Dimension-independent bounds on the degree of approximation by neural networks

by H. N. Mhaskar
C. A. Micchelli

Let ϕ be a univariate 2π-periodic function. Suppose that $s \geq 1$ and f is a 2π-periodic function of s real variables. We study sufficient conditions in order that a neural network having a single hidden layer consisting of n neurons, each with an activation function ϕ, can be constructed so as to give a mean square approximation to f within a given accuracy ϵ_n, independent of the number of variables. We also discuss the case in which the activation function ϕ is not 2π-periodic.

1. Introduction

In recent years, many authors have studied the problem of approximation by neural networks (e.g., [1–4]). A neural (mapping) network is a device for highly parallel computation of functions. In this paper, we are concerned with neural networks consisting of three layers, one of them hidden. The *hidden layer* consists of a number of processors, or *neurons*, working in parallel. Each of these neurons is equipped with a local memory and is capable of performing some simple computations. A neuron is *trained* by setting the contents of its local memory. The numbers in the local memory are called the *weights*. A neuron accepts a number of real-valued inputs and evaluates a weighted sum of these inputs with the weights stored in its memory. It then calculates a *transfer function* (or *activation function*), typically nonlinear, of this weighted sum and puts out the result. Usually, we assume that one of the inputs is always 1. If the remaining inputs are represented by a vector $\mathbf{x} \in \mathbf{R}^s$, and if the activation function is $\phi : \mathbf{R} \to \mathbf{R}$, then the output of a neuron is $\phi(\mathbf{w} \cdot \mathbf{x} + b)$, where the vector $\mathbf{w}$ and the number b are stored in the local memory and $\mathbf{w} \cdot \mathbf{x}$ denotes the inner product of $\mathbf{w}$ and $\mathbf{x}$. In many models, the function ϕ is the Heaviside function, assuming the value 1 if its argument is positive and 0 otherwise. The neuron can then be thought of as a decision-maker, which *fires* if and only if the weighted sum of the inputs exceeds $-b$, a preset threshold. However, other functions are also used often and are sometimes more efficient for various applications.

The *input layer* of a network consists of a simple device that fans the input $\mathbf{x}$ to each of the neurons in the hidden layer. The *output layer* consists of a single device similar to a neuron, except that it puts out a weighted sum of its inputs without evaluating a transfer function of this sum. For a more extensive introduction to our concept of a neural network, we refer to [5].

Mathematically, the output of a network with a single hidden layer with n neurons, each evaluating a transfer function ϕ, is a special function of form $\Sigma_{k=1}^{n} c_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k)$. (The weights c_k are associated with the output-layer device.) The primary goal in the construction of a mapping network is to approximate an *arbitrary* function by such special functions. In this context, we often use the term *network* to denote the function evaluated by the network. A typical problem in this area is the *density problem*, where one seeks conditions on ϕ that ensure that an arbitrarily accurate approximation of a certain class of functions is possible. The more difficult *complexity problem* is to determine how many neurons are necessary to yield a prescribed *degree of approximation*, defined below in Equation (2.6), for every function in a given class.

A typical problem can be described more precisely as follows. Let $s \geq 1$ be an integer and $K \subset \mathbf{R}^s$ be a compact set. Given a locally square integrable function $f : \mathbf{R}^s \to \mathbf{R}$, a suitable activation function ϕ and a *tolerance* $\epsilon > 0$, one seeks an integer n, *weights* $\mathbf{w}_k \in \mathbf{R}^s$, *thresholds* $b_k \in \mathbf{R}$, and *coefficients* $c_k \in \mathbf{R}$ $(1 \leq k \leq n)$ such that

$$\left\| f(\mathbf{x}) - \sum_{k=1}^{n} c_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k) \right\|_K \leq \epsilon, \tag{1.1}$$

where $\| \cdot \|_K$ denotes the usual L^2 norm on K with respect to the s-dimensional Lebesgue measure.

There is a large amount of literature on the density problem, i.e., the possibility of such an approximation; we refer the reader to [6] for some of the references. The complexity problem, which has been studied less, is to determine the relationship between ϵ and n. Equivalently, given n, one seeks to estimate the smallest possible value of ϵ in approximating every function from a given class. For the case in which the class of functions being approximated is the class of all functions having a locally square integrable gradient, and for which the activation function ϕ is a *bounded sigmoidal function*, a particular case of our results in [6] gives $\mathcal{O}(n^{-1/(2s+2)})$ as an upper bound for the quantity ϵ in (1.1). For a different class of functions, defined in terms of the Fourier transform rather than the bounds on the gradient, Barron [1] has obtained the upper bound $\mathcal{O}(n^{-1/2})$, again for the case in which the activation function is a bounded sigmoidal function. An interesting feature of this bound is that it is independent of the number of input variables s. When the class of functions being approximated is defined in the classical manner, in terms of the bounds on the partial derivatives, it is known [7] that such a dimension-independent bound for the degree of approximation is not possible.

