\alpha/s}N^{1/D})^{-1}, \quad i \neq j, \quad i, j = 1, \dots, N+1. \quad (3.26)$$

Then the functions

$$f_i := \mathbb{I}(\mathbf{a}_i) \quad (3.27)$$

are in  $K$ , and

$$\|f_i - f_j\|_2 = |\mathbf{a}_i - \mathbf{a}_j| \geq (2D^{\alpha/s}N^{1/D})^{-1} =: 2\delta, \quad i \neq j, \quad i, j = 1, \dots, N+1. \quad (3.28)$$

Next, we define for  $f \in L^2$ ,  $i = 1, \dots, N+1$ ,

$$F_i(f) := \begin{cases} \Omega(\delta - \|f - f_i\|_2), & \text{if } \|f - f_i\|_2 < \delta \\ 0, & \text{otherwise.} \end{cases} \quad (3.29)$$

Then each  $F_i$  is a bounded, continuous, real function on  $L^2$ . If  $Y$  is the linear span of  $F_i$ , then  $Y$  has dimension  $N+1$ . Let  $F = \sum d_i F_i$  be an arbitrary element of  $Y$ . Then

$$\|F\| := \sup_{f \in K} |F(f)| = \Omega(\delta) \max_{1 \leq i \leq N+1} |d_i|. \quad (3.30)$$

Using the monotonicity and subadditivity of  $\Omega$ , if  $\|f - f_i\|_2, \|g - f_i\|_2 \leq \delta$ , then

$$\begin{aligned} |F(f) - F(g)| &= |d_i| \left| \Omega(\delta - \|f - f_i\|_2) - \Omega(\delta - \|g - f_i\|_2) \right| \\ &\leq |d_i| \Omega \left( \left| \|g - f_i\|_2 - \|f - f_i\|_2 \right| \right) \\ &\leq \frac{\|F\|}{\Omega(\delta)} \Omega(\|f - g\|_2). \end{aligned} \quad (3.31)$$

If  $\|f - f_i\|_2, \|g - f_j\|_2 \leq \delta$  and  $i \neq j$ , then we pick  $h, \tilde{h}$  with  $\|h - f_i\|_2 = \delta, \|\tilde{h} - f_j\|_2 = \delta, \|f - h\|_2, \|g - \tilde{h}\|_2 \leq \|f - g\|_2$ . (For example, we may take the ‘‘line’’ joining  $f$  and  $g$ , and let  $h$  and  $\tilde{h}$  be the ‘‘points’’ where

this “line” intersects the balls around  $f_i$  and  $f_j$  respectively.) Then  $F(h) = F(\tilde{h}) = 0$  and, using the estimate just proved,

$$\begin{aligned} |F(f) - F(g)| &\leq |F(f) - F(h)| + |F(\tilde{h}) - F(g)| \\ &\leq \frac{\|F\|}{\Omega(\delta)} \Omega(\|f - h\|_2) + \frac{\|F\|}{\Omega(\delta)} \Omega(\|\tilde{h} - g\|_2) \\ &\leq \frac{2\|F\|}{\Omega(\delta)} \Omega(\|f - g\|_2). \end{aligned} \tag{3.32}$$

The case when at most one of  $f$  and  $g$  are in a ball of radius  $\delta$  around one of the  $f_i$ 's is simpler. We have therefore proved that

$$|F(f) - F(g)| \leq \frac{2\|F\|}{\Omega(\delta)} \Omega(\|f - g\|_2) \tag{3.33}$$

for all  $f, g \in L^2$ . Consequently, any  $F \in Y$  with  $\|F\| \leq \Omega(\delta)/2$  is in  $\mathcal{F}$ ; i.e.,

$$b_N(\mathcal{F}) \geq (1/2)\Omega(\delta). \tag{3.34}$$

From the definition of  $D$  and  $\delta$  (3.28), we see that  $N^{1/D} \sim 1$  and  $\delta \sim (\log N)^{-\alpha/s}$ . Therefore, the proof of Theorem 2.2 is complete in view of the estimate (3.24).  $\square$

## 4 Conclusions

Given a continuous, not necessarily linear, functional  $F$  on an  $L^p$  space, we have constructed generalized translation networks which provide a near optimal approximation to  $F$  in terms of the number of neurons involved. Our networks evaluate a functional of the form  $\sum_{k=1}^N \gamma_k \phi(A_k(L_m(f)) + \mathbf{b})$ , where  $f$  is the input signal in the  $L^p$ -space, and  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  is the activation function. This setup covers the conventional neural networks ( $d = 1$ ), as well as radial (elliptical) basis function networks, where  $d = s$  and  $\phi$  is a radially symmetric function.

In our construction,  $L_m$  is a linear operator mapping  $f$  to a judiciously chosen number of coefficients in a polynomial approximation to  $f$ . In particular, this operator is independent of both the input signals and the system to be approximated. The remaining parameters are obtained as in [17] to approximate a function of finitely many real variables, constructed from  $F$  and  $L_m$ . The parameters  $A_k$ 's are full-rank matrices depending only on a norm of  $F$ , but are the same for all  $F$  having the same norm bound. The parameter  $\mathbf{b}$  depends only on  $\phi$ . The parameters  $\gamma_k$  are continuous linear functionals of  $F$ . Explicit formulas are given for all of these parameters; one does not need to train the network in the usual sense. In particular, our constructions are explicit, deterministic, and do not involve any iterative and/or optimization techniques, such as back-propagation.

