Journal of Approximation Theory **95**, 215–228 (1998)
Article No. AT973217

Uniform Approximation by Neural Networks*

Y. Makovoz

Department of Mathematics, University of Massachusetts at Lowell, Lowell, Massachusetts 01854

Communicated by András Kroó

Received February 17, 1997; accepted October 29, 1997

Let $D \subset \mathbf{R}^d$ be a compact set and let Φ be a uniformly bounded set of $D \to \mathbf{R}$ functions. For a given real-valued function f defined on D and a given natural number n, we are looking for a good uniform approximation to f of the form $\sum_{i=1}^n a_i \phi_i$, with $\phi_i \in \Phi$, $a_i \in \mathbf{R}$. Two main cases are considered: (1) when D is a finite set and (2) when the set Φ is formed by the functions $\phi_{v,b}(x) := s(v \cdot x + b)$, where $v \in \mathbf{R}^d$, $b \in \mathbf{R}$, and s is a fixed $\mathbf{R} \to \mathbf{R}$ function.

1

We consider the following nonlinear approximation problem. Let X be a Banach space, $f, \phi_k \in X$, $c_k \in \mathbf{R}$ $(k = 1, 2, \ldots)$, and

$$f = \sum_k c_k \phi_k, \tag{1}$$

where the sum can be finite or infinite or more generally, f can be of the form $f = \int c_\lambda \phi_\lambda \, d\mu(\lambda)$, in an appropriate setting. Given a natural number n, we want to find, *based only on* (1), a good approximation to f by a linear combination

$$g_n = \sum_{i=1}^{n} a_i \phi_{k_i} \tag{2}$$

of at most n of the ϕ_k. Maurey (see [9]) has proved that if $\Phi := \{\phi_k\}$ is a bounded set in a Hilbert space X and if $f \in \overline{\mathrm{co}}(\Phi \cup (-\Phi))$, then for every n there is a g_n for which $\|f - g_n\| = O(n^{-1/2})$. The author [7] has proved independently the same estimate for L_q, $q < \infty$, assuming that the set Φ is

* This work was supported in part by the NSF under Grant DMS-9505199.

0021-9045/98 $25.00

bounded in L_∞. Moreover (see [8]), in a Hilbert space there exists a g_n for which

$$\|f - g_n\| \leqslant 2\varepsilon_n(\Phi)\, n^{-1/2} \sum_k |c_k|, \tag{3}$$

where

$$\varepsilon_n(\Phi) := \inf\{\varepsilon > 0 : \Phi \text{ can be covered by } n \text{ sets of diameter } \leqslant \varepsilon\}. \tag{4}$$

An estimate of the same nature was proved in [8] for L_q, $q < \infty$.

An important example of the above scheme is approximation by neural networks. For $x, y \in \mathbf{R}^d$, we shall write $x \cdot y$ for the scalar product and $|x|$ for the Euclidean norm. The unit sphere $\{x \in \mathbf{R}^d : |x| = 1\}$ will be denoted by S_d. We shall denote by C any constant that does not depend on n (but may depend on d), so that C may have different values in different places, even within the same chain of equalities or inequalities. In our proofs we shall use random numbers, vectors, and functions, which we shall mark by a tilde (as in $\tilde{v}$) when we want to distinguish them from the ordinary, non-random ones.

Given a bounded set $D \subset \mathbf{R}^d$, a function $f: D \to \mathbf{R}$ called the *target function*, a function $s: \mathbf{R} \to \mathbf{R}$ called the *activation function*, and a natural number n, we want to approximate f by a function $g_n: D \to \mathbf{R}$ of the form

$$g_n(x) = \sum_{i=1}^{n} a_i s(v_i \cdot x + b_i), \tag{5}$$

with $a_i, b_i \in \mathbf{R}$, $v_i \in \mathbf{R}^d$ (a "single hidden layer feedforward neural network with n neurons"). We suppose that we already know a representation of f as a neural network,

$$f(x) = \sum_k c_k s(v_k \cdot x + b_k), \tag{6}$$

but with more than n, possibly infinitely many, terms. Of particular interest is the case when the activation function is the unit step function σ,

$$\sigma(t) := \begin{cases} 1 & \text{for} \quad t \geqslant 0 \\ 0 & \text{for} \quad t < 0. \end{cases} \tag{7}$$

In this case the quantity $\varepsilon_n(\Phi)$ of (4) can be easily estimated for $L_2(D)$, and one can prove [8] the existence of g_n of the form (5), for which $\|f - g_n\|_{L_2(D)} = O(n^{-1/2 - 1/(2d)})$ for any f of (6) with $\sum |c_i| < \infty$, an improvement over $O(n^{-1/2})$ of Barron [2]. A better than $O(n^{-1/2})$ estimate is possible [8] also for $L_q(D)$, $q < \infty$.

The case of the *uniform norm*, treated in this paper, is substantially different. The following simple example shows that in the general situation one cannot expect in this case an estimate similar to (3) even if $\varepsilon_n(\Phi)\, n^{-1/2}$ is replaced by any sequence $C_n \to 0$ independent of f.