In this paper, we obtain an analogue of Barron's result for a large class of activation functions, not necessarily sigmoidal. As pointed out by Hecht-Nielson [3] (see also [6]), the problem of approximating any function on a compact set can be reduced to one in which the function being approximated is 2π-periodic in each of its variables. Accordingly, we consider only the approximation of 2π-periodic functions on $Q^s := [-\pi, \pi]^s$. It is then convenient to assume that the activation function ϕ is also a 2π-periodic function of one variable. We establish sufficient conditions to ensure a dimension-independent bound on the degree of approximation with the activation function ϕ. We emphasize that the actual bound itself is not the critical issue here; the novelty of our results is that the bound is dimension-independent and is valid for a large class of activation functions, not necessarily sigmoidal. We also illustrate with examples two techniques that may be used to apply our results to the case in which the activation function is not periodic.

In the next section, we formulate our main results. The proofs of all of the new results in Section 2 are given in Section 3.

2. Main results

In order to describe our main result, we need some notation. In the sequel, $s \geq 1$ is a fixed integer, $Q^s := [-\pi, \pi]^s$. For a Lebesgue-measurable function $f : Q^s \to \mathbf{R}$, we denote

$$\|f\|_s := \left\{ \frac{1}{(2\pi)^s} \int_{Q^s} |f(\mathbf{t})|^2 d\mathbf{t} \right\}^{1/2}. \tag{2.1}$$

The class of all Lebesgue-measurable functions $f : \mathbf{R}^s \to \mathbf{R}$ that are 2π-periodic in each of the s variables and for which $\|f\|_s < \infty$ is denoted by L_s^2, with the usual convention that functions which are equal almost everywhere are identified. If $f \in L_s^2$, its Fourier coefficients are defined by

$$\hat{f}(\mathbf{k}) := \frac{1}{(2\pi)^s} \int_{Q^s} f(\mathbf{t}) e^{-i\mathbf{k}\cdot\mathbf{t}} d\mathbf{t}, \qquad \mathbf{k} \in \mathbf{Z}^s. \tag{2.2}$$

If $f \in L_s^2$, we set

$$\|f\|_{SF,s} := \sum_{\mathbf{k} \in \mathbf{Z}^s} |\hat{f}(\mathbf{k})| \tag{2.3}$$

and define

$$SF_s := \{ f \in L_s^2 : \|f\|_{SF,s} < \infty \}. \tag{2.4}$$

We observe that functions in SF_s are actually continuous, but not necessarily absolutely continuous. Therefore, the

condition that $f \in SF_s$ is weaker than the periodic version of the condition in Barron's work [1]. If $\phi \in L_1^2$ and $n \geq 1$ is an integer, we define

$$\prod_{\phi,n,s} := \left\{ \sum_{k=1}^{n} a_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k) : \quad a_k, b_k \in \mathbf{R}, \quad \mathbf{w}_k \in \mathbf{Z}^s, \quad k = 1, \cdots, n \right\}. \tag{2.5}$$

The class $\Pi_{\phi,n,s}$ is the class of all possible functions that can be represented as outputs of a neural network with one hidden layer consisting of n neurons, each with an activation function ϕ, and each receiving the same input from $\mathbf{R}^s$. For the sake of convenience in proving our theorems, we assume that ϕ is 2π-periodic; therefore, to maintain this periodicity, the weights $\mathbf{w}_k$ are restricted to integers. Later, we discuss a few examples to demonstrate how these restrictions may be removed in the case of certain commonly used activation functions.

In this paper, we are interested in obtaining bounds on the degree of approximation

$$E_{\phi,n,s}(f) := \inf_{P \in \Pi_{\phi,n,s}} \|f - P\|_s, \qquad f \in SF_s. \tag{2.6}$$

The bound on $E_{\phi,n,s}(f)$ depends not just on $\|f\|_{SF,s}$ but also on ϕ. If $\Lambda \subseteq \mathbf{Z}^s$, we denote the class of all expressions of the form $\Sigma_{\mathbf{k} \in \Lambda} a_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}}$ by H_Λ. The number of elements of Λ is denoted $|\Lambda|$. For $f \in L_s^2$, its Fourier projection on H_Λ is defined by

$$P_\Lambda(f; x) := P_{\Lambda,s}(f; x) := \sum_{\mathbf{k} \in \Lambda} \hat{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} \tag{2.7}$$

and the degree of approximation from H_Λ by

$$\epsilon_\Lambda(f) := \epsilon_{\Lambda,s}(f) := \inf_{P \in H_\Lambda} \|f - P\|_s. \tag{2.8}$$

It is well known that the unique trigonometric polynomial $P \in H_\Lambda$ that attains the infimum in (2.8) is given by the Fourier projection $P_\Lambda(f)$. We define, for $f \in L_s^2$,

$$\epsilon_{n,s}(f) := \inf_{\Lambda \in \mathbf{Z}^s, |\Lambda| \leq n} \epsilon_{\Lambda,s}(f). \tag{2.9}$$

