NEURAL NETWORKS FOR OPTIMAL APPROXIMATION OF SMOOTH AND ANALYTIC FUNCTIONS

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

Neural Computation, **8** (1996), 164- 177.

Abstract. We prove that neural networks with a single hidden layer are capable of providing an optimal order of approximation for functions assumed to possess a given number of derivatives, if the activation function evaluated by each principal element satisfies certain technical conditions. Under these conditions, it is also possible to construct networks that provide a geometric order of approximation for analytic target functions. The permissible activation functions include the squashing function $(1+e^{-x})^{-1}$ as well as a variety of radial basis functions. Our proofs are constructive. The weights and thresholds of our networks are chosen independently of the target function; we give explicit formulas for the coefficients as simple, continuous, linear functionals of the target function.

1. Introduction. In recent years, there has been a great deal of research in the theory of approximation of real valued functions using artificial neural networks with one or more hidden layers, with each principal element (*neuron*) evaluating a sigmoidal or radial basis function ([1, 2, 3, 5, 7, 8, 9, 15, 17, 18]). A typical density result shows that a network can approximate an arbitrary function in a given function class to any degree of accuracy. Such theorems are proved for instance in [5, 8] in the case of sigmoidal activation functions and in [16, 19] for radial basis functions. Very general theorems of this nature can be found in [9, 12].

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 example, the now classical result of Barron [1] shows that if the function is assumed to satisfy certain conditions expressed in terms of its Fourier transform, and each of the neurons evaluates a sigmoidal activation function, then at most $\mathcal{O}(\epsilon^{-2})$ neurons are needed to achieve the order of approximation ϵ. An interesting aspect of this result is that the order of magnitude of the number of neurons is independent of the number of variables on which the function depends. Other bounds of this nature are obtained in [13] when the activation function is not necessarily sigmoidal.

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 functions of s real variables having a continuous gradient. By a suitable normalization, one may assume that the gradient is bounded by 1. It is known (e.g. [6]) that any reasonable approximation scheme to provide an approximation order ϵ for all functions in this class must depend upon at least $\Omega(\epsilon^{-s})$ parameters. In [10], we showed how to construct networks with two hidden layers, each neuron evaluating a bounded sigmoidal function, to accomplish such an approximation order with $\mathcal{O}(\epsilon^{-s})$ neurons. Along with Micchelli [14] we have studied this problem in much greater detail. The best result known so far for networks with a single hidden layer is that $\mathcal{O}(\epsilon^{-s-1}\log(1/\epsilon))$ neurons are enough if the activation function is the squashing function $1/(1+e^{-x})$. In our work with Chui and Li [4] we have shown that if $s>1$ and the approximation is required to be "localized", then at least $\Omega(\epsilon^{-s}\log(1/\epsilon))$ neurons are necessary, even if different neurons may evaluate different activation functions. A detailed discussion of the notion of localized approximation is not relevant within the context of this paper; we refer the reader to [4]. We made a conjecture in [11] that with a sigmoidal activation function, the number of neurons necessary to provide the approximation order ϵ to all functions in this class, with or without localization, cannot be $\mathcal{O}(\epsilon^{-s})$.

In this paper, we disprove this conjecture. We prove that if the activation function satisfies certain technical conditions then the optimal order of approximation for this class (and other similar classes) can be achieved with a neural network with a single hidden layer. Our results will be formulated for neural networks more general than the traditional networks evaluating a univariate activation function. In particular, our results will include estimates on the order of approximation by generalized regularization networks introduced in [7, 17, 18]. The precise definitions and results will be given in the next section. The proofs of all the new results in Section 2 will be given in Section 3.

2. Main Results. Let $1 \le d \le s$, $n \ge 1$ be integers, $f : \mathbf{R}^s \to \mathbf{R}$ and $\phi : \mathbf{R}^d \to \mathbf{R}$. 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 $d \times s$ real matrices, the *thresholds* $\mathbf{b}_k \in \mathbf{R}^d$ and the *coefficients* $a_k \in \mathbf{R}$ $(1 \le k \le n)$. The set of all such functions (with a fixed n) will be denoted by $\Pi_{\phi;n,s}$. We are interested in approximating the target function f by elements of $\Pi_{\phi;n,s}$ on $[-1,1]^s$. In the case when $d = 1$, the class $\Pi_{\phi;n,s}$ denotes the outputs of the classical neural networks with one hidden layer consisting of n neurons, each evaluating the univariate activation function ϕ. In [7, 17, 18] Girosi, Poggio and Jones 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 applications such as image processing and graphics as solutions of certain extremal problems. Our approximations will not be constructed as in [7, 17, 18] as solutions of extremal problems, but rather will be given explicitly. They will not provide the *best approximation,* but will nevertheless provide the optimal order of approximation.