Our main objective is to obtain estimates on the number of neurons in terms of the desired accuracy in approximation, and the properties of the functional  $F$  and the compact set on which the approximation takes place. Our estimates are valid also when a uniform approximation of a uniformly continuous operator on  $L^p$ , taking values in a class of bounded, continuous functions, is desired. Lower bounds for the degree of approximation are also obtained.

Although the paper is mainly of a theoretical nature, the proofs suggest actual constructions based on classical polynomial approximation theory. We have not conducted any numerical experiments, but the approximation theory tools used here are well tested over the last several decades.

## References

- [1] A. R. BARRON, *Universal approximation bounds for superposition of a sigmoidal function*, IEEE Trans. Information Theory, **39** (1993), 930-945.
- [2] T. CHEN AND H. CHEN, *Approximation of continuous functionals by neural networks with application to dynamical systems*, IEEE Trans. Neural Networks, **4** (1993), 910-918.

- [3] T. CHEN AND H. CHEN, *Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems*, IEEE Trans. Neural Networks, **6** (1995), 911-917.
- [4] T. CHEN AND H. CHEN, *Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks*, IEEE Trans. Neural Networks, **6** (1995), 904-910.
- [5] G. CYBENKO, *Approximation by superposition of sigmoidal functions*, Mathematics of Control, Signal and Systems, **2** (1989), 303-314.
- [6] R. DEVORE, R. HOWARD AND C. A. MICCHELLI, *Optimal nonlinear approximation*, Manuscripta Mathematica, **63** (1989), 469-478.
- [7] A. T. DINGANKAR, "On Applications of Approximation Theory to Identification, Control, and Classification", Ph. D. Dissertation, Univ. of Texas at Austin, 1995.
- [8] F. GIROSI, M. JONES AND T. POGGIO, *Regularization theory and neural networks architectures*, Neural Computation, **7** (1995), 219-269.
- [9] C. GOFFMAN AND G. PEDRICK, "First course in Functional Analysis", Prentice Hall, Engelwood Cliffs, 1965.
- [10] A. J. HELMICKI, C. A. JACOBSEN, AND C. N. NETT, *Control oriented system identification : a worst case deterministic approach in  $\mathcal{H}_\infty$* , IEEE Trans. Auto. Control, **3** (1991), 1163-1176.
- [11] K. HORNIK, M. STINCHCOMBE AND H. WHITE, *Multilayer feedforward networks are universal approximators*, Neural Networks, **2** (1989), 359-366.
- [12] A. U. LEVIN AND K. S. NARENDRA, *Identification using feedforward networks*, Neural Computation, **7** (1995), 349-357.
- [13] G. G. LORENTZ, "Approximation of Functions", Holt, Rinehart and Winston, New York, 1966.
- [14] G. G. LORENTZ, "Bernstein Polynomials", University of Toronto Press, Toronto, 1953.
- [15] H. N. MHASKAR, *Versatile Gaussian networks*, Proceedings of IEEE Workshop on Nonlinear Image and Signal Processing, (I. Pitas Editor), Halkidiki, Greece, June, 1995, IEEE, pp.70-73.
- [16] H. N. MHASKAR, *On smooth activation functions*, To appear in the Annals of Mathematics in Artificial Intelligence.
- [17] H. N. MHASKAR, *Neural networks for optimal approximation of smooth and analytic functions*, Neural Computation, **8** (1996), 164- 177.
- [18] H. N. MHASKAR AND C. A. MICCHELLI, *Degree of approximation by neural and translation networks with a single hidden layer*, Advances in Applied Mathematics, **16** (1995), 151-183.
- [19] D. S. MODHA AND R. HECHT-NIELSEN, *Multilayer functionals*, in "Mathematical Approaches to Neural Networks" (J. G. Taylor Ed.), Elsevier Science Publ., 1993, pp. 235-260.
- [20] K. S. NARENDRA AND K. PARTHASARATHY, *Identification and control of dynamic systems using neural networks*, IEEE Trans. Neural Networks, **1** (1990), 4-27.
- [21] P. NEVAI, "Orthogonal Polynomials", Memoirs of Amer. Math. Soc., **18**(213), Amer. Math. Soc., Providence, Rhode Island, 1979.
- [22] D. J. NEWMAN AND H. S. SHAPIRO, *Jackson's theorem in higher dimensions*, in "Approximation Theory, Proc. Conf. Math. Res. Inst., Oberwolfach, 1963", (P. L. Butzer and J. Korevaar Eds.), ISNM **5**, Birkhäuser, Basel, pp.208-219.

- [23] J. PARK AND I. W. SANDBERG, *Universal approximation using radial basis function networks*, Neural Computation, **3** (1991), 246-257.
- [24] F. W. J. OLVER, "Asymptotics and special functions", Academic Press, New York, 1974.
- [25] I. W. SANDBERG, *Approximation theorems for discrete time systems*, IEEE Trans. Circuits and Systems, **38** (1991), 564-566.
- [26] I. W. SANDBERG, *Gaussian basis functions and approximations for nonlinear systems*, Digest of the Ninth Kobe International Symposium on Electronics and Information Sciences, Kobe, Japan, June 1991, pp. 3-1-3-6.
- [27] S. L. SOBOLEV, "Applications of Functional Analysis in Mathematical Physics", Trans. Math. Monographs, **7**, Amer. Math. Soc., Providence, Rhode Island, 1963.
- [28] E. D. SONTAG, *Some topics in neural networks and control*, Siemens Corporate Research Inc., Report number LS93- 02.
- [29] G. SZEGÖ, "Orthogonal Polynomials", Amer. Math. Soc. Coll. Publ., **23**, Providence, Rhode Island, 1975.
- [30] A. F. TIMAN, "Theory of Approximation of Functions of a Real Variable", Macmillan Co., New York, 1963.