JOURNAL OF APPROXIMATION THEORY **85**, 98–109 (1996)
ARTICLE NO. 0031

Random Approximants and Neural Networks

Y. MAKOVOZ

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

Communicated by Will Light

Received June 22, 1994; accepted in revised form April 4, 1995

Let D be a set with a probability measure μ, $\mu(D)=1$, and let K be a compact subset of $L_q(D,\mu)$, $1\leqslant q<\infty$. For $f\in L_q$, $n=1,2,...$, let $\rho_n(f,K)=\inf\|f-g_n\|_q$, where the infimum is taken over all g_n of the form $g_n=\sum_{i=1}^n a_i\phi_i$, with arbitrary $\phi_i\in K$ and $a_i\in\mathbf{R}$. It is shown that for $f\in\overline{\operatorname{conv}(K\cup(-K))}$, under some mild restrictions, $\rho_n(f,K)\leqslant C_q\varepsilon_n(K)\,n^{-1/2}$, where $\varepsilon_n(K)\to 0$ as $n\to\infty$. This fact is used to estimate the errors of certain neural net approximations. For the latter, also the lower estimates of errors are given.

1

Let K be a bounded set in a real Banach space X. Given $f\in X$ and a natural number n, we consider approximations to f of the form $\sum_{i=1}^n a_i\phi_i$, with arbitrary $\phi_i\in K$ and real coefficients a_i. Approximations by splines with free knots or by rational functions with free poles can be interpreted in this way. Here we study approximations by linear combinations of the so-called sigmoidal functions which appear in the theory of neural networks.

A possible approach to finding a good approximation of the above type is to start with an approximation of the form $\sum_{i\in I}c_i\phi_i$ with a set I of arbitrary, possibly infinite, cardinality, and then reduce the cardinality to n. The most obvious idea is to aggregate the terms by replacing clusters of close ϕ_i by single representatives. Another idea is to aggregate the terms in such a way that would cause mutual cancellation of errors. The latter approach is realized in [10], where we prove, for $X=L_q$, $q<\infty$, that under rather general assumptions, for each n one can find $g_n=\sum_{i=1}^n a_i\phi_i$ for which $\|f-g_n\|=O(n^{-1/2})$. Recently the author learned that the case $q=2$ had been considered earlier by Maurey (see [11]). Here we combine both aggregating ideas and obtain refinements of the above results.

Our proofs are based on some elementary probabilistic considerations (although the problem in question is, of course, non-probabilistic). We deal

0021-9045/96 $18.00

with random variables ξ that can take only a finite number of real values $x_1, ..., x_n$ with the probabilities $p_1, ..., p_n$, $p_i > 0$, $\sum_{i=1}^n p_i = 1$. The set of (x_i, p_i) is called the *distribution* of ξ; each x_i is a *realization* of ξ, and $E(\xi) = \sum_{i=1}^n x_i p_i$ is the *expectation* of ξ. Given another random variable η with the values $y_1, ..., y_m$ and probabilities $q_1, ..., q_m$, one may consider the set of all couples (x_i, y_j) and corresponding probabilities $p_{i,j}$. If $p_{i,j} = p_i q_j$, the variables ξ, η are called *independent*. The sum $\xi + \eta$ is defined as a random variable taking values $x_i + y_j$ with the probabilities $p_{i,j}$; the product $\xi\eta$, as well as sums and products of more than two random variables, are defined in the same way. If $\xi = \sum \xi_v$, $\eta = \sum \eta_j$, and each ξ_i is independent of each η_j, then ξ and η are also independent. For arbitrary ξ, η, one has $E(\xi + \eta) = E(\xi) + E(\eta)$. If ξ, η are independent, then also $E(\xi\eta) = E(\xi) \cdot E(\eta)$. These identities remain valid for ξ, η taking values in the Hilbert space H, with x_i, y_j, $E(\xi)$, $E(\eta) \in H$ and with $\xi\eta$ and $E(\xi) \cdot E(\eta)$ treated as scalar products. The number $var(\xi) = \sum_i \|x_i - E(x)\|^2 p_i$ is called the *variance* of ξ. For a constant (non-random) c, $var(c\xi) = c^2\, var(\xi)$, $var(c + \xi) = var(\xi)$. For independent ξ, η, $var(\xi + \eta) = var(\xi) + var(\eta)$.