An additional advantage of our networks is that the *weights* A_k's and the *thresholds* $\mathbf{b}_k$'s will be determined independently of the target function f. We observe in this connection that the determination of these quantities is typically a major problem in most traditional training algorithms such as backpropagation. In fact, the only "training" required for our networks consists of evaluating the coefficients a_k. We give explicit formulas for these coefficients as linear combinations of the Fourier-Chebyshev coefficients of the target function. Alternative formulas based on the values of the target function can also be given, but we do not present these alternative constructions here, since a good discussion of this issue would require us to elaborate upon some very techincal background material. From a practical perspective, we observe that we are assuming that the target function can be sampled without noise at prescribed points. Our constructions are extremely simple, use no optimization, and avoid all the problems, for example, local minima, stability, etc., associated with the classical, optimization-based training paradigms such as backpropagation. We fully expect the constructions to be robust under noise, but have not developed any theory to deal with this question.

First, we introduce some notations. If $A \subseteq \mathbf{R}^s$ is Lebesgue measurable, and $f : A \to \mathbf{R}$ is a measurable function, we define the $L^p(A)$ norms of f as follows.

$$||f||_{p,A} := \begin{cases} \left\{ \displaystyle\int_A |f(\mathbf{x})|^p d\mathbf{x} \right\}^{1/p}, & \text{if } 1 \le p < \infty, \\ \operatorname*{ess\ sup}_{\mathbf{x}\in A} |f(\mathbf{x})|, & \text{if } p = \infty. \end{cases} \tag{2.1}$$

The class of all functions f for which $||f||_{p,A} < \infty$ is denoted by $L^p(A)$. It is customary (and in fact essential from a theoretical point of view) to adopt the convention that if two functions are equal almost everywhere in the measure-theoretic sense then they should be considered as equal elements of $L^p(A)$. We make two notational simplifications. The symbol $L^\infty(A)$ will denote the class of continuous functions on A. In this paper, we have no occasion to consider discontinuous functions in what is normally denoted by $L^\infty(A)$, and using this symbol for the class of continuous functions will simplify the statements of our theorems. Second, when the set $A = [-1,1]^s$, we will not mention the set in the notation. Thus, $||f||_p$ will mean $||f||_{p,[-1,1]^s}$ etc. We measure the *degree of approximation* of f by the expression

$$E_{\phi;n,p}(f) := \inf\{||f-g||_p \ : \ g \in \Pi_{\phi;n,s}\}. \tag{2.2}$$

The quantity $E_{\phi;n,p}(f)$ denotes the theoretically minimal error that can be achieved in approximating the function f in the L^p norm by generalized translation networks with n neurons each evaluating the activation function ϕ. The complexity problem is clearly equivalent to obtaining sharp estimates on $E_{\phi;n,p}(f)$.

In theoretical investigations of the degree of approximation, one typically makes an a priori assumption that the target function f, although itself unknown, belongs to some known class of functions. In this paper, we are interested in the Sobolev classes, which we define as follows. Let $r \ge 1$ be an integer and Q be a cube in $\mathbf{R}^s$. The class $W^p_{r,s}(Q)$ consists of all functions with $r-1$ continuous partial derivatives on Q which in turn can be expressed (almost everywhere on Q) as indefinite integrals of functions in $L^p(Q)$. Alternatively, the class $W^p_{r,s}(Q)$ consists of functions which have, at almost all points of Q, all partial derivatives up to order r such that all of these derivatives are in $L^p(Q)$. The Sobolev norm of $f \in W^p_{r,s}(Q)$ is defined by

$$||f||_{W^p_{r,s}(Q)} := \sum_{0 \le \mathbf{k} \le r} ||D^{\mathbf{k}} f||_{p,Q} \tag{2.3}$$

where for the multi-integer $\mathbf{k} = (k_1, \ldots, k_s) \in \mathbf{Z}^s$, $0 \le \mathbf{k} \le r$ means that each component of $\mathbf{k}$ is nonnegative and does not exceed r, $|\mathbf{k}| := \sum_{j=1}^s |k_j|$ and

$$D^{\mathbf{k}} f = \frac{\partial^{|\mathbf{k}|} f}{\partial x_1^{k_1} \cdots \partial x_s^{k_s}}, \qquad \mathbf{k} \ge 0.$$

Again, $W^\infty_{r,s}(Q)$ will denote the class of functions which have continuous derivatives of order r and lower. As before, if $Q = [-1,1]^s$, we will not mention it in the notation. Thus, we write $W^p_{r,s} = W^p_{r,s}([-1,1]^s)$ etc.