One may think of $\epsilon_{n,s}(f)$ as $E_{\psi,n,s}(f)$, where $\psi(x) = e^{ix}$. If $\lambda \geq 1$ is a real number and $\lfloor \lambda \rfloor$ denotes the largest integer not exceeding λ, we define

$$\epsilon_{\lambda,s}(f) := \epsilon_{\lfloor \lambda \rfloor,s}(f), \qquad E_{\phi,\lambda,s}(f) := E_{\phi,\lfloor \lambda \rfloor,s}(f). \tag{2.10}$$

The bounds on $E_{\phi,n,s}(f)$ are given in the following theorem, in which it is convenient to introduce one more parameter N and to estimate $E_{\phi,2nN,s}(f)$.

Theorem 2.1 *Let $s \geq 1$ be an integer, $f \in SF_s$, $\phi \in L_1^2$, and $\hat{\phi}(1) \neq 0$. Then, for integers $n, N \geq 1$,*

$$E_{\phi,2nN,s}(f) \leq \left\{ \frac{\delta_n}{\sqrt{n+1}} + \frac{2\epsilon_{N,1}(\phi)}{|\hat{\phi}(1)|} \right\} \|f\|_{SF,s}, \tag{2.11}$$

where $\{\delta_n\}$ is a sequence of positive numbers, $0 < \delta_n \leq 2$, depending upon f such that $\delta_n \to 0$ as $n \to \infty$. Moreover, the coefficients in the network that yields (2.11) are bounded, with the bound being independent of n and N.

We discuss a few immediate consequences of this theorem. From the proof of Theorem 3.1 below, it is clear that if $\phi \in SF_1$,

$$\epsilon_{N,1}(\phi) \leq \frac{\eta_N}{\sqrt{N+1}} \|\phi\|_{SF,1},$$

where $\{\eta_n\}$ is a sequence of numbers in the interval (0, 2], depending upon ϕ, and $\lim_{n\to\infty} \eta_n = 0$. Therefore, choosing $N = n$ in Theorem 2.1 leads to the following estimate.

Corollary 2.2 *If $\phi \in SF_1$,*

$$E_{\phi,2n^2,s}(f) \leq \frac{\delta_n}{\sqrt{n+1}} \left\{ 1 + \frac{\|\phi\|_{SF,1}}{|\hat{\phi}(1)|} \right\} \|f\|_{SF,s}, \tag{2.12}$$

where $\delta_n \in (0, 2]$ converges to 0 as $n \to \infty$ but may depend upon ϕ as well as f.

The error bound in (2.12) is weaker than the one given by Barron [1] but applies for a large class of activation functions that may not necessarily be sigmoidal. Moreover, the conditions on the target function f are weaker than the periodic analogue of the conditions required in [1]. An important aspect of the estimate (2.12) is that it is independent of the dimension s. We observe that this is no contradiction to the saturation results in [7], because the class of functions being approximated here is different from the class for which the results of [7] are applicable. Moreover, our proof is constructive in nature, if we know all of the Fourier coefficients of the target function and are able to preprocess them.

If ϕ is known to be a smooth function, it is possible to improve upon the bound (2.12). Thus, if

$$\hat{\phi}(k) = \mathcal{O}(e^{-\alpha|k|}) \tag{2.13}$$

for some constant $\alpha > 0$, then a simple estimation of the Fourier expansion of ϕ shows that

$$\epsilon_{c \log n,1}(\phi) = \mathcal{O}(1/n) \tag{2.14}$$

with a properly chosen constant c. In the sequel, the letter c, appearing in various formulas, denotes a positive constant independent of s, n, and f. Its value may be different at different occurrences, even within the same formula. There are standard results in approximation theory that guarantee (2.13) under certain analyticity

conditions on the function ϕ. We refer the reader to [8] for details and merely state the following corollary of (2.14).

Corollary 2.3 *If* $\hat{\phi}(k) = \mathcal{O}(e^{-\alpha|k|})$ *for some constant* $\alpha > 0$,

$$E_{\phi,n\log n,s}(f) \le \frac{\delta_n}{\sqrt{n}} \|f\|_{SF,s}, \tag{2.15}$$

where $\delta_n \in (0, 2]$ *now depends on* ϕ *as well as* f, *and* $\lim_{n\to\infty} \delta_n = 0$.

We give two applications of Theorem 2.1 in order to illustrate two techniques that can be used to apply this theorem to the more usual case, in which the activation function is not periodic.