EXAMPLE 1. In the space $C[0, 1]$ let

$$f := (2n)^{-1}\,(\phi_1 + \cdots + \phi_{2n}), \tag{8}$$

with ϕ_i defined as follows. For $m := \binom{2n}{n}$, consider the matrix $(a_{i,j})$, $i = 1, ..., 2n$, $j = 1, ...m$, in which the columns are formed by all possible $2n$-dimensional vectors with n coordinates equal to 1, the other n to zero. Let $(t_j)_1^m \subset [0, 1]$ be some fixed points and let $\phi_i \in C[0, 1]$ be any function for which $\phi_i(t_j) = a_{i,j}$. Then $f(t_j) = 1/2$ for all j. On the other hand, for any linear combination $g = \sum_{\nu=1}^n c_\nu \phi_{i_\nu}$ one has $g(t_j) = 0$ for some j, so that $\|f - g\|_C \geqslant 1/2$.

Thus good estimates for $\|f - g_n\|$ in the uniform norm cannot be as universal as in L_q, $q < \infty$. They can be valid only under some restrictions on f and $\{\phi_i\}$. Barron [1] considers the approximation in the space $L_\infty(D)$. He proves that if the ϕ_k in (1) are indicator functions of sets $D_k \subset D$ and if the family $\{D_k\}$ satisfies certain combinatorial conditions, then for every f of the form (1) with $\sum_k |c_k| \leqslant 1$ there is a g_n for which $\|f - g_n\| \leqslant Cn^{-1/2}$. This is true, in particular, when the D_k are half-spaces $v_k \cdot x + b_k \geqslant 0$, that is, in the case of neural networks with the activation function σ. Yukich, Stinchcombe, and White [12] extend Barron's result to neural networks with rather general activation functions. Moreover, they consider also the error of approximation $\|D^\alpha f - D^\alpha g_n\|$ of partial derivatives, up to a certain order.

Our main results are stated in two theorems. Theorem 1 establishes a finite dimensional analogue of (3) in the uniform norm. Let l_∞^N be the space of vectors $y = (y_1, ..., y_N)$ with the norm $\|y\| = \max_i |y_i|$, let $\Phi \subset l_\infty^N$ be a bounded set, and let $\varepsilon_n(\Phi)$ be defined by (4) in the l_∞^N metric.

THEOREM 1. *For any $f \in l_\infty^N$ of the form*

$$f = \sum_k c_k \phi_k, \qquad \phi_k \in \Phi, \qquad \sum_k |c_k| \leqslant 1 \tag{9}$$

and every natural number n there is a $g_n = \sum_{i=1}^n a_i \phi_{k_i}$ with $\sum_{i=1}^n |a_i| \leqslant 1$ for which

$$\|f - g_n\| \leqslant 4\varepsilon_n(\Phi)\, n^{-1/2} \sqrt{\log(N+1)}. \tag{10}$$

We prove this theorem in Section 2 and show how it can be used in the analysis of neural networks with continuous target and activation functions. We also discuss briefly a related problem of approximation by sparse trigonometric polynomials. In Section 3 we address specifically the case of the activation function σ. It will be convenient to take D to be the unit ball $|x| \leqslant 1$ of $\mathbf{R}^d$. Since $\sigma(\lambda t) = \sigma(t)$ for $\lambda > 0$, the v in $\sigma(v \cdot x + b)$ can be restricted to the unit sphere S_d, and we can assume, without loss of generality, that $|b| \leqslant 1$. Indeed, otherwise the neuron $\sigma(v \cdot x + b)$ is constant on D, which means that all such neurons can be represented by just one term in (5). Thus, the set $\{(v, b)\}$ of parameters can be identified with the cartesian product $Q = Q_d := S_d \times [-1, 1]$. We define the product measure μ on Q by setting $\mu := \mu_1 \times \mu_2$, where μ_1 is the (unique) rotation-invariant measure on S_d normalized by $\mu_1(S_d) = 1$ and μ_2 is the Lebesgue measure on $[-1, 1]$.

THEOREM 2. *Let $f: D \to \mathbf{R}^d$ be of the form*

$$f(x) = \int_Q c(v, b)\, \sigma(v \cdot x + b)\, d\mu, \tag{11}$$

where $c(\cdot, \cdot) \in L_\infty(Q, \mu)$, $\|c\|_\infty \leqslant 1$. Then for any natural number n there exist v_k, b_k, a_k, $k = 1, ..., n$, for which

$$\sup_{|x| \leqslant 1} \left| f(x) - \sum_{k=1}^{n} a_k \sigma(v_k \cdot x + b_k) \right| \leqslant C n^{-1/2 - 1/(2d)} \sqrt{\log n}. \tag{12}$$

Remarks. (1) In the case of uniform norm one has to distinguish between continuous linear combinations (11) and those of the form $\sum_k c_k \sigma(v_k \cdot x + b_k)$ because one cannot "round off" the parameters in this case (the set of functions $\{\sigma_{v,b} : \sigma_{v,b}(x) = \sigma(v \cdot x + b) : (v, b) \in Q\}$ is not precompact in $L_\infty(D)$).

(2) Assuming that $\|c\|_1 \leqslant 1$, Barron [2] proves for the functions (11) an estimate $\leqslant C n^{-1/2}$ in (12). Since $\mu(Q) = 2$, we have $\|c\|_1 \leqslant 2\, \|c\|_\infty$. Thus our result is incomparable, generally speaking, with that of Barron as we obtain a better estimate under stronger assumptions. However, typically $c(v, b)$ is a continuous or piecewise continuous function on Q, in which case our result is, of course, stronger.