Since the target function itself is unknown, the quantity of interest is

$$E_{\phi;n,p,r,s} := \sup\{E_{\phi;n,p}(f) \ : \ ||f||_{W^p_{r,s}} \le 1\}. \tag{2.4}$$

We observe that any function in $W^p_{r,s}$ can be normalized so that $||f||_{W^p_{r,s}} \le 1$. Hence, $E_{\phi;n,p,r,s}$ measures the "worst case" degree of approximation by generalized translation networks with n neurons under the assumption that $f \in W^p_{r,s}$ and is properly normalized.

Since any element of $\Pi_{\phi;n,s}$ depends upon $(ds+d+1)n$ parameters, the general results by DeVore, Howard and Micchelli [6] *indicate* that

$$E_{\phi;n,p,r,s} \ge cn^{-r/s} \tag{2.5}$$

The general results in [6] are not exactly applicable here since the definition of the degree of approximation does not preclude the possibility that the parameters involved in the approximation may be discontinuous functionals on the class in question. Therefore, (2.5) is only a conjecture, rather than a known fact. In our constructions below, the parameters are continuous functionals of the class and hence (2.5) is applicable and shows that the networks provide an optimal order of approximation subject to the continuity requirement.

In the sequel, we make the following convention regarding constants. The letters $c, c_1, c_2, \cdots$ will denote positive constants which may depend upon p, r, s and other explicitly indicated quantities. Their value may be different at different occurrences, even within a single formula.

We now formulate our main theorem.

THEOREM 2.1. *Let $1 \le d \le s$, $r \ge 1$, $n \ge 1$ be integers, $1 \le p \le \infty$, $\phi : \mathbf{R}^d \to \mathbf{R}$ be infinitely many times continuously differentiable in some open sphere in $\mathbf{R}^d$. We further assume that there exists $\mathbf{b}$ in this sphere such that*

$$D^{\mathbf{k}}\phi(\mathbf{b}) \ne 0, \qquad \mathbf{k} \in \mathbf{Z}^d, \ \mathbf{k} \ge 0. \tag{2.6}$$

Then there exist $d \times s$ matrices $\{A_j\}_{j=1}^n$ with the following property. For any $f \in W^p_{r,s}$, there exist coefficients $a_j(f)$ such that

$$||f - \sum_{j=1}^n a_j(f)\phi(A_j(\cdot) + \mathbf{b})||_p \le cn^{-r/s}||f||_{W^p_{r,s}}. \tag{2.7}$$

The functionals a_j are continuous linear functionals on $W^p_{r,s}$. In particular,

$$E_{\phi;n,p,r,s} \le cn^{-r/s}. \tag{2.8}$$

We observe that the condition (2.6) implies that ϕ is not a polynomial. For the function $\phi(\mathbf{x}) := \cos x_1 + \cos x_2$, ($d = 2$), we have $D^{(1,1)}\phi \equiv 0$. Thus, when $d > 1$, the assumption (2.6) is stronger than the assumption that ϕ is not a polynomial. We suspect that it is a stronger assumption also in the case when $d = 1$. The following Proposition 2.2 shows that (2.6) is nevertheless satisfied by a large class of functions. In light of the first part of this proposition, we doubt that in the case when $d = 1$, a nonpolynomial function that is infinitely many times differentiable but does not satisfy (2.6) would be of any practical interest whatever.

PROPOSITION 2.2. *Let $d \geq 1$ be an integer and $\phi : \mathbf{R}^d \to \mathbf{R}$ be infinitely many times continuously differentiable on an open sphere B. If (2.6) is not satisfied, i.e., at every point of B some derivative of ϕ is zero, then for every closed sphere $U \subset B$, there exists a multi-integer $\mathbf{r} \geq 0$, a sphere $N \subseteq U$ and functions $h_{i,j,N}$ of $d-1$ real variables such that*

$$\phi(\mathbf{x}) = \sum_{i=1}^{d} \sum_{j=1}^{r_i - 1} h_{i,j,N}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) x_i^j, \qquad \mathbf{x} \in N. \tag{2.9}$$

If $d = 1$ and ϕ is analytic in a (complex) neighborhood of some point in B but not a polynomial, then (2.6) is satisfied.

Some of the important examples where (2.6) is satisfied are the following, where for $\mathbf{x} \in \mathbf{R}^d$, we write $||\mathbf{x}|| := \big(\sum_{j=1}^{d} x_j^2\big)^{1/2}$:

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

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

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

and

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

If the target function is merely assumed to be in L^p rather than in $W^p_{r,s}$, the estimate (2.7) leads to a similar estimate in terms of the modulus of smoothness of the function. This is a fairly standard argument in approximation theory, and does not add any new insight to the problem. Since a formulation of this result would require us to introduce a great deal more notation, we omit this apparent generalization.