The paper is organized as follows. In 2 we prove a refinement of Maurey's theorem for sets in the Hilbert space. In 3 we obtain an L_q result of the same nature for $q < \infty$, which is a refinement of a similar statement in [10], with a new, self-contained, and simpler proof. In 4 we consider applications of these results to neural net approximations. Finally, in 5 we show that the results of 4 cannot be essentially improved.

2

Let K be a bounded set in the Hilbert space H and let

$$\varepsilon_n(K) = \inf\{\varepsilon > 0 : K \text{ can be covered by at most } n \text{ sets of diameter } \leqslant \varepsilon\}. \tag{1}$$

THEOREM 1. *Let $\Phi := \{\phi_1, \phi_2, ...\}$ be an arbitrary bounded sequence of elements of H. For every $f \in H$ of the form*

$$f = \sum_i c_i \phi_i, \qquad \sum_i |c_i| < \infty, \tag{2}$$

and for every natural number n, there is a $g = \sum_i a_i \phi_i$ with at most n non-zero coefficients a_i and with $\sum_i |a_i| \leqslant \sum_i |c_i|$, for which

$$\|f - g\| \leqslant 2\varepsilon_n(\Phi)\, n^{-1/2} \sum_i |c_i|. \tag{3}$$

Proof. Without loss of generality, we may assume that the sum in (2) has only a finite number of terms, $f = \sum_{i=1}^{N} \phi_i$. Moreover, we may assume that $c_i > 0$, $i = 1, ..., N$ (since f either has this property or is a difference of two functions that have it), and that $\sum_i c_i = 1$. For a given n and some fixed $\varepsilon > \varepsilon_n(\Phi)$, we can break the set $\{1, 2, ..., N\}$ into n non-empty subsets I_ν, $\nu = 1, ..., n$, so that the sets $\Phi_\nu := \{\phi_i : i \in I_\nu\}$ are of diameter $\leqslant \varepsilon$. We approximate each $f_\nu := \sum_{i \in I_\nu} c_i \phi_i$ by a linear combination $\sum_{i \in I_\nu} a_i \phi_i$ with a small number m_ν of non-zero a_i. To this end, we set $S_\nu := \sum_{i \in I_\nu} c_i$, $m_\nu := [nS_\nu] + 1$, and define the random elements

$$\hat{f}_\nu := (S_\nu / m_\nu)(\hat{\psi}_1^{(\nu)} + \cdots + \hat{\psi}_{m_\nu}^{(\nu)}), \qquad \hat{f} := \hat{f}_1 + \cdots + \hat{f}_n, \tag{4}$$

where the $\hat{\psi}_k^{(\nu)}$, $k = 1, ..., m_\nu$, are identically distributed; each $\hat{\psi}_k^{(\nu)}$ equals one of the $\phi_i \in \Phi_\nu$ with the probability $p_i^{(\nu)} := c_i / S_\nu$. We further assume that all the $\hat{\psi}_k^{(\nu)}$, $\nu = 1, ..., n$, $k = 1, ..., m_\nu$, are pairwise independent. We have

$$E(\hat{f}_\nu) = \frac{S_\nu}{m_\nu} \sum_{k=1}^{m_\nu} E(\hat{\psi}_k^{(\nu)}) = \frac{S_\nu}{m_\nu} m_\nu \sum_{i \in I_\nu} \frac{c_i}{S_\nu} \phi_i = f_\nu,$$

so that $E(f_\nu - \hat{f}_\nu) = 0$, hence $E(f - \hat{f}) = 0$.