Example 1: The squashing activation function
In this example, let

$$\sigma(x) := \frac{1}{1 + e^{-x}}. \tag{2.16}$$

Then σ is a bounded sigmoidal function. The function

$$\lambda(x) := \sigma(x + 1) - \sigma(x - 1) \tag{2.17}$$

is a hump function that satisfies

$$|\lambda(x)| = \mathcal{O}(e^{-|x|}) \qquad \text{as } |x| \to \infty. \tag{2.18}$$

Hence,

$$\phi(x) := \sum_{k\in\mathbf{Z}} \lambda(x + 2k\pi) \tag{2.19}$$

is a 2π-periodic continuous function. Using contour integration, one can easily calculate the Fourier coefficients of ϕ so as to verify that ϕ satisfies condition (2.13). Therefore, for $f \in SF_s$, (2.15) yields a function (network)

$$g(\mathbf{x}) := \sum_{1\le k\le n\log n} a_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k), \tag{2.20}$$

with properly chosen coefficients, weights, and thresholds, such that

$$\|f - g\|_s \le \frac{c}{n^{1/2}} \|f\|_{SF,s}. \tag{2.21}$$

Using (2.18), we may obtain a constant $a > 0$, such that with $N := \lfloor a \log n \rfloor$,

$$|\phi(x) - \sum_{j\in\mathbf{Z},|j|\le N} \lambda(x + 2j\pi)| \le \frac{c}{n^{1/2}}, \qquad x \in \mathbf{R}. \tag{2.22}$$

Since the coefficients a_k in the network g are bounded, independent of n, the network defined by

$$h(\mathbf{x}) := \sum_{1\le k\le n\log n} a_k \sum_{j\in\mathbf{Z},|j|\le N} \lambda(\mathbf{w}_k \cdot \mathbf{x} + b_k + 2j\pi),$$

consisting of $\mathcal{O}(n \log^2 n)$ neurons, satisfies the dimension-independent bound

$$\|f - h\|_s \le \frac{c}{n^{1/2}} \|f\|_{SF,s}.$$

Example 2: The truncated power function
In this example, let m be a fixed integer and

$$\sigma(x) := \begin{cases} x^m, & \text{if } x \ge 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{2.23}$$

Then (see [6]) the B-spline

$$B_m(x) := \frac{1}{m!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} \sigma((m + 1)x - j) \tag{2.24}$$

is an m-times continuously differentiable function and vanishes outside of $[0, 1]$, in particular, at $\pm\pi$. Therefore, one may extend B_m to $\mathbf{R}$ as a 2π-periodic function ϕ. The direct theorems of approximation theory [8] imply that

$$\epsilon_{N,1}(\phi) \le \frac{c}{N^m}.$$

Let $\nu := \lfloor n^{1+1/2m} \rfloor$ and $f \in SF_s$. Theorem 2.1 yields a network defined by

$$g(\mathbf{x}) := \sum_{k=1}^{\nu} a_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k), \tag{2.25}$$

with properly chosen coefficients, weights, and thresholds, such that

$$\|f - g\|_s \le \frac{c}{n^{1/2}} \|f\|_{SF,s}, \tag{2.26}$$

where the constant c depends upon m. The network defined by

$$h(\mathbf{x}) := \sum_{k=1}^{\nu} a_k B_m((\mathbf{w}_k \cdot \mathbf{x} + b_k) \bmod 2\pi),$$

containing $(m + 1)\nu$ neurons, then satisfies

$$\|f - h\|_s \le \frac{c}{n^{1/2}} \|f\|_{SF,s}, \tag{2.27}$$

where $c > 0$ is a constant depending on m only. We observe that the larger the value of m, the smaller (asymptotically) the number of neurons in the network h.

The proof of Theorem 3.1 below shows that when $\phi(x) = e^{ix}$ [equivalently, $\phi(x) = \cos x$ or $\phi(x) = \sin x$], it is possible to construct a network of size n to yield an approximation power of $n^{-1/2}$; i.e., in this case

$$E_{\phi,n,s}(f) \le \frac{1}{\sqrt{n+1}} \|f\|_{SF,s} . \tag{2.28}$$

We observe that a network in $\Pi_{\phi,n,s}$ is defined by $(s+2)n$ parameters. In the construction given in the proof of Theorem 2.1, these parameters do not necessarily depend continuously on the function being approximated. Theorem 2.4, given below, shows that the order of approximation given by (2.28) is the best possible for the whole class SF_s, if these parameters are to be chosen continuously.