The idea behind the proof of Theorem 2.1 is simple. It is well known that for every integer $m \geq r$, there exists a polynomial $P_m(f)$ of coordinatewise degree not exceeding m such that for every $f \in W^p_{r,s}$,

$$||f - P_m(f)||_p \leq c m^{-r} ||f||_{W^p_{r,s}}. \tag{2.10}$$

Following [9] we express each monomial in $P_m(f)$ in terms of a suitable derivative of ϕ. In turn, this derivative can be approximated by an appropriate divided difference, involving $\mathcal{O}(m^s)$ evaluations of ϕ. A careful book-keeping then yields Theorem 2.1.

If the target function f is analytic in the poly-ellipse

$$\mathcal{E}_\rho := \{\mathbf{z} = (z_1, \ldots, z_s) \in \mathbf{C}^s : \ |z_j + \sqrt{z_j^2 - 1}| \leq \rho,\ j = 1, \ldots, s\} \tag{2.11}$$

for some $\rho > 1$ and $1 < \rho_1 < \rho$ then for every integer $m \geq 1$ there exists a polynomial ([20]) $L_m(f)$ (different from the polynomials described above) with coordinatewise degree not exceeding m such that

$$||f - L_m(f)||_p \leq c_{\rho,\rho_1} \rho_1^{-m} \max_{\mathbf{z} \in \mathcal{E}_\rho} |f(\mathbf{z})|. \tag{2.12}$$

Approximating these polynomials by networks as above, we get the following Theorem 2.3.

THEOREM 2.3. *Let $1 \le d \le s$, $n \ge 1$ be integers, $1 < \rho_1 < \rho$, $1 \le p \le \infty$ and f be analytic in the poly-ellipse $\mathcal{E}_\rho$. Further, let ϕ be as in Theorem 2.1. Then*

$$E_{\phi;n,p}(f) \le c_{\rho,\rho_1}\rho_1^{-n^{1/s}} \max_{\mathbf{z}\in\mathcal{E}_\rho} |f(\mathbf{z})|. \tag{2.13}$$

It is possible to obtain some estimates on the degree of approximation under substantially weaker assumptions on ϕ than those assumed in Theorems 2.1, 2.3. One strategy, as in [9], would be to take the convolution of ϕ with a suitable, infinitely many times continuously differentiable function; apply Theorem 2.1 (or Theorem 2.3) to the resulting function and use a quadrature formula. We have not yet worked out the details of this argument, but it seems unlikely that these estimates would be optimal under the weak assumptions on ϕ. Using the ideas in the proof of Theorem 2.1, it is also possible to obtain estimates for simultaneous approximation of derivatives of the target function. This would follow from the corresponding theorems in the theory of trigonometric approximation (cf. [14]). Although the technical details in these generalizations are expected to be of some interest, we do not wish to pursue these ideas further in this paper.

3. Proofs. In order to prove Theorem 2.1, we first recall some well known facts from the theory of trigonometric approximation. These will be used to construct the polynomial operator in (2.10). The subspace of 2π-periodic functions in $L^p([-\pi,\pi]^s)$ (respectively $W^p_{r,s}([-\pi,\pi]^s)$) will be denoted by L^{p^*} (respectively $W^{p^*}_{r,s}$). If $g \in L^{p^*}$, its Fourier coefficients are defined by

$$\hat{g}(\mathbf{k}) := \frac{1}{(2\pi)^s}\int_{[-\pi,\pi]^s} g(\mathbf{t})e^{-i\mathbf{k}\cdot\mathbf{t}}d\mathbf{t}, \qquad \mathbf{k}\in\mathbf{Z}^s. \tag{3.1}$$

The partial sums of the Fourier series of g are defined by

$$s_{\mathbf{m}}(g,\mathbf{t}) := \sum_{-\mathbf{m}\le\mathbf{k}\le\mathbf{m}} \hat{g}(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{t}}, \qquad \mathbf{m}\in\mathbf{Z}^s,\ \mathbf{m}\ge 0,\ \mathbf{t}\in[-\pi,\pi]^s, \tag{3.2}$$

where the notation $\mathbf{k} \le \mathbf{m}$ means $k_j \le m_j$, $1 \le j \le s$. The de la Valleé Poussin operator is defined by

$$v_n(g,\mathbf{t}) := \frac{1}{(n+1)^s}\sum_{n\le\mathbf{m}\le 2n} s_{\mathbf{m}}(g,\mathbf{t}), \qquad n\in\mathbf{Z},\ n\ge 0,\ \mathbf{t}\in[-\pi,\pi]^s. \tag{3.3}$$

The de la Valleé Poussin operator has the following important property.