It follows from the properties of the variance, since $f_\nu - \hat{f}_\nu$ are obviously independent, that

$$E(\|f - \hat{f}\|^2) = var(f - \hat{f}) = \sum_{\nu=1}^{n} var(f_\nu - \hat{f}_\nu) = \sum_{\nu=1}^{n} var(\hat{f}_\nu).$$

All possible realizations of each $\hat{\psi}_k^{(\nu)}$ are in the corresponding set Φ_ν of diameter $\leqslant \varepsilon$, hence $var(\hat{\psi}_k^{(\nu)}) \leqslant \varepsilon^2$ and

$$var(\hat{f}_\nu) = \frac{S_\nu^2}{m_\nu^2} \sum_{k=1}^{m_\nu} var(\hat{\psi}_k^{(\nu)}) \leqslant \frac{\varepsilon^2 S_\nu^2}{m_\nu} \leqslant \frac{\varepsilon^2}{n} S_\nu,$$

consequently, $E(\|f - \hat{f}\|^2) \leqslant (\varepsilon^2/n) \sum_{\nu=1}^{n} S_\nu = \varepsilon^2/n$. Therefore for some realization f^* of $\hat{f}$ must be $\|f - f^*\| \leqslant \varepsilon/\sqrt{n}$. This completes the proof since f^* is a linear combination of at most $n \sum S_\nu + n = 2n$ elements ϕ_i and ε can be chosen arbitrarily close to $\varepsilon_n(\Phi)$. ∎

The above proof can be, of course, carried out without recourse to probability theory. One may say that a good approximation to f of (2) is chosen from the finite set of possible candidates of the form $\hat{f} := \hat{f}_1 + \cdots + \hat{f}_n$, with each $\hat{f}_\nu$ given by (4) and each $\hat{\psi}_k^{(\nu)}$ in (4) selected arbitrarily from the set Φ_ν (thus, there are $|I_1|^{m_1} \cdots |I_n|^{m_n}$ candidates). To show that there exists an f^* with a small norm $\|f - f^*\|$, we assign a weight $\lambda(\hat{f}) > 0$ to each $\hat{f}$ and estimate the sum $\sum \lambda(\hat{f}) \|f - \hat{f}\|^2$ over all possible $\hat{f}$. In this context, the requirement of independence of the $\hat{\psi}_k^{(v)}$ is just a special way of defining $\lambda(\hat{f})$ by means of the numbers $p_i^{(v)} = c_i/S_v$.

As we have already mentioned, Maurey established (3) without the factor $\varepsilon_n(\Phi)$. For a precompact Φ and $n \to \infty$, we have $\varepsilon_n(\Phi) \to 0$, so our estimate is better. Lee Jones [7] gave a non-probabilistic proof of Maurey's result; he found an iterative algorithm that produces successively, for $n = 1, 2, \ldots$, the functions g of Theorem 1 with $\|f - g\| = O(n^{-1/2})$.

3

Let D be a set with a probability measure μ. We shall prove an analogue of Theorem 1 for the space $L_q = L_q(D, \mu)$, $1 \leqslant q < \infty$. We assume here that $\Phi = \{\phi_1, \phi_2, \ldots\}$ is a bounded set in L_∞ (and therefore in all L_q): $\|\phi\|_\infty \leqslant 1$, $i = 1, 2, \ldots$. Let $\varepsilon_n(\Phi)$ be the quantity (1), with the diameters of sets in the L_2 norm.

THEOREM 2. *For $1 \leqslant q < \infty$, every $f \in L_q$ of the form (2) and every natural number n, there is a $g = \sum_i a_i \phi_i$, with at most n non-zero coefficients a_i, $\sum_i |a_i| \leqslant \sum_i |c_i|$, for which*

$$\|f - g\|_q \leqslant C_q \varepsilon_n(\Phi)^{2/q^*} n^{-1/2} \sum_i |c_i|, \tag{5}$$

where q^* is the minimal even integer satisfying $q^* \geqslant q$.

LEMMA 1. *Let ξ, η be two independent, identically distributed random variables, $E(\xi) = E(\eta) = 0$. Then $E(|\xi|^q) \leqslant E(|\xi - \eta|^q)$, $1 \leqslant q < \infty$.*

Proof. We have

$$E(|\xi|^q) = \sum_i |x_i|^q p_i = \sum_i \left| x_i - \sum_j x_j p_j \right|^q p_i = \sum_i \left| \sum_j (x_i - x_j) p_j \right|^q p_i.$$

Applying Jensen's inequality to the inner sum, we get

$$E(|\xi|^q) \leqslant \sum_i \sum_j |x_i - x_j|^q p_j p_i = E(|\xi - \eta|^q). \quad \blacksquare$$