To state this theorem, we recall some terminology from [7]. Let M be any mapping from $\mathbf{R}^N$ into L_s^2 and $\mathcal{M}_N$ the corresponding N-dimensional manifold:

$$\mathcal{M}_N := \{M(\mathbf{a}) : \mathbf{a} \in \mathbf{R}^N\}.$$

For instance, $\Pi_{\phi,n,s}$ is an $(s+2)n$-dimensional manifold in L_s^2 when $\phi \in L_s^2$. We let

$$K_s := \{f : \|f\|_{SF,s} \le 1\}. \tag{2.29}$$

The continuous n-width of K_s in L_s^2 is defined as

$$d_N^C(K_s)_{L_s^2} := \inf_{g,M} \sup_{f \in K_s} \|f - M(g(f))\|_s , \tag{2.30}$$

where the infimum is taken over all *continuous* functions $g : L_s^2 \to \mathbf{R}^N$ and manifolds (mappings) $M : \mathbf{R}^N \to L_s^2$. Thus, $d_N^C(K_s)_{L_s^2}$ measures how well we can approximate all of K_s by a continuous selection of parameters from N-dimensional manifolds in L_s^2. In particular, for $N = (s+2)n$, it gives a lower bound for continuous selections from $\Pi_{\phi,n,s}$ for *any* activation function $\phi \in L_s^2$.

Theorem 2.4 *We have*

$$d_N^C(K_s)_{L_s^2} \ge \frac{1}{\sqrt{N+1}}, \qquad N = 1, 2, \cdots . \tag{2.31}$$

To summarize our discussion intuitively, we have shown that for the approximation of a function in SF_s, the function e^{ix} is in some sense the "best" choice for a periodic activation function. Moreover, the closer an activation function is to this ideal function, the better order of approximation one obtains for the class SF_s.

3. Proofs

A crucial ingredient in our proof of Theorem 2.1 is a theorem that is similar in spirit to what is sometimes known as Jones's lemma (see [9]). Let $\mathcal{H}$ be an arbitrary, separable Hilbert space, let $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathcal{H}$, and let $\| \cdot \|$ denote the corresponding norm. Let $H := \{h_k\}_{k=1}^{\infty}$ be a complete orthonormal family in $\mathcal{H}$. Any $f \in \mathcal{H}$ can then be written in the form

$$f = \sum_{k=1}^{\infty} a_k(f) h_k , \tag{3.1}$$

where the series converges in the norm of $\mathcal{H}$. We define the set

$$S_H := \left\{ f \in \mathcal{H} : \sum_{k=1}^{\infty} |a_k(f)| \le 1 \right\}. \tag{3.2}$$

If $\Lambda \subseteq \mathbf{Z}$, we define U_Λ to be the linear span of $\{h_k : k \in \Lambda\}$, and we let T_Λ denote the projection operator onto U_Λ. We write

$$\mathcal{E}_\Lambda(f) := \inf_{h \in U_\Lambda} \|f - h\|, \qquad f \in \mathcal{H} \tag{3.3}$$

and recall that

$$\mathcal{E}_\Lambda(f) = \|f - T_\Lambda(f)\|. \tag{3.4}$$

We are interested in the quantity

$$\Delta_n := \Delta_{n,\mathcal{H},H} := \sup_{f \in S_H} \inf_{\Lambda \subseteq \mathbf{Z}, |\Lambda| \le n} \mathcal{E}_\Lambda(f), \qquad n = 1, 2, \cdots . \tag{3.5}$$

Theorem 3.1 *We have*

$$\frac{1}{2\sqrt{n}} \le \Delta_n \le \frac{1}{\sqrt{n+1}}, \qquad n = 1, 2, \cdots . \tag{3.6}$$

Moreover, if $f \in S_H$, there is a sequence $\{\delta_n\}$ of numbers such that each $\delta_n \in (0, 2]$, $\delta_n \to 0$ as $n \to \infty$, and

$$\inf_{\Lambda \subseteq \mathbf{Z}, |\Lambda| \le n} \mathcal{E}_\Lambda(f) \le \frac{\delta_n}{\sqrt{n}}, \qquad n = 1, 2, \cdots . \tag{3.7}$$

Proof Let $f \in S_H$ be arbitrary. We observe that all rearrangements of the expansion (3.1) converge in $\mathcal{H}$ to f. Therefore, we may rearrange this expansion and write

$$f = \sum_{k=1}^{\infty} d_k g_k , \tag{3.8}$$

where the set $\{g_k\}$ is the same as H and the coefficients d_k satisfy

$$|d_k| \ge |d_{k+1}| \ge 0, \qquad k = 1, 2, \cdots, \qquad \text{and} \sum_{k=1}^{\infty} |d_k| \le 1. \tag{3.9}$$

Using Parseval's identity and (3.9), we obtain

$$\begin{aligned} \inf_{\Lambda \subseteq \mathbf{Z}, |\Lambda| \le n} [\mathcal{E}_\Lambda(f)]^2 &\le \|f - \sum_{k=1}^{n} d_k g_k\|^2 \\ &= \sum_{k=n+1}^{\infty} |d_k|^2 \le |d_{n+1}| \sum_{k=1}^{\infty} |d_k| \\ &\le |d_{n+1}| . \end{aligned} \tag{3.10}$$

Using the fact [see (3.9)] that $\{|d_k|\}$ is a decreasing sequence, we obtain

$$\sum_{n/2\le k\le n+1} |d_k| \ge \frac{n|d_{n+1}|}{2}.$$

Since the series $\Sigma\,|d_k|$ converges, the left-hand side of the above inequality tends to 0 as $n \to \infty$. Moreover,

$$\sum_{n/2\le k\le n+1} |d_k| \le 1.$$

Therefore, (3.10) leads to (3.7). Again, in view of (3.9), we obtain

$$1 \ge \sum_{k=1}^{n+1} |d_k| \ge (n+1)|d_{n+1}|.$$

Together with (3.10) and the fact that $f \in S_H$ is arbitrary, this leads to the second inequality in (3.6).