PROPOSITION 3.1. cf. ([22, 14]) *If $r \ge 1$, $s, m \ge 1$ are integers, $1 \le p \le \infty$ and $g \in W^{p^*}_{r,s}$ then $v_m(g)$ is a trigonometric polynomial of coordinatewise order at most $2m$ and*

$$||g - v_m(g)||_{p,[-\pi,\pi]^s} \le \frac{c}{m^r}||g||_{W^{p^*}_{r,s}} \tag{3.4}$$

Further,

$$\sum_{0\le\mathbf{k}\le 2m} |\widehat{v_m(g)}(\mathbf{k})| \le cm^\alpha ||g||_{W^{p^*}_{r,s}} \tag{3.5}$$

where $\alpha := s/\min(p,2)$.

The standard way to construct a periodic function from a function on $[-1,1]^s$ is to make the substitution $x_j =: \cos t_j$, $1 \le j \le s$, for $\mathbf{x} \in [-1,1]^s$ and $\mathbf{t} \in [-\pi,\pi]^s$. Obviously, the integrals defining the L^p norms are no longer equal under this substitution. Therefore, we make the following construction.

According to [21, §VI.3.1], there exists a continuous linear operator $T : W^p_{r,s} \to W^p_{r,s}([-2,2]^s)$ such that the restriction of $T(f)$ to $[-1,1]^s$ is (almost everywhere) equal to f. The continuity of the operator T means that

$$||T(f)||_{W^p_{r,s}([-2,2]^s)} \le c||f||_{W^p_{r,s}} \tag{3.6}$$

for every f in $W^p_{r,s}$. In practical applications, f itself may be defined on $[-2,2]^s$. We may then choose to work with f itself rather than $T(f)$. However, the bounds in (2.7) will then depend upon $||f||_{W^p_{r,s}([-2,2]^s)}$ rather than $||f||_{W^p_{r,s}}$. Next, let ψ be an infinitely many times continuously differentiable function that takes the value 1 on $[-1,1]^s$ and 0 outside of $[-3/2,3/2]^s$. Then the function $T(f)\psi$ coincides with f on $[-1,1]^s$, is identically 0 outside $[-3/2,3/2]^s$ and

$$||T(f)\psi||_{W^p_{r,s}([-2,2]^s)} \le c||f||_{W^p_{r,s}}. \tag{3.7}$$

In the sequel, we denote the extension $T(f)\psi$ of f again by the symbol f.

We define a 2π-periodic function from the function f (extended as above) by the formula

$$f^*(\mathbf{t}) := f(2\cos t_1, \ldots, 2\cos t_s), \qquad \mathbf{t} \in [-\pi,\pi]^s. \tag{3.8}$$

Then $f^* \in W^{p^*}_{r,s}$ and, using induction and the fact that f is identically 0 outside of $[-3/2,3/2]^s$, we conclude from (3.7) that

$$c_1||f||_{W^p_{r,s}} \le ||f^*||_{W^{p^*}_{r,s}} \le c_2||f||_{W^p_{r,s}}. \tag{3.9}$$

Now, it is easy to check that for any integer m, $v_m(f^*)$ is an even function and hence we may write

$$v_m(f^*,\mathbf{t}) =: \sum_{0\le\mathbf{k}\le 2m} V_{\mathbf{k}}(f)\prod_{j=1}^{s}\cos(k_j t_j). \tag{3.10}$$

For integer $k \ge 0$, let T_k be the Chebyshev polynomial adapted to the interval $[-2,2]$, defined by (cf. [22])

$$T_k(2\cos t) = \cos(kt), \qquad t \in [-\pi,\pi] \tag{3.11}$$

and for a multi-integer $\mathbf{k} \ge 0$, let

$$T_{\mathbf{k}}(\mathbf{x}) := \prod_{j=1}^{s} T_{k_j}(x_j), \qquad \mathbf{x} \in \mathbf{R}^s. \tag{3.12}$$

The polynomial $P_m(f)$ defined by

$$P_m(f,\mathbf{x}) := \sum_{0\le\mathbf{k}\le 2m} V_{\mathbf{k}}(f)T_{\mathbf{k}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbf{R}^s \tag{3.13}$$

is an algebraic polynomial of coordinatewise degree at most $2m$ and is related to $v_m(f^*)$ by the formula

$$P_m(f,(2\cos t_1,\ldots,2\cos t_s)) = v_m(f^*,\mathbf{t}), \qquad \mathbf{t} \in [-\pi,\pi]^s.$$

Consequently, we obtain from (3.4), (3.9) that

$$||f - P_m(f)||_p \le \frac{c}{m^r}||f||_{W^p_{r,s}}. \tag{3.14}$$

Also, in view of (3.5), (3.9), we have

$$\sum_{0\le\mathbf{k}\le 2m} |V_{\mathbf{k}}(f)| \le cm^{\alpha}||f||_{W^p_{r,s}}. \tag{3.15}$$

The next step in the proof of Theorem 2.1 is to construct an approximation to every polynomial. This is summarized in the following Lemma 3.2.