To prove Theorem 2, we proceed as in the proof of Theorem 1. We assume that $c_i > 0$, $\sum_i c_i = 1$. Then we fix some $\varepsilon > \varepsilon_n(\Phi)$, define Φ_v, f_v, and $\hat{f}_v$, and approximate the given f by the random element

$$\hat{f} = \sum_{v=1}^{n} \hat{f}_v = \sum_{v=1}^{n} (S_v/m_v)(\hat{\psi}_1^{(v)} + \cdots + \hat{\psi}_{m_v}^{(v)}).$$

For notational convenience, we now relabel arbitrarily the elements $(S_\nu/m_\nu)\,\hat{\psi}_k^{(\nu)}$, $\nu=1, ..., n$, $k=1, ..., m_\nu$, into a single-index sequence $\hat{\xi}_j$, so that $\hat{f}=\hat{\xi}_1+\cdots+\hat{\xi}_m$, where $m:=m_1+\cdots+m_n$. To estimate $\|f-\hat{f}\|_q$, we consider independent random elements η_j, $j=1, ..., m$, distributed identically with the corresponding $\hat{\xi}_j$ and independent of the latter. Let $\hat{g}:=\sum_j \hat{\eta}_j$, $\hat{u}_j:=\hat{\xi}_j-\hat{\eta}_j$, $\hat{u}:=\hat{f}-\hat{g}=\sum_{j=1}^m \hat{u}_j$. Since the random element $f-\hat{f}$ has only a finite set of realizations, we obviously have

$$E\left(\int_D |f(t)-\hat{f}(t)|^q\,d\mu\right)=\int_D E(|f(t)-\hat{f}(t)|^q)\,d\mu.$$

By Lemma 1, since $E(f-\hat{f})=E(f-\hat{g})=0$, for $t\in D$,

$$E(|f(t)-\hat{f}(t)|^q)\leqslant E(|(f(t)-\hat{f}(t))-(f(t)-\hat{g}(t))|^q)=E(|\hat{u}(t)|^q).$$

To prove (5), we may assume that $q=q^*$, that is, that q itself is an even integer. Then

$$|\hat{u}(t)|^q=\sum \frac{q!}{q_1!\cdots q_m!}\,\hat{u}_1(t)^{q_1}\cdots\hat{u}_m(t)^{q_m},$$

where the sum is extended to all combinations $(q_1, ..., q_m)$ of non-negative integers with $q_1+\cdots+q_m=q$. Since $\{\hat{u}_j(t)\}$ are independent random variables, we have

$$E(|\hat{u}(t)|^q)=\sum \frac{q!}{q_1!\cdots q_m!}\,E(\hat{u}_1(t)^{q_1})\cdots E(\hat{u}_m(t)^{q_m}). \tag{6}$$

For each random function $\hat{u}_j(t)$, its possible values for a fixed t are of the form $(S_\nu/m_\nu)(\phi_i(t)-\phi_{i'}(t))$, with the probabilities $p_i^{(\nu)}\cdot p_{i'}^{(\nu)}$ and with ϕ_i, $\phi_{i'}$ belonging to the same Φ_ν. It follows that $\hat{u}_j(t)$ is a symmetric random variable, that is, if $y\in\mathbf{R}$ is one of its possible realizations, then so is $-y$, with the same probability. If q_j is odd, then $\hat{u}_j(t)^{q_j}$ is also symmetric, hence $E(\hat{u}_j(t)^{q_j})=0$. Therefore, in (6) only those terms are non-zero in which all $q_1, ..., q_m$ are even. Since $\|\phi_i\|_\infty\leqslant 1$, for every realization $u_j(t)$ of $\hat{u}_j(t)$ we have $\|u_j\|_\infty\leqslant 2S_\nu/m_\nu\leqslant 2/n$ for $j=1, ..., m$. At the same time, since ϕ_i, $\phi_{i'}$ belong to the same Φ_ν, we have $\|\phi_i-\phi_{i'}\|_2\leqslant\varepsilon$, hence $\|u_j\|_2\leqslant\varepsilon/n$ for each j. In every non-zero term of the sum (6) we have $q_j\geqslant 2$ for at least one j. Consequently, in view of the above estimates,