Let

$$g := \frac{1}{2n}\sum_{k=1}^{2n} h_k. \tag{3.11}$$

Then $g \in S_H$, and it is obvious, using Parseval's identity, that

$$\inf_{\Lambda\subseteq\mathbf{Z},|\Lambda|\le n} [\mathcal{E}_\Lambda(g)]^2 = \sum_{k=n+1}^{2n}\left(\frac{1}{2n}\right)^2 = \frac{1}{4n}. \tag{3.12}$$

This leads to the first inequality in (3.6) and completes the proof. ■

Our proof of Theorem 2.1 requires the use of a quadrature formula in order to express certain integrals involving exponential functions as finite sums. The following lemma (see [10], Exercise 2.5.8, p. 100) provides the necessary details. For the convenience of the reader, we sketch a proof.

Lemma 3.2 *Let $n \ge 1$ be an integer. Given any continuous (complex-valued) functions $\{g_k\}_{k=1}^n$ on Q^s, there exist nonnegative numbers λ_j and vectors $\mathbf{t}_j \in Q^s$, $j = 1, \cdots, 2n+1$ (depending upon the family of functions) such that $\Sigma_{j=1}^{2n+1}\lambda_j = 1$ and*

$$\frac{1}{(2\pi)^s}\int_{Q^s} g_k(\mathbf{t})d\mathbf{t} = \sum_{j=1}^{2n+1}\lambda_j g_k(\mathbf{t}_j), \qquad k = 1, \cdots, n. \tag{3.13}$$

Proof of Lemma 3.2 First, we assume that the functions g_k are real-valued. Let $G \subseteq \mathbf{R}^n$ be the set defined by

$$G := \{(g_1(\mathbf{t}), \cdots, g_n(\mathbf{t})) : \mathbf{t} \in Q^s\}$$

and $\mathrm{co}(G)$ be its convex hull. We observe that $\mathrm{co}(G)$ is necessarily closed; hence, the definition of the Riemann integral implies that the point

$$\left(\frac{1}{(2\pi)^s}\int_{Q^s} g_1(\mathbf{t})d\mathbf{t}, \cdots, \frac{1}{(2\pi)^s}\int_{Q^s} g_n(\mathbf{t})d\mathbf{t}\right) \in \mathbf{R}^n$$

is in $\mathrm{co}(G)$. In view of the Caratheodory theorem (see [10], Theorem 2.2, p. 69), there exist $n+1$ points $\{\mathbf{u}_j\}_{j=1}^{n+1} \in \mathrm{Q}^s$ and nonnegative numbers $\{\mu_j\}_{j=1}^{n+1}$ with $\Sigma_{j=1}^{n+1}\mu_j = 1$ such that

$$\frac{1}{(2\pi)^s}\int_{Q^s} g_k(\mathbf{t})d\mathbf{t} = \sum_{j=1}^{n+1}\mu_j g_k(\mathbf{u}_j), \qquad k = 1, \cdots, n.$$

The lemma follows by considering the real and imaginary parts of g_k as separate functions. ■

Proof of Theorem 2.1 Using Theorem 3.1 and the definition of $\epsilon_{N,1}(\phi)$, we find sets $\Lambda_1 \subset \mathbf{Z}^s$ and $\Lambda_2 \subset \mathbf{Z}$ such that $|\Lambda_1| = n$, $|\Lambda_2| = N$, and, with $P_1 := P_{\Lambda_1,s}(f)$ and $P_2 := P_{\Lambda_2,1}(\phi)$,

$$\|f - P_1\|_s \le \frac{\delta_n\|f\|_{SF,s}}{\sqrt{n+1}}, \qquad \|\phi - P_2\|_1 \le 2\epsilon_{N,1}(\phi), \tag{3.14}$$

where $\delta_n \to 0$ and $\delta_n \in (0, 2]$. Without loss of generality, we may assume that $1 \in \Lambda_2$ and $\mathbf{0} \in \Lambda_1$. Let

$$\Lambda := \{1 - k : k \in \Lambda_2\}$$

and let numbers $\lambda_j \ge 0$, $t_j \in [-\pi, \pi]$ be found as in Lemma 3.2, so that $\Sigma_{j=1}^{2N}\lambda_j = 1$ and