LEMMA 3.2. *Let ϕ satisfy the conditions of Theorem 2.1, $m \ge 1$ be an integer and $\mathbf{k} \ge 0$ be any multi-integer in $\mathbf{Z}^s$ with $\max_{1 \le j \le s} |k_j| \le m$. Then for every $\epsilon > 0$, there exists $G_{\mathbf{k},m,\epsilon} \in \Pi_{\phi;(6m+1)^s,s}$ such that*

$$||T_{\mathbf{k}} - G_{\mathbf{k},m,\epsilon}||_\infty \le \epsilon. \tag{3.16}$$

The weights and thresholds of each $G_{\mathbf{k},m,\epsilon}$ may be chosen from a fixed set with cardinality not exceeding $(6m+1)^s$.

Proof. First, we consider the case when $d = 1$. The point $\mathbf{b}$ in (2.6) is a real number in this case and accordingly, will be denoted by b. Let ϕ be infinitely many times continuously differentiable on $[b-\delta, b+\delta]$. For a multi-integer $\mathbf{p} = (p_1, \ldots, p_s)$, and $\mathbf{x} \in \mathbf{R}^s$, we write $\mathbf{x}^{\mathbf{p}} := \prod_{j=1}^{s} x_j^{p_j}$, where 0^0 is interpreted as 1. From the formula

$$\phi_{\mathbf{p}}(\mathbf{w}; \mathbf{x}) := \frac{\partial^{|\mathbf{p}|}}{\partial w_1^{p_1} \ldots \partial w_s^{p_s}} \phi(\mathbf{w} \cdot \mathbf{x} + b) = \mathbf{x}^{\mathbf{p}} \phi^{(|\mathbf{p}|)}(\mathbf{w} \cdot \mathbf{x} + b), \tag{3.17}$$

we conclude that

$$\mathbf{x}^{\mathbf{p}} = \left(\phi^{(|\mathbf{p}|)}(b)\right)^{-1} \phi_{\mathbf{p}}(\mathbf{0}; x). \tag{3.18}$$

Following the ideas in [9], we now replace the partial derivative $\phi_{\mathbf{p}}(\mathbf{0}; \mathbf{x})$ by an appropriate divided difference. For multi-integers $\mathbf{p}$ and $\mathbf{r}$, we write

$$\binom{\mathbf{p}}{\mathbf{r}} := \prod_{j=1}^{s} \binom{p_j}{r_j}.$$

For any $h > 0$, the network defined by the formula

$$\Phi_{\mathbf{p},h}(\mathbf{x}) := h^{-|\mathbf{p}|} \sum_{0 \le \mathbf{r} \le \mathbf{p}} (-1)^{|\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} \phi\Big(h(2\mathbf{r} - \mathbf{p}) \cdot \mathbf{x} + b\Big) \tag{3.19}$$

is in $\Pi_{\phi;(p_1+1)\cdots(p_s+1),s}$ and represents a divided difference for $\phi_{\mathbf{p}}(\mathbf{0}; \mathbf{x})$. Further, we have

$$||\Phi_{\mathbf{p},h} - \phi_{\mathbf{p}}(\mathbf{0}; \cdot)||_\infty \le M_{\phi;m,s} h^2, \qquad \max_{1 \le j \le s} |p_j| \le m, \ |h| \le \delta/(3ms) \tag{3.20}$$

where $M_{\phi;m,s}$ is a positive constant depending only on the indicated variables.

Now, we write $T_{\mathbf{k}}(\mathbf{x}) := \sum_{0 \le \mathbf{p} \le \mathbf{k}} \tau_{\mathbf{k},\mathbf{p}} \mathbf{x}^{\mathbf{p}}$, and choose

$$h := h_{\phi;m,s} := \min\left\{\frac{\delta}{3ms}, \min_{0 \le \mathbf{k} \le 2m} \left(\frac{\epsilon}{M_{\phi;m,s} \sum_{0 \le \mathbf{p} \le \mathbf{k}} \left|\phi^{(|\mathbf{p}|)}(b)\right|^{-1} |\tau_{\mathbf{k},\mathbf{p}}|}\right)^{1/2}\right\}.$$

Then (3.20) implies that the network $G_{\mathbf{k},m,\epsilon}$ defined by

$$G_{\mathbf{k},m,\epsilon}(\mathbf{x}) := \sum_{0 \le \mathbf{p} \le \mathbf{k}} \tau_{\mathbf{k},\mathbf{p}} \left(\phi^{(|\mathbf{p}|)}(b)\right)^{-1} \Phi_{\mathbf{p},h_{\phi;m,s}}(\mathbf{x}), \qquad \mathbf{x} \in [-1,1]^s \tag{3.21}$$

satisfies (3.16). For each $\mathbf{k}$, the weights and thresholds in $G_{\mathbf{k},m,\epsilon}$ are chosen from the set

$$\{(h_{\phi;m,s}\mathbf{r}, b) \ : \ \mathbf{r} \in \mathbf{Z}^s, \ |r_j| \le 3m, 1 \le j \le s\}.$$

The cardinality of this set is $(6m+1)^s$. Therefore, $G_{\mathbf{k},m,\epsilon} \in \Pi_{\phi;(6m+1)^s,s}$.

Next, if $d > 1$, and $\mathbf{b}$ is as in (2.6), then we consider the univariate function

$$\sigma(x) := \phi(x, b_2, \cdots, b_s).$$

The function σ satisfies all the hypothesis of Theorem 2.1, with b_1 in place of $\mathbf{b}$ in (2.6). Taking into account the fact that $\sigma(\mathbf{w} \cdot \mathbf{x} + b_1) = \phi(A_{\mathbf{w}}\mathbf{x} + \mathbf{b})$ with

$$A_{\mathbf{w}} := \begin{pmatrix} \mathbf{w} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix},$$

any network in $\Pi_{\sigma;n,s}$ is also a network in $\Pi_{\phi;n,s}$. Therefore, the case $d = 1$ implies the lemma also when $d > 1$. ■

Proof of Theorem 2.1. Without loss of generality, we may assume that $n \geq 13^s$. Let $m \geq 1$ be the largest integer such that $(12m + 1)^s \leq n$. We define $P_m(f) = \sum_{0 \leq \mathbf{k} \leq 2m} V_{\mathbf{k}}(f)T_{\mathbf{k}}$ as in (3.13). In view of (3.15), the network

$$\mathcal{N}_n(f, \mathbf{x}) := \sum_{0 \leq \mathbf{k} \leq 2m} V_{\mathbf{k}}(f) G_{\mathbf{k},2m,m^{-r-\alpha}}(\mathbf{x}) \tag{3.22}$$

is in $\Pi_{\phi;n,s}$ and satisfies

$$||P_m(f) - \mathcal{N}_n(f)||_\infty \leq cm^{-r}||f||_{W^p_{r,s}}.$$

Since $||g||_p \leq 2^{s/p}||g||_\infty$ for all Lebesgue measurable functions g on $[-1, 1]^s$ we get from (3.14) that

$$||f - \mathcal{N}_n(f)||_p \leq cn^{-r/s}||f||_{W^p_{r,s}}$$

as required. Further, it is quite clear that the coefficients $V_{\mathbf{k}}$ are continuous linear functionals on L^p. Hence, the continuity assertion follows. ■

We will prove Proposition 2.2 after the proof of Theorem 2.3.

Proof of Theorem 2.3. Again, we may assume that $n \geq 7^s$ and let $m \geq 1$ be the largest integer such that $(6m + 1)^s \leq n$. We write $x_{j,n} := \cos((2j + 1)\pi/(2m))$, $0 \leq j \leq m$, and use the Lagrange interpolation polynomial $L_m(f)$ at the points $\{(x_{k_1,m}, \ldots, x_{k_s,m})\}$, $0 \leq \mathbf{k} \leq m$, in place of $P_m(f)$ in the proof of Theorem 2.1. According to [20], this polynomial satisfies (2.12). Theorem 2.3 then follows as an application of Lemma 3.2 in exactly the same way as Theorem 2.1. ■

We end this section with a proof of Proposition 2.2.

Proof of Proposition 2.2. For multi-integer $\mathbf{k} \geq 0$, let

$$Z_{\mathbf{k}} := \{\mathbf{x} \in U \ : \ D^{\mathbf{k}}(\mathbf{x}) = 0\}.$$

Since (2.6) is not satisfied, we have $U = \bigcup_{\mathbf{k} \in \mathbf{Z}^d, \mathbf{k} \geq 0} Z_{\mathbf{k}}$. Now, each $Z_{\mathbf{k}}$ is a closed set and U being a closed sphere, is a complete metric space. Therefore, Baire's category theorem implies that for some multi- integer $\mathbf{r} \geq 0$, $Z_{\mathbf{r}}$ contains a nonempty interior. Hence, there exists an open sphere $N \subseteq U$ such that $D^{\mathbf{r}}\phi(\mathbf{x}) = 0$ for every $\mathbf{x} \in N$. The formula (2.9) expresses ϕ as a solution of this differential equation on N. If $d = 1$, ϕ is analytic in a closed neighborhood U of some point $x_0 \in B$ and (2.6) is not satisfied, then we have proved that ϕ is equal to a polynomial on some interval contained in U. The identity theorem of complex analysis then shows that ϕ itself is a polynomial. ■