$$\int_D E(\hat{u}_1(t)^{q_1})\cdots E(\hat{u}_m(t)^{q_m})\,d\mu\leqslant(\varepsilon/n)^2\,(2/n)^{q-2}=2^{q-2}n^{-q}\varepsilon^2,$$

and

$$E(\|f-\hat{f}\|_q^q) \leqslant E(\|\hat{u}\|_q^q) \leqslant 2^{q-2} n^{-q} \varepsilon^2 \sum \frac{q!}{q_1! \cdots q_m!}, \tag{7}$$

where the sum is over the set Q of all combinations $(q_1, ..., q_m)$ of non-negative even q_j with $\sum_1^m q_j = q$. This sum is obviously $\leqslant q!\,|Q|$. Now $|Q|$ is also the number of terms in the expansion of $(x_1^2 + \cdots + x_m^2)^{q/2}$, hence $|Q| \leqslant (1 + \cdots + 1)^{q/2} = m^{q/2} \leqslant (2n)^{q/2}$. With these estimates, we obtain from (7) $E(\|f-\hat{f}\|_q^q) \leqslant C\varepsilon^2 n^{-q/2}$, with C depending only on q. The proof can be now concluded as in Theorem 1. ∎

More accurate estimates (see, for example, [12], Ch. 5, §8) show that one can take $C_q \leqslant 2\sqrt{q}$ in (5).

In a weaker form, without the factor $\varepsilon_n(\Phi)$, the inequality (5) was obtained in [10]. A generalization of this weaker result to a class of Banach spaces that includes L_q, $q < \infty$, can be found in [5]. The proofs in [10] and [5] are based on deeper probabilistic arguments.

In some cases one can use a flexible strategy, applying the above results only to some selected parts of the expansion (2).

Example (from [10]). Let $f(t) := \sum_{k=1}^{\infty} k^{-r} \cos kt$, $r > 0$. If $r > 1$, then for a given n we write $f = \sum_{k=1}^{[n/2]} + \sum_{k=[n/2]}^{\infty}$, and for $q \geqslant 2$ apply Theorem 2, with $n/2$ instead of n, to the second sum. As a result, we obtain a function $g(t) := \sum_{v=1}^{n} a_v \cos k_v t$ for which $\|f-g\|_q = O(n^{-r+1/2})$, $q \geqslant 2$, while approximation of f by conventional (that is, with $k_v = v$) trigonometric polynomials of order n gives only $O(n^{-r+1-1/q})$. For $0 < r < 1$, we write $f = \sum_{k=1}^{[n/2]} + \sum_{k=[n/2]}^{m^q} + \sum_{k=m^q}^{\infty}$ and apply Theorem 2 to the second sum; this leads to $\|f-g\|_q = O(n^{(q/2)(1-r)-1/2})$. Both estimates provide the best possible orders for the error of approximation of f by trigonometric polynomials with $\leqslant n$ frequencies. The factor $\varepsilon_n(\Phi)$ of (5) yields no improvement here since the set $\{\cos kt\}_{k=1}^{\infty}$ is not precompact in L_q. For further results in this direction see [3].

Let Ω, K be two sets in a Banach space X, and let

$$\rho_n(\Omega) := \rho_n(\Omega, K, X) := \sup_{f \in \Omega} \inf \|f - g_n\|_X,$$

with the infimum over all g_n of the form $g_n = \sum_{i=1}^{n} a_i \phi_i$, $\phi_i \in K$, $a_i \in \mathbf{R}$. Let $\rho_n^*(\Omega)$ be the same quantity, with the additional condition $\sum_{i=1}^{n} |a_i| \leqslant 1$. Obviously, $\rho_n \leqslant \rho_n^*$.

From Theorem 2 immediately follows

COROLLARY. *For an arbitrary set* $K \subset X$, *let* $K^c := \overline{\operatorname{conv}(K \cup (-K))}$. *Then*

$$\rho_n^*(K^c, K, L_q) \leqslant C_q \varepsilon_n(K)^{2/q^*} n^{-1/2}, \qquad 1 \leqslant q < \infty \tag{8}$$

where C_q *depends only on* q.

Note that under the assumptions of Theorem 2, $\varepsilon_n(K) \leqslant 2$ for all n.