The theorems of this paper, as well as all the above mentioned results, are proved by probabilistic methods. Again, there is a significant difference between the proofs for L_q, $q < \infty$, and for L_∞. The available proofs for $q < \infty$ deal only with *averages* (expectations or variances) of the random quantities involved. In contrast, for $q = \infty$ we shall need estimates of *probabilities* of certain events, which usually requires more sophisticated techniques. In [1] Barron derives his result from the uniform central limit theorem of Dudley whereas the authors of [12] use also some other facts from the so-called theory of empirical processes. We were unable, however, to find general results that could similarly match our needs. Instead, in our proof of Theorem 2 we modify and adapt to our construction the method of Vapnik and Chervonenkis [10] by which they prove their uniform law of large numbers.

We try to keep our exposition elementary and essentially self-contained. Although our goal is to establish only the existence of desired approximations, the proofs can serve as an outline for Monte Carlo type algorithms. The logarithmic factors that appear in (12) and other L_∞ esimates below can probably be removed. It is certainly the case in (14).

2

The following estimate (13) is well known. It belongs to the family of the so-called exponential bounds for large deviations (see, for example, [5, p. 266]).

LEMMA 1. *Let $\xi = \beta_1 \xi_1 + \cdots + \beta_n \xi_n$, where $\beta_1, ..., \beta_n$ are real numbers and $\xi_1, ..., \xi_n$ are independent random variables with $|\xi_j| \leqslant 1$, $E\xi_j = 0$, $j = 1, ..., n$. Then for every $z > 0$,*

$$P(|\xi| > z) \leqslant 2 \exp(-z^2/(4B)), \qquad B := \sum_{j=1}^{n} \beta_j^2. \tag{13}$$

Proof. The inequality $e^t - t \leqslant e^{t^2}$ is valid for all real t. It is obvious for $t \geqslant 1$ and can be easily proved for $t < 1$ using power series expansions. From this we get, for every real s,

$$E(e^{s\xi_j}) = E(e^{s\xi_j} - s\xi_j) \leqslant E(e^{s^2 \xi_j^2}) \leqslant e^{s^2}.$$

Due to the independence of the ξ_j,

$$E(e^{s\xi}) = \prod_{j=1}^{n} E(e^{s\beta_j \xi_j}) \leqslant e^{s^2 B}.$$

By the Chebyshev inequality $P(f(\xi) \geqslant a) \leqslant Ef/a$, valid for every random variable ξ, every non-negative function f, and $a > 0$, we have, for $s > 0$,

$$P(\xi \geqslant z) = P(e^{s\xi} \geqslant e^{sz}) \leqslant \exp(s^2 B - sz).$$

Taking $s = z/(2B)$, we get $P(\xi \geqslant z) \leqslant e^{-z^2/(4B)}$. Replacing ξ_j with $(-\xi_j)$ we similarly get $P(\xi \leqslant -z) \leqslant e^{-z^2/(4B)}$, and (13) follows. ∎

We illustrate the use of this lemma in the type of problems under consideration by the following example.

EXAMPLE 2. We shall prove that for the Bernoulli function $f(x) = \sum_{k=1}^{\infty} k^{-r} \cos kx$, $r > 1$, and $n = 1, 2, \dots$ there is a function $g(x) = \sum_{j=1}^{n} a_j \cos k_j x$ with at most n harmonics for which

$$\|f - g\|_{C[0, 2\pi]} \leqslant C n^{-r+1/2} \sqrt{\log n}. \tag{14}$$

For the proof we set $f = f_1 + f_2 + f_3$, where $f_1(x) := \sum_{k=1}^{n-1} k^{-r} \cos kx$, $f_2 = \sum_{k=n}^{N}$, $N := n^{(r-1/2)/(r-1)}$. Then

$$\|f_3\| \leqslant \sum_{k=N+1}^{\infty} k^{-r} \leqslant C N^{-r+1} \leqslant C n^{-r+1/2}.$$

We approximate f_2 by the random function $\tilde{g}_2$,

$$\tilde{g}_2(x) := \frac{S}{n} \sum_{i=1}^{n} \tilde{\psi}_i(x), \qquad S := \sum_{k=n}^{N} k^{-r},$$

where $\tilde{\psi}_i$, $i = 1, \dots, n$, are independent, identically distributed random functions. Each $\tilde{\psi}_i$ equals one of the $\cos k(\cdot)$ $(k = n, \dots, N)$ with the probability k^{-r}/S (more formally, the subscript i is a random variable with the range $(n, \dots, N)$). Then for every i and every fixed x we have $E(\tilde{\psi}_i(x)) = S^{-1} f_2(x)$. For a fixed x, let

$$\tilde{\xi} := f_2(x) - \tilde{g}_2(x) = \frac{2S}{n} \sum_{i=1}^{n} \tilde{\xi}_i,$$

where

$$\tilde{\xi}_i = \tilde{\xi}_i(x) := \frac{f_2(x)}{2S} - \frac{1}{2} \tilde{\psi}_i(x).$$

Then $E\tilde{\xi}_i = 0$. Since obviously $|\tilde{\psi}_i(x)| \leqslant 1$, $(1/S)\, |f_2(x)| \leqslant 1$, we have $|\tilde{\xi}_i| \leqslant 1$. Therefore, by (13), for every x and every $z > 0$,

$$P(|\tilde{\xi}| > z) \leqslant 2 \exp\left(-\frac{z^2 n}{16 S^2}\right). \tag{15}$$

Let Ω_N be the set of $4N$ points $\pi v/(2N)$, $-2N \leqslant v \leqslant 2N - 1$. It follows from (15) that

$$P(\max_{x \in \Omega_N} |\tilde{\xi}| > z) \leqslant 8N \exp\left(-\frac{z^2 n}{16 S^2}\right).$$

The latter probability can be made <1 by setting $z=CS\sqrt{(\log N)/n}$ with sufficiently large C. This means that there exists a function $g_2(x)=(S/n)\times \sum_{j=1}^{n}\cos k_j x$ (some k_j may be repeating), for which

$$\max_{x\in\Omega_N}|f_2(x)-g_2(x)|\leqslant CS\sqrt{(\log N)/n}=O(n^{-r+1/2}\sqrt{\log n}).$$

This estimate can be extended from $x\in\Omega_N$ to all x since

$$\max|T_N(x)|\leqslant A\max_{\Omega_N}|T_N(x)|$$

for some absolute constant A and any trigonometric polynomial T_N of order $\leqslant N$ (see [13, Chap. 10, (7.30)]). We obtain a desired approximation g (with $2n$ harmonics) if we set $g=f_1+g_2$. ∎

The exact order in this problem, $O(n^{-r+1/2})$, is only slightly better than (14) but it has been established with the help of much stronger tools (for references and the latest in approximation by sparse trigonometric polynomials, see [3, 4]).

Proof of Theorem 1. We use a construction similar to that of the proof of Theorem 1 in [8]. We assume without loss of generality that $f=\sum_{k=1}^{m}c_k\phi_k$, and that $m>n$ (otherwise there is nothing to prove), $c_k>0$, and $\sum_{k=1}^{m}c_k=1$. We fix some $\varepsilon>\varepsilon_n(\Phi)$ and represent the set $\Phi:=\{\phi_k\}_1^m$ as the union of n disjoint non-empty subsets Φ_ν of diameter $\leqslant\varepsilon$ (in the l_∞^N metric), so that $\Phi_\nu=\{\phi_k\colon k\in I_\nu\}$, $\bigcup_{\nu=1}^{m}I_\nu=\{1,...,m\}$. Let $f_\nu:=\sum_{k\in I_\nu}c_k\phi_k$, $S_\nu:=\sum_{k\in I_\nu}c_k$, $n_\nu:=[nS_\nu]+1$, and let

$$\tilde g_\nu:=\frac{S_\nu}{n_\nu}(\tilde\psi_1^{(\nu)}+\cdots+\tilde\psi_{n_\nu}^{(\nu)}),\qquad \tilde g:=\tilde g_1+\cdots+\tilde g_n,$$

where all the $\tilde\psi_k^{(\nu)}$, $\nu=1,...,n$, $k=1,...,n_\nu$, are independent random elements. For a fixed ν, all the $\tilde\psi_k^{(\nu)}$, $k=1,...,n_\nu$, are identically distributed; namely, each $\tilde\psi_k^{(\nu)}$ is equal to some $\phi_i\in\Phi_\nu$ with the probability $p_i^{(\nu)}:=c_i/S_\nu$.

It will be convenient to treat the elements of l_∞^N as real-valued functions of the argument $x\in\{1,...,N\}$. We have

$$f(x)-\tilde g(x)=\sum_{\nu=1}^{n}\frac{\varepsilon S_\nu}{n_\nu}\sum_{k=1}^{n_\nu}\tilde\xi_k^{(\nu)}(x),\qquad \tilde\xi_k^{(\nu)}(x):=\left[\frac{1}{S_\nu}f_\nu(x)-\tilde\psi_k^{(\nu)}(x)\right]\cdot\frac{1}{\varepsilon}. \tag{16}$$

By a straightforward computation, $E(\tilde\psi_k^{(\nu)})=(1/S_\nu)f_\nu$, hence $E(\tilde\xi_k^{(\nu)}(x))=0$ for every x. Furthermore, for each fixed ν and each x, the values of $\tilde\psi_k^{(\nu)}(x)$ are within a distance $\leqslant\varepsilon$ from each other, consequently, at a distance $\leqslant\varepsilon$

from their expectation. Hence $|\tilde{\xi}_k^{(\nu)}(x)| \leqslant 1$. We now apply Lemma 1 to the double sum representing $f(x)-\tilde{g}(x)$ in (16). We have

$$B=\sum_{\nu=1}^{n} n_\nu \cdot \frac{\varepsilon^2 S_\nu{}^2}{n_\nu{}^2} \leqslant \sum_{\nu=1}^{n} \frac{\varepsilon^2 S_\nu{}^2}{n S_\nu} = \frac{\varepsilon^2}{n}.$$

By Lemma 1, for $z>0$,

$$P(\max_{x=1, ..., N} |f(x)-\tilde{g}(x)|>z) \leqslant 2N \exp\left(-\frac{z^2 n}{4\varepsilon^2}\right).$$

For $z=4\varepsilon\sqrt{\log(N+1)/n}$ this probability is <1, which proves (10) since ε can be taken arbitrarily close to $\varepsilon_n(\Phi)$. ∎

EXAMPLE 3. Consider approximation of a function $f\colon D \to \mathbf{R}$, $D=\{x \in \mathbf{R}^d\colon |x| \leqslant 1\}$, by a neural network whose activation function $s(t)=s_h(t)$, $0<h\leqslant 1$, is a smoothed unit step function defined as a continuous function equal to zero for $t<0$, to 1 for $t>h$, and linear on $[0, h]$. Suppose that

$$f(x)=\sum_k c_k s(v_k \cdot x+b_k), \qquad |v_k|=1, \quad k=1, 2, ..., \qquad \sum_k |c_k| \leqslant 1.$$

We may assume that $|b_k| \leqslant 2$ for all k for otherwise $s(v_k \cdot x+b_k)$ is constant on D. Let $s_{v,b}(x):=s(v \cdot x+b)$. Since s satisfies the Lipschitz condition with the constant $1/h$,

$$\|s_{v,b}-s_{v',b'}\|_{C(D)} \leqslant \frac{1}{h}(|v-v'|+|b-b'|).$$

It is not hard to derive from this by a standard argument the existence of an ε-net in $C(D)$ for the set $\Phi:=\{s_{v,b}\colon |v|=1, |b| \leqslant 2\}$ containing $\sim(\varepsilon h)^{-d}$ elements. Equivalently, $\varepsilon_n(\Phi) \sim (1/h)\, n^{-1/d}$.

Similarly, for $\delta:=hn^{-3/2}$ there is a δ-net D_δ in D of cardinality $N \sim (n^{3/2}/h)^d$. By Theorem 1, we can approximate $f(x)$ by a $g(x)=\sum_{i=1}^n a_i s(v_{k_i} \cdot x+b_{k_i})$ with $\sum |a_i| \leqslant 1$ so that

$$\sup_{x \in D_\delta} |f(x)-g(x)| \leqslant (C/h)\, n^{-1/2-1/d} \log(n^{3/2}/h). \tag{17}$$

Every function $s_{v,b}$ satisfies the Lipschitz condition on D with the constant $1/h$. Since $\sum_k |c_k| \leqslant 1$ and $\sum_i |a_i| \leqslant 1$, the same is true for f and g. It follows that with our choice of δ the estimate (17) can be extended to all $x \in D$ (possibly with a different C).

Comparing (17) with the estimate (12) for the unit step function σ, we see that (17) gives a better order for $n \to \infty$ but deteriorates when $h \to 0$.

3

We shall consider random functions $\tilde{g}: D \to \mathbf{R}$, $D := \{x \in \mathbf{R}^d : |x| \leqslant 1\}$. More precisely, we shall introduce a probability space $(G, \mathscr{F}, P)$, where G is a set of $D \to \mathbf{R}$ functions, $\mathscr{F}$ is a σ-field of subsets of G, and P is a probability measure on $\mathscr{F}$. We shall consider only those functions g for which $\|g\| := \sup_{x \in D} |g(x)| < \infty$. We assume that for every $x \in D$ and $y \in \mathbf{R}$, the set $\{g \in G: g(x) < y\}$ is measurable. We shall also deal with the space of couples $(\tilde{g}, \tilde{g}^*)$ of independent random functions equipped with the product measure $P' = P \times P$. We shall assume that $\|\tilde{g}\|$ and $\|\tilde{g} - \tilde{g}^*\|$ are random variables on G or $G \times G$, respectively. In our proof of Theorem 2 these measurability assumptions will be trivially satisfied. The following lemma is rather general; the set D in it can be arbitrary.

LEMMA 2. *Let* $\sup_{g \in G} \|g\| < \infty$, *and for a fixed* $x \in D$ *let* $f(x) := E\tilde{g}(x)$ *and* $\operatorname{var} \tilde{g}(x)$ *be the expectation and variance of the random variable* $\tilde{g}(x)$. *If*

$$\varepsilon \geqslant 3\sqrt{V}, \qquad V := \sup_{x \in D} \operatorname{var} \tilde{g}(x),$$

then

$$P\{\tilde{g}: \|f - \tilde{g}\| > \varepsilon\} \leqslant 2P'\{(\tilde{g}, \tilde{g}^*): \|\tilde{g} - \tilde{g}^*\| > \varepsilon/2\}. \tag{18}$$

Proof. In the space $G \times G$ of couples $(\tilde{g}, \tilde{g}^*)$ consider two events, A and B,

$$A := \{\|\tilde{g} - \tilde{g}^*\| > \varepsilon/2\}, \qquad B := \{\|f - \tilde{g}\| > \varepsilon\}.$$

From the Chebyshev inequality

$$\operatorname{Prob}\{|\tilde{\xi} - E\tilde{\xi}| \geqslant \lambda\} \leqslant \operatorname{var}(\tilde{\xi})/\lambda^2$$

follows for every fixed $x \in D$

$$P\{\tilde{g}^*: |f(x) - \tilde{g}^*(x)| > \varepsilon/2\} \leqslant \frac{4V}{\varepsilon^2} < \frac{1}{2}. \tag{19}$$

If $\|f - g\| > \varepsilon$ for some $g \in G$, then $|f(x_0) - g(x_0)| > \varepsilon$ for some $x_0 \in D$. On the other hand, by (19),

$$P\{\tilde{g}^*: |f(x_0) - \tilde{g}^*(x_0)| \leqslant \varepsilon/2\} \geqslant 1/2,$$

so that for every such g

$$P\{\tilde{g}^*: \|g - \tilde{g}^*\| > \varepsilon/2\} \geqslant 1/2.$$

This implies, due to the independence of $\tilde{g}$ and $\tilde{g}^*$, the estimate for the conditional probability: $P'(A/B) \geqslant 1/2$. Since the event B involves only $\tilde{g}$ but not $\tilde{g}^*$, we have $P'(B) = P(B)$. Hence

$$P'(A) \geqslant P'(BA) = P'(B) \cdot P'(A/B) \geqslant P(B) \cdot (1/2),$$

as claimed. ▮

A hyperplane in $\mathbf{R}^d$ is defined by some $v \in \mathbf{R}^d$, $b \in \mathbf{R}$ as the set $\{x\colon v \cdot x + b = 0\}$. The next lemma is well known (see, for example, [6, p. 385]).

LEMMA 3. *The largest number of connected components into which n hyperplanes can split the space $\mathbf{R}^d$ does not exceed $(4en/d)^d$.*

Proof of Theorem 2. It will obviously suffice to establish (12) for some subsequence $n_m \sim m^d$, $m = 1, 2, \ldots$. We may also assume that $c(v, b) \geqslant 0$ in (11) since $c = c_1 - c_2$ with $0 \leqslant c_1, c_2 \leqslant 1$. We break the set $Q := \{(v, b)\}$ into certain subsets with disjoint interiors, which we shall call *clusters*, so that if (v, b) and (v', b') belong to the same cluster, then $|v - v'| \leqslant 1/m$, $|b - b'| \leqslant 1/m$. To this end, we cover the sphere S_d by $\sim m^{d-1}$ balls of radius $1/(2m)$ centered on S_d. That this is possible can be easily deduced from the fact that the ball $|v| \leqslant 1$ can be covered by $3^d m^d$ balls of radius $1/m$ (see [6, p. 487]). By eliminating the overlaps we obtain a covering of S_d by $\sim m^{d-1}$ disjoint subsets $A_j \subset S_d$, each of diameter $\leqslant 1/m$. The ($(d-1)$-dimensional) area of each A_j satisfies $\mu_1(A_j) \leqslant (1/m)^{d-1}$ since it obviously does not exceed the area of the sphere $|v| = 1/m$ equal to $(1/m)^{d-1} \mu_1(S_d)$. We now define clusters Q_i as the cartesian products $A_j \times \Delta_k$, where Δ_k are the intervals $[k/m, (k+1)/m]$, $k = -m, \ldots, m-1$. This gives the total of $n \sim m^{d-1} \cdot (2m) \sim m^d$ clusters Q_i, and $\mu(Q_i) \leqslant (1/m)^{d-1} \cdot (1/m) = m^{-d}$ for each i. With each cluster Q_i we associate the number

$$a_i := \int_{Q_i} c(v, b)\, d\mu, \qquad i = 1, \ldots, n. \tag{20}$$

Since $0 \leqslant c(v, b) \leqslant 1$, we have $0 \leqslant a_i \leqslant m^{-d}$. We may assume that $a_i \neq 0$ for each i for this can be always achieved by an arbitrarily small perturbation of $c(v, b)$. Let $(\tilde{v}_i, \tilde{b}_i)$ be the random point distributed on Q continuously, with the density

$$\rho_i(v, b) := \begin{cases} c(v, b)/a_i & \text{if} \quad (v, b) \in Q_i \\ 0 & \text{if} \quad (v, b) \notin Q_i. \end{cases} \tag{21}$$

We define a random approximation $\tilde{g}$ to f by setting

$$\tilde{g}(x)=\tilde{g}_n(x):=\sum_{i=1}^{n} a_i\tilde{\sigma}_i(x), \qquad \tilde{\sigma}_i(x):=\sigma(\tilde{v}_i\cdot x+\tilde{b}_i). \tag{22}$$

We assume that the $2n$ random variables

$$\tilde{v}_1, \tilde{b}_1, ..., \tilde{v}_n, \tilde{b}_n, \tilde{v}_1^*, \tilde{b}_1^*, ..., \tilde{v}_n^*, \tilde{b}_n^*$$

are independent and let P denote the corresponding product measure. We have

$$E(\tilde{\sigma}_i(x))=(1/a_i)\int_{Q_i} c(v, b)\,\sigma(v\cdot x+b)\,d\mu,$$

hence $E(\tilde{g}(x))=f(x)$ for every $x\in D$.

We now want to esimate the probability $P(\tilde{g}\colon \|f-\tilde{g}\|>\varepsilon)$ for some special ε. In order to apply Lemma 2, we estimate the variance $\operatorname{var}(\tilde{g}(x))$ for an arbitrary point $x\in D$. We have, due to the independence of the $\tilde{\sigma}_i$,

$$\operatorname{var}\tilde{g}(x)=\sum_{i=1}^{n} a_i^2\operatorname{var}(\tilde{\sigma}_i(x))\leqslant m^{-2d}\sum_{i=1}^{n}\operatorname{var}(\tilde{\sigma}_i(x)). \tag{23}$$

Since $\tilde{\sigma}_i(x)$ can take only two values, 0 or 1, we have $\operatorname{var}(\tilde{\sigma}_i(x))\leqslant 1$ for all i. Moreover, if some cluster Q_i does not contain a point (v, b) for which $v\cdot x+b=0$, then the realizations of $\tilde{\sigma}_i$ are either all 1 at x or all 0, so that $\operatorname{var}(\tilde{\sigma}_i(x))=0$. If for some $x\in D$ and some v, v', b, b' we have $v\cdot x+b=0$, $v'\cdot x+b'=0$, and $|v-v'|\leqslant 1/m$, then we also have $|b-b'|\leqslant 1/m$. It follows that for each A_j there are at most three intervals Δ_k for which the cluster $A_j\times\Delta_k$ contributes a non-zero term to the sum (23). Thus of the total number $n\sim m^d$ of summands in (23), only at most $\sim m^{d-1}$ are non-zero (the subset of non-zero summands varies with x), so that for every x

$$\operatorname{var}\tilde{g}(x)\leqslant Cm^{d-1}\cdot m^{-2d}\leqslant Cn^{-1-1/d}, \tag{24}$$

with C independent of x. This justifies the application of Lemma 2 for any $\varepsilon=Cn^{-1/2-1/(2d)}$ with sufficiently large C. According to this lemma, we need to estimate the quantity $2P'\{(\tilde{g}, \tilde{g}^*)\colon \|\tilde{g}-\tilde{g}^*\|>\varepsilon/2\}$, where $\tilde{g}$ and $\tilde{g}^*$ are two independent samples. For a fixed $x\in D$,

$$\tilde{g}(x)-\tilde{g}^*(x)=\sum_{i=1}^{n} a_i[\sigma(\tilde{v}_i\cdot x+\tilde{b}_i)-\sigma(\tilde{v}_i^*\cdot x+\tilde{b}_i^*)]. \tag{25}$$

The parameter multivector in (25),

$$w:=(\tilde{v}_1, \tilde{b}_1, ..., \tilde{v}_n, \tilde{b}_n, \tilde{v}_1^*, \tilde{b}_1^*, ..., \tilde{v}_n^*, \tilde{b}_n^*), \tag{26}$$

is a random variable distributed continuously on the cartesian product Q^{2n} with the density

$$\rho(w) := \rho_1(v_1, b_1) \cdots \rho_n(v_n, b_n)\, \rho_1(v_1^*, b_1^*) \cdots \rho_n(v_n^*, b_n^*).$$

The value of $\rho(w)$ remains invariant if any point (v_i, b_i) is interchanged with its counterpart (v_i^*, b_i^*), $i=1, ..., n$. This fact enables us to treat the choice of parameters (26) as a two-step procedure. We (1) select in each cluster Q_i two points, (v_i', b_i') and (v_i'', b_i''), and then (2) arbitrarily designate one of them as (v_i, b_i) and the other as (v_i^*, b_i^*). For every outcome of the first step there are 2^n possible outcomes of the second step, all with the same probability 2^{-n}. For a fixed x and fixed v_i', b_i', v_i'', b_i'', $i=1, ..., n$, we have

$$\tilde{g}(x) - \tilde{g}^*(x) = \sum_{i=1}^{n} \beta_i \tilde{\theta}_i, \qquad \beta_i := a_i[\sigma(v_i' \cdot x + b_i') - \sigma(v_i'' \cdot x + b_i'')],$$

where $\tilde{\theta}_1, ..., \tilde{\theta}_n$ are independent random variables equal to 1 or to -1, each with the probability 1/2, with the corresponding probability measure P_0 defined on the vectors $\theta := (\theta_1, ..., \theta_n)$ by setting $P_0(\theta) = 2^{-n}$ for every θ. By Lemma 1 we obtain for fixed β_i

$$2P_0\{|\tilde{g}(x) - \tilde{g}^*(x)| > \varepsilon/2\} \leqslant 4 \exp\left(-\frac{(\varepsilon/2)^2}{4B}\right). \tag{27}$$

We have $|\beta_i| \leqslant m^{-d}$, and the number of non-zero β_i is at most $\sim m^{d-1}$ (for the reason explained in the derivation of (24)), hence $B = \sum_{i=1}^{n} \beta_i^2 \leqslant Cm^{-d-1}$. From this and (27),

$$2P_0\{|\tilde{g}(x) - \tilde{g}^*(x)| > \varepsilon/2\} \leqslant 4 \exp(-A\varepsilon^2 m^{d+1}), \tag{28}$$

where A depends only on d (but does not depend on x or m). For fixed v_i', b_i', v_i'', b_i'', $i=1, ..., n$, the function $\tilde{g}(x) - \tilde{g}^*(x)$ is piecewise constant on D, and by Lemma 3, the number N of subsets on which it is constant does not exceed $(8en/d)^d$. Consequently, the norm $\|\tilde{g} - \tilde{g}^*\|$ equals the maximum of $|\tilde{g}(x) - \tilde{g}^*(x)|$ on some set of N points which can be considered fixed while $\{v_i', b_i', v_i'', b_i''\}_1^n$ remain fixed. In view of (28),

$$2P_0\{\|\tilde{g} - \tilde{g}^*\| > \varepsilon/2\} \leqslant 4N \exp(-A\varepsilon^2 m^{d+1}). \tag{29}$$

Since this estimate of conditional probability does not depend on the condition (that is, on the choice of v_i', b_i', v_i'', b_i'', $i=1, ..., n$), P_0 in (29) can be replaced by P'. Therefore by Lemma 2,

$$P(\tilde{g}\colon \|f - \tilde{g}\| > \varepsilon) \leqslant 4N \exp(-A\varepsilon^2 m^{d+1}). \tag{30}$$

Since $n = Cm^d$, the probability (30) will be <1 if we set

$$\varepsilon := Cn^{-1/2-1/(2d)}\sqrt{\log N} = Cn^{-1/2-1/(2d)}\sqrt{\log n}$$

with sufficiently large C. This fact implies the existence of $g(x) = \sum_{i=1}^{n} a_i\sigma(v_i \cdot x + b_i)$ satisfying (12). ▌

Remarks. (1) A similar approach can be used in the case of more general networks $\sum_k c_k\sigma(P_k(x))$, where P_k are polynomials in d variables of degrees not exceeding l. A relevant result is the following generalization [11] of Lemma 3: For any set of n polynomials, the number of connected components into which the surface $P_1 \cdots P_n = 0$ splits $\mathbf{R}^d$ is at most $(4eln/d)^d$.

(2) Sufficient conditions for the validity of certain error estimates in neural network approximation can be expressed in terms of the Fourier transform. Barron [1] proves that if a function $f\colon D \to \mathbf{R}$ can be extended to $f \in L_1(\mathbf{R}^d)$ with

$$\int_{\mathbf{R}^d} |\omega|\,|\hat{f}(\omega)|\,d\omega < \infty, \tag{31}$$

then for each $n = 1, 2, \ldots$ there is a $g_n = \sum_k a_k\sigma(v_k\dot{x} + b_k)$ for which $\|f - g_n\|_{L_\infty(D)} = O(n^{-1/2})$. In the spherical coordinates (31) becomes

$$\int_{S_d} d\mu_1(v)\int_0^\infty r^d\,|\hat{f}(rv)|\,dr < \infty.$$

We claim that under a stronger condition

$$\sup_{v \in S_d}\int_0^\infty r^d\,|\hat{f}(rv)|\,dr < \infty, \tag{32}$$

there is a g_n for which $\|f - g_n\| = O(n^{-1/2-1/(2d)}\sqrt{\log n})$. Indeed, for every $a > 0$ and $|t| \leqslant a$,

$$e^{it} = e^{-ia} + i\int_{-a}^{a} \sigma(t-\tau)\,e^{i\tau}\,d\tau.$$

Hence for $|x| \leqslant 1$, $\omega \neq 0$,

$$e^{i\omega\cdot x} = e^{-i|\omega|} + i\int_{-|\omega|}^{|\omega|} \sigma(\omega\cdot x - \tau)\,e^{i\tau}\,d\tau.$$

By a change of variables, $r := |\omega|$, $v := \omega/r$, and since $\sigma(\lambda t) = \sigma(t)$, $\lambda > 0$, we get

$$e^{irv \cdot x} = e^{-ir} + ir \int_{-1}^{1} \sigma(v \cdot x - \tau)\, e^{ir\tau}\, d\tau.$$

Substituting this into the inverse transform

$$f(x) = \int_{\mathbf{R}^d} \tilde{f}(\omega)\, e^{i\omega \cdot x}\, d\omega = \int_{S_d} d\mu_1(v) \int_0^\infty r^{d-1} \hat{f}(rv)\, e^{irv \cdot x}\, dr,$$

we obtain

$$f(x) = C_f + \int_Q c(v, \tau)\, \sigma(v \cdot x - \tau)\, d\mu, \qquad c(v, \tau) := i \int_0^\infty r^d \hat{f}(rv)\, e^{ir\tau}\, dr.$$

Therefore, due to (32), the function $f(x) - C_f$ satisfies, up to a constant factor, the conditions of Theorem 2, which justifies our claim.

REFERENCES

1. A. R. Barron, Neural net approximation, *in* "Yale Workshop on Adaptive and Learning Systems," Yale Univ. Press, New Haven, CT, 1992.
2. A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory* **39**, No. 3 (1993), 930–945.
3. E. Belinskii, Decomposition theorems and approximation by a "floating" system of exponentials, *Trans. Amer. Math. Soc.*, to appear.
4. R. A. DeVore and V. N. Temlyakov, Nonlinear approximation by trigonometric sums, *J. Fourier Anal. Appl.* **2**, No. 1 (1995), 29–48.
5. M. Loève, "Probability Theory," Springer-Verlag, Berlin, 1987.
6. G. G. Lorentz, M. v. Golitschek, and Y. Makovoz, "Constructive Approximation. Advanced Problems," Springer-Verlag, Berlin, 1996.
7. Y. Makovoz, On trigonometric n-widths and their generalization, *J. Approx. Theory* **41**, No. 4 (1984), 361–366.
8. Y. Makovoz, Random approximants and neural networks, *J. Approx. Theory* **85**, No. 1 (1996), 98–109.
9. G. Pisier, Remarques sur un résultat non publié de B. Maurey, *in* "Séminaire d'analyse fonctionelle 1980–1981, École Polytechnique, Centre de Mathématiques, Palaiseau."
10. V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Teor. Veroyatnost. i Primenen.* **16**, No. 2 (1971), 264–279. [Russian; English transl., *Theory Probab. Appl.*]
11. H. E. Warren, Lower bounds for approximation by non-linear manifolds, *Trans. Amer. Math. Soc.* **133**, No. 2 (1968), 167–179.
12. J. E. Yukich, M. B. Stinchcombe, and H. White, Sup-norm approximation bounds for networks through probabilistic methods, *IEEE Trans. Inform. Theory* **41**, No. 4 (1995), 1021–1027.
13. A. Zygmund, "Trigonometric Series," Cambridge Univ. Press, Cambridge, 1959.