4. Conclusions.

We have constructed generalized translation networks with a single hidden layer that provide an optimal order of approximation for functions in Sobolev classes similar to the order obtained in the classical polynomial approximation theory. If the target function is analytic, then it is possible to get a geometric rate of approximation, again similar to polynomial approximation. The weights and thresholds of our networks are chosen independently of the target function. We give explicit formulas for the coefficients, so that the "training" consists of calculating certain simple, coninuous linear functionals on the target function. The activation function for the network is fairly general, but has to satisfy certain smoothness conditions. Among the activation functions for which our theorems are applicable are the squashing function, the Gaussian function, thin plate splines and generalized multiquadric functions.

Acknowledgments. I wish to thank Professor F. Girosi and T. Poggio, MIT Artificial Intelligence Laboratories, for their kind encouragement in this research. The research was supported in part by National Science Foundation Grant DMS 9404513 and Air Force Office of Scientific Research Grant F49620-93-1-0150.

REFERENCES

1. A. R. Barron, *Universal approximation bounds for superposition of a sigmoidal function,* IEEE Trans. Information Theory, **39** (1993), 930-945.
2. A. R. Barron and R. L. Barron, *Statistical learning networks: a unified view,* in "Symposium on the Interface: Statistics and Computing Science", Reston, Virginia, April, 1988.
3. D. S. Broomhead and D. Lowe, *Multivariable functional interpolation and adaptive networks,* Complex Systems, **2** (1988), 321-355.
4. C. K. Chui, X. Li and H. N. Mhaskar, *Some limitations on neural networks with one hidden layer,* Submitted for publication.
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. F. Girosi, M. Jones and T. Poggio, *Regularization theory and neural networks architectures,* Neural Computation, **7** (1995), 219-269.
8. K. Hornik, M. Stinchcombe and H. White, *Multilayer feedforward networks are universal approximators,* Neural Networks, **2** (1989), 359-366.
9. M. Leshno, V. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function,* Neural Networks, **6** (1993), 861-867.
10. H. N. Mhaskar, *Approximation properties of a multilayered feedforward artificial neural network,* Advances in Computational Mathematics, **1** (1993), 61-80.
11. H. N. Mhaskar, *Approximation of real functions using neural networks,* in Proc. of Int. Conf. on Computational Mathematics, New Delhi, India, 1993, (H. P. Dikshit and C. A. Micchelli Eds.), World Scientific Press, 1994.
12. H. N. Mhaskar and C. A. Micchelli, *Approximation by superposition of a sigmoidal function and radial basis functions,* Advances in Applied Mathematics, **13** (1992), 350-373.
13. H. N. Mhaskar and C. A. Micchelli, *Dimension independent bounds on the degree of approximation by neural networks,* IBM Journal of Research and Development, **38** (1994), 277-284.
14. H. N. Mhaskar and C. A. Micchelli, *Degree of approximation by neural and translation networks with a single hidden layer,* to appear in Advances in Applied Mathematics.
15. J. Moody and C. Darken, *Fast learning in networks of locally tuned processing units,* Neural Computation, **1**(2) (1989), 282-294.
16. J. Park and I. W. Sandberg, *Universal approximation using radial basis function networks,* Neural Computation, **3** (1991), 246-257.
17. T. Poggio and F. Girosi, *Networks for approximation and learning,* in Proceedings of the IEEE, **78**(9), (1990).
18. T. Poggio, F. Girosi and M. Jones, *From regularization to radial, tensor and additive splines,* in "Neural Networks for Signal Processing, III", 1993, (C. A. Kamm, G. M. Kuhn, B. Yoon, R. Chellappa, S. Y. Kung Eds.), IEEE, New York, 1993, pp.3- 10.
19. M. J. D. Powell, *The theory of radial basis function approximation,* in " Advances in Numerical Analysis III, Wavelets, Subdivision Algorithms and Radial Basis Functions", (W. A. Light Ed.), Clarendon Press, Oxford, 1992, pp. 105-210.
20. J. Siciak, *On some extremal functions and their applications in the theory of analytic functions of several complex variables,* Trans. Amer. Math. Soc., **105** (1962), 322-357.

21. E. M. Stein, “Singular integrals and differentiability properties of functions”, Princeton Univ. Press, Princeton, 1970.

22. A. F. Timan, “Theory of Approximation of Functions of a Real Variable”, Macmillan Co., New York, 1963.