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ikt}dt = \sum_{j=1}^{2N}\lambda_j e^{ikt_j}, \qquad k \in \Lambda. \tag{3.15}$$

[Since $0 \in \Lambda$, we may use Lemma 3.2 with the functions $\{e^{ikt}\}$ for $k \in \Lambda\setminus\{0\}$ and thus obtain (3.15)—in fact, with $2N-1$ summands on the right-hand side of (3.15), rather than $2N$ summands as above.] From the formula

$$e^{it} = \frac{1}{2\pi\hat{\phi}(1)}\int_{-\pi}^{\pi} e^{iu}\phi(t-u)du$$

$$= \frac{1}{2\pi\hat{\phi}(1)}\int_{-\pi}^{\pi} e^{iu}P_2(t-u)du \tag{3.16}$$

and (3.15), we obtain

$$P_1(\mathbf{x}) = \hat{f}(\mathbf{0}) + \sum_{\mathbf{k}\in\Lambda_1\setminus\{\mathbf{0}\}}\hat{f}(\mathbf{k})e^{i\mathbf{k}\mathbf{x}}$$

$$= \hat{f}(\mathbf{0}) + \frac{1}{2\pi\hat{\phi}(1)}\sum_{\mathbf{k}\in\Lambda_1\setminus\{\mathbf{0}\}}\hat{f}(\mathbf{k})\int_{-\pi}^{\pi} e^{iu}P_2(\mathbf{k}\cdot\mathbf{x}-u)du$$

$$= \hat{f}(\mathbf{0}) + \frac{1}{\hat{\phi}(1)}\sum_{\mathbf{k}\in\Lambda_1\setminus\{\mathbf{0}\}}\sum_{j=1}^{2N}\lambda_j\hat{f}(\mathbf{k})e^{it_j}P_2(\mathbf{k}\cdot\mathbf{x}-t_j). \tag{3.17}$$

Let

$$g(\mathbf{x}) := \hat{f}(\mathbf{0}) + \frac{1}{\hat{\phi}(1)} \sum_{\mathbf{k} \in \Lambda_1 \setminus \{\mathbf{0}\}} \sum_{j=1}^{2N} \lambda_j \hat{f}(\mathbf{k}) e^{it_j} \phi(\mathbf{k} \cdot \mathbf{x} - t_j). \quad (3.18)$$

Using (3.14) and Parseval's identity, one may easily check that for the difference $\psi(\mathbf{x}) := P_2(\mathbf{k} \cdot \mathbf{x} - t_j) - \phi(\mathbf{k} \cdot \mathbf{x} - t_j)$, where $\mathbf{k} \in \Lambda_1 \setminus \{\mathbf{0}\}$,

$$\|\psi\|_s \leq 2\epsilon_{N,1}(\phi).$$

Since

$$\sum_{k \in \Lambda_1} \sum_{j=1}^{2N} |\hat{f}(\mathbf{k})| \lambda_j \leq \|f\|_{SF,s},$$

Equation (3.17) implies that

$$\|P_1 - g\|_s \leq \frac{2\|f\|_{SF,s} \epsilon_{N,1}(\phi)}{|\hat{\phi}(1)|}.$$

Using (3.14), we obtain

$$\|f - g\|_s \leq \left\{ \frac{\delta_n}{\sqrt{n+1}} + \frac{2\epsilon_{N,1}(\phi)}{|\hat{\phi}(1)|} \right\} \|f\|_{SF,s}.$$

Since $\hat{\phi}(1) \neq 0$, it follows that $\phi(a) \neq 0$ for some $a \in [-\pi, \pi]$. Hence, we may write

$$\hat{f}(\mathbf{0}) = (\phi(a))^{-1} f(\mathbf{0}) \phi(\mathbf{0} \cdot \mathbf{x} + a).$$

Therefore, $g \in \Pi_{\phi, 2nN, s}$ and the proof is complete. ■

Finally, we prove Theorem 2.4. This is done in the more general context of a Hilbert space, as in Theorem 3.1. Thus, continuing the notation as before, we define the continuous n-width

$$d_N^C := d_N^C(S_H)_{\mathcal{H}} := \inf_{g, M} \sup_{f \in S_H} \|f - M(g(f))\|, \quad (3.19)$$

where the infimum is taken over all *continuous* functions $g : \mathcal{H} \to \mathbf{R}^N$ and manifolds (mappings) $M : \mathbf{R}^N \to \mathcal{H}$. Theorem 2.4 is then a special case of the following theorem.

Theorem 3.3 *We have*

$$d_N^C(S_H)_{\mathcal{H}} \geq \frac{1}{\sqrt{N+1}}, \qquad N = 1, 2, \cdots. \quad (3.20)$$

Proof Let X be any $N+1$-dimensional subspace of $\mathcal{H}$, and let $\rho := \rho(X)$ be the largest constant such that

$$\|f\| \geq \rho \|f\|_{S_H},$$

where, with the notation as in (3.1), $\|f\|_{S_H}$ denotes $\Sigma_{k=1}^{\infty} |a_k(f)|$. Theorem 3.1 of [7] establishes that

$$d_N^C(S_H)_{\mathcal{H}} \geq \sup\{\rho(X)\}, \quad (3.21)$$

where the supremum is taken over all $N+1$-dimensional subspaces X of $\mathcal{H}$. Now, if $\Lambda \subset \mathbf{Z}^s$ and $|\Lambda| = N+1$, we have for the expression $f = \Sigma_{k \in \Lambda} a_k h_k$,

$$\|f\|^2 = \sum_{k \in \Lambda} |a_k|^2$$

$$\geq \frac{1}{N+1} \left(\sum_{k \in \Lambda} |a_k| \right)^2 = \frac{1}{N+1} \|f\|_{S_H}^2. \quad (3.22)$$

Therefore, for the $N+1$-dimensional space $X = \text{span}\{h_k : k \in \Lambda\}$, $\rho(X) \geq (N+1)^{-1/2}$. The inequality (3.20) now follows from (3.21). ■

4. Conclusions

We have considered the problem of obtaining dimension-independent bounds for the degree of approximation of a periodic function using a neural network with a single hidden layer. Our results are applicable for a large class of target functions and activation functions. We have also obtained a lower bound for the degree of approximation. We have illustrated the application of our theory by discussing two cases in which the activation function is not periodic. Among the activation functions considered is the standard squashing activation function.

Acknowledgment

Research for this work was supported in part by Air Force Office of Scientific Research Grant No. 2-26113 and in part by the Alexander von Humboldt Foundation.

References

1. A. R. Barron, "Universal Approximation Bounds for Superposition of a Sigmoidal Function," *IEEE Trans. Info Theory* **39,** 930–945 (1993).
2. G. Cybenko, "Approximation by Superposition of Sigmoidal Functions," *Math. Control, Signals & Syst.* **2,** No. 4, 303–314 (1989).
3. R. Hecht-Nielson, "Theory of the Backpropagation Neural Network," *Proceedings of the IEEE International Conference on Neural Networks* **1,** 593–605 (1988).
4. K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks* **2,** 359–366 (1989).
5. H. N. Mhaskar, "Approximation Properties of a Multilayered Feedforward Artificial Neural Network," *Adv. Computational Math.* **1,** 61–80 (1993).
6. H. N. Mhaskar and C. A. Micchelli, "Approximation by Superposition of a Sigmoidal Function and Radial Basis Functions," *Adv. Appl. Math.* **13,** 350–373 (1992).
7. R. DeVore, R. Howard, and C. A. Micchelli, "Optimal Nonlinear Approximation," *Manuscripta Mathematica* **63,** 469–478 (1989).
8. A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Pergamon Press, Oxford (translated by J. Berry from the Russian edition), 1963.
9. L. K. Jones, "A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training," *Ann. Statist.* **20,** 608–613 (1992).
10. T. J. Rivlin, *Chebyshev Polynomials*, John Wiley & Sons, Inc., New York, 1990.

Received June 24, 1993; accepted for publication December 21, 1994

Hrushikesh N. Mhaskar *Department of Mathematics and Computer Science, California State University, Los Angeles, California 90032 (hmhaska@atss.calstatela.edu).* Dr. Mhaskar graduated in 1980 from Ohio State University, Columbus, and immediately joined California State University in Los Angeles as a faculty member. He became a full professor in 1990. He has held visiting positions at the University of Michigan, University of South Florida, Indian Institute of Technology in Bombay, Bowling Green State University in Ohio, Katholische Universität in Eichstätt, Germany, and Texas A&M University in College Station, and has lectured widely on his research. He has authored and coauthored nearly 60 papers. Dr. Mhaskar's research interests are in neural networks, orthogonal polynomials, and potential theory. In 1985, he was included in *Outstanding Young Men of America*, and in 1992 he was awarded a Humboldt Fellowship.

Charles A. Micchelli *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (CAM at YKTVMV, cam@watson.ibm.com).* Dr. Micchelli graduated in 1969 from Stanford University with a Ph.D. in mathematics, and since 1970 has been a member of the research staff of the IBM Thomas J. Watson Research Center at Yorktown Heights, NY. He has held visiting and adjunct positions at numerous universities in the U.S., Europe, South America, and Israel and has lectured widely on his research. In 1983, he was invited to the International Congress of Mathematicians in Warsaw and in 1990 was a CBMS lecturer at Kent State University. He serves on the editorial board of seven mathematical journals, is the author of the book *Mathematical Aspects of Geometric Modeling*, to be published by SIAM this year, and has written or coauthored more than 180 papers. His research interests are in computational mathematics. In 1992, Dr. Micchelli received a Humboldt Award; in 1994, he is on sabbatical at RWTH–Aachen as a visiting professor.