4

Given a real-valued function f defined on a bounded set $D \subset \mathbf{R}^d$ and a natural number n, consider approximations to f of the form

$$g(x) = \sum_{i=1}^{n} a_i s(v_i x + b_i), \qquad x \in D, \quad a_i, b_i \in \mathbf{R}, \quad v_i \in \mathbf{R}^d, \tag{9}$$

where s: $\mathbf{R} \to \mathbf{R}$ is some fixed function. Approximations of this form appear in the theory of neural networks. Here we assume for simplicity that D is an open convex set in $\mathbf{R}^d$ equipped with the Lebesgue measure. A bounded measurable function s: $\mathbf{R} \to \mathbf{R}$ is called *sigmoidal*, if $s(t) \to 1$ for $t \to +\infty$, $s(t) \to 0$ for $t \to -\infty$. One can prove ([4], see also [6]) that for a sigmoidal s, every $f \in C(D)$ can be uniformly approximated, with arbitrarily small error, by functions (9) with suitable n, a_i, v_i, b_i. The most important s is the unit step function

$$\sigma(t) := \begin{cases} 1 & \text{if } \ t \geqslant 0, \\ 0 & \text{if } \ t < 0. \end{cases}$$

Since obviously $\sigma(\lambda t) = \sigma(t)$, $\lambda > 0$, for $s = \sigma$ we may assume $|v_i| = 1$ in (9) (here and below $|\cdot|$ is the Euclidean norm in $\mathbf{R}^d$).

With Barron, we consider the class $V = V_D$, the closure in $L_q(D)$ of the set of all functions f: $\mathbf{R}^d \to \mathbf{R}$ of the form

$$f(x) = \sum_i c_i \sigma(v_i x + b_i), \qquad \sum_i |c_i| \leqslant 1, \quad |v_i| = 1.$$

For $d = 1$, D is an interval, and if $f \in V$, then f is a function of bounded variation: $\operatorname{Var}(f) \leqslant 1$ on D. Conversely, every f: $\mathbf{R} \to \mathbf{R}$ with $\operatorname{Var}(f) \leqslant 1$ is of the form $f = f_0 + \text{const.}$, $f_0 \in V$. Moreover, for $d \geqslant 1$, if g: $\mathbf{R} \to \mathbf{R}$ and $\operatorname{Var}(g) \leqslant 1$ on a sufficiently large interval (depending on D), then, for a sufficiently small $\gamma = \gamma(D) > 0$, all functions $\gamma g(vx + b)$, $|v| = 1$, belong to V. In particular, for some $\gamma > 0$, we have $\gamma\, |\omega|^{-1} e^{i\omega x} \in V$ for all $\omega \in \mathbf{R}^d$, $\omega \neq 0$.

From this one immediately deduces, since V is a convex, symmetric and closed set, that if the Fourier transform of some f satisfies $C_f := \gamma^{-1} \int_{\mathbf{R}^d} |\omega| \, |\hat{f}(\omega)| \, d\omega < 1$, then $f \in V$. Functions f with this property have been extensively studied in [2].

For a given function s: $\mathbf{R} \to \mathbf{R}$, we define the set

$$\mathscr{A}_s := \{ s_{v,b} : s_{v,b}(x) = s(vx + b), \, v \in \mathbf{R}^d, \, b \in \mathbf{R} \}.$$

THEOREM 3. *For $1 \leqslant q < \infty$, $n = 1, 2, \ldots$, and every sigmoidal function s,*

$$\rho_n^*(V, \mathscr{A}_s, L_q) \leqslant C n^{-1/2 - 1/(q^* d)}, \qquad 1 \leqslant q < \infty, \tag{10}$$

where C depends only on D, $s(t)$, and q.

This theorem does not cover the case $q = \infty$. However, in [1] Barron proves, using a deep combinatorial theorem of Dudley, that $\rho_n^*(V, \mathscr{A}_s, L_\infty) = O(n^{-1/2})$, which implies $\rho_n^* = O(n^{-1/2})$ for all L_q, $q < \infty$. As we see, for $q < \infty$ this estimate can be improved. The improvement is significant for small d and disappears when $d \to \infty$. It should be noted that the estimate (10) is given for the *whole* class $\mathscr{A}_s$. For individual functions, Theorem 2 can give a better rate depending on the behavior of the corresponding $\varepsilon_n(\Phi)$.

It is sufficient to prove (10) only for the case $s = \sigma$. Indeed, if s is an arbitrary sigmoidal function, then $s(\lambda t) \to \sigma(t)$, $\lambda \to +\infty$, uniformly on every set $|t| \geqslant a > 0$; on $[-a, a]$ the difference $\sigma(t) - s(\lambda t)$ remains bounded. It follows that $\|\sigma_{v,b} - s_{v,b}\|_{L_q(D)}$ can be made arbitrarily small by taking a sufficiently large $\|v\|$.

For $s = \sigma$ we can use (8) since obviously $V = (\mathscr{A}_\sigma)^c$. We need an estimate for $\varepsilon_n(\mathscr{A}_\sigma)$. We may consider only the $\sigma_{v,b}$ with $|v| = 1$. If D is contained in some ball $|x| \leqslant r$, then we may assume that $|b| \leqslant r$ for otherwise $\sigma_{v,b}$ is identically 1 or 0 on D. Suppose that $|v - v_1| < \varepsilon$, $|b - b_1| < \varepsilon$ for some $\varepsilon > 0$. If $v = v_1$ and, say, $b > b_1$, then $\sigma_{v,b} - \sigma_{v_1,b_1}$ is equal to ± 1 on the strip $-b \leqslant vx \leqslant -b_1$ of width $\leqslant \varepsilon$, and to zero elsewhere. Similarly, if $b = b_1$, then $\sigma_{v,b} - \sigma_{v_1,b_1} \neq 0$ only on a strip of width $O(\varepsilon)$. It follows that $\|\sigma_{v,b} - \sigma_{v_1,b_1}\| \leqslant C\sqrt{\varepsilon}$ in L_2. (Here and below C stands for various constants independent of n). Therefore we obtain an $O(\sqrt{\varepsilon})$-net for $\mathscr{A}_\sigma$ in $L_2(D)$ if we find an ε-net for the set $P := \{(v, b) \in \mathbf{R}^{d+1} : |v| = 1, \, |b| \leqslant r\}$. By a standard volume ratio argument, one needs $O((1/\varepsilon)^{d-1})$ elements to build an ε-net for the sphere $|v| = 1$ and $O(1/\varepsilon)$ elements for the interval $[-r, r]$, which gives $O(\varepsilon^{-d})$ elements for P. Consequently, one can find an ε-net for $\mathscr{A}_\sigma$ in L_2 consisting of $O(\varepsilon^{-2d})$ elements. Thus $\varepsilon_n(\mathscr{A}_\sigma) = O(n^{-1/(2d)})$, and (10) now follows from (8).

5

The estimate (10) cannot be essentially improved. We show this for $q=2$, in which case the right-hand side of (10) (with $q^*=2$) is the smallest.

If $d=1$, $D=[0, 2\pi]$, we take $f_0(x)=(2n)^{-1}$ sign sin $nx \in V_{[0, 2\pi]}$. One can easily see that $\|f_0-g\|_2 \geqslant Cn^{-1}$ for any piecewise constant function g with $\leqslant n$ breaks, which shows, for $s=\sigma$, that in this case $\rho_n^* \geqslant \rho_n \geqslant Cn^{-1}$, matching the upper estimate (10). The same is true for more general s. It is not clear, however, how to construct a similar extremal function for $d \geqslant 2$, so we use an indirect approach based on the concept of metric entropy.

For a precompact set K in a metric space and $\varepsilon>0$, the ε-entropy is defined by $H_\varepsilon(K):=\log_2 N_\varepsilon$, where N_ε is the minimal n for which there exists an ε-net for K consisting of n points. To estimate $H_\varepsilon(K)$ from below, one can find in K a large number M_ε of elements that are 2ε-distinguishable, that is, are at a distance $>2\varepsilon$ from each other. Then, clearly, $H_\varepsilon(K) \geqslant \log_2 M_\varepsilon$.

We say that a sigmoidal function s: $\mathbf{R} \to \mathbf{R}$ belongs to the class S if (a) s satisfies a Lipschitz condition $|s(t)-s(t')| \leqslant M\,|t-t'|$ for some M and all t, t' and (b) $|s(t)-\sigma(t)| \leqslant C\,|t|^{-\gamma}$ for some C, $\gamma>0$ and all $t \neq 0$.

Lemma 2. *Let $s=\sigma$ or $s \in S$. Then for any $\varepsilon>0$, the set $\mathscr{A}_s$ has a finite ε-net in $L_2(D)$ with the number of elements that grows polynomially in $1/\varepsilon$ for $\varepsilon \to 0$.*

Proof. The case $s=\sigma$ has been already considered in the proof of Theorem 3, with the ε-net of cardinality $O(\varepsilon^{-2d})$. If $s \in S$, then $\|s_{v,b}-\sigma_{v,b}\|_2 \leqslant \varepsilon$ if $|v| \geqslant R=R(\varepsilon)=O(|\varepsilon|^{-\gamma})$. It follows that the set $\{s_{v,b}: |v| \geqslant R\}$ has an ε-net of cardinality $O(\varepsilon^{-2d})$. On the other hand, if $s \in S$, then $\|s_{v,b}-s_{v',b'}\|_2 \leqslant \varepsilon$ for $|v-v'| \leqslant C\varepsilon$, $|b-b'| \leqslant C\varepsilon$. Therefore the set $\{s_{v,b}: |v| \leqslant R\}$ has an ε-net of $(R/\varepsilon)^d$ elements. Thus for some $l>0$ and every $\varepsilon>0$ the whole set $\mathscr{A}_s$ has an ε-net of $O(|\varepsilon|^{-l})$ elements. ∎

Let D be an open and convex subset of $\mathbf{R}^d$.

Theorem 4. *If $s=\sigma$ or $s \in S$, then for $d \geqslant 2$*

$$\rho_n^*(V, \mathscr{A}_s, L_2(D)) \geqslant Cn^{-1/2-1/d-\eta}, \tag{11}$$

where $\eta>0$ can be taken arbitrarily small, $C=C(D,\eta)$. For $d=2$ a better estimate is available:

$$\rho_n^* \geqslant Cn^{-3/4-\eta}. \tag{12}$$

The estimate (11) can be found in Barron's paper [1]. The outlined proof (by reducing (11) to a statistical problem of non-parametric estimation) seems to be rather involved. Our derivation of (11) and (12) below is more transparent and essentially self-contained. Note that for $d=2$ our lower estimate (12) is an almost exact match to the upper estimate (10).

We need a simple lemma about the entropy of almost orthogonal sequences in the Hilbert space H.

LEMMA 3. *Let $K \subset H$ be a set containing m elements $\phi_1, ..., \phi_m$ with the property*

$$\sum_{k=1, k \neq i}^{m} |(\phi_i, \phi_k)| \leqslant (1/2) \|\phi_i\|^2, \qquad i=1, ..., m, \tag{13}$$

and let $K^c := \overline{\mathrm{conv}(K \cup (-K))}$. If $\varepsilon := m^{-1/2} \min_i \|\phi_i\|$, then $H_\varepsilon(K^c) \geqslant Cm$, where C is an absolute constant.

Proof. Consider the 2^m elements $g_\theta \in K^c$ of the form

$$g_\theta := m^{-1}(\theta_1 \phi_1 + \cdots + \theta_m \phi_m), \qquad \theta_i = \pm 1.$$

We use the following elementary fact (see, for example, [9]): for each sufficiently large m, there is a set Σ_m consisting of $\geqslant (4/3)^m$ sign vectors $\theta = (\theta_i)_1^m$, so that any two vectors in Σ_m are different in more than $[m/8]$ places. If $\theta, \theta' \in \Sigma_m$, then

$$g_\theta - g_{\theta'} = m^{-1}(\xi_1 \psi_1 + \cdots + \xi_r \psi_r), \qquad \xi_i = \pm 2, \quad i=1, ..., r, \quad r \geqslant [m/8],$$

where $\{\psi_1, ..., \psi_r\}$ is a subset of $\{\phi_1, ..., \phi_m\}$. Hence

$$\|g_\theta - g_{\theta'}\|^2 = m^{-2} \sum_{j,k=1}^{r} a_{j,k} \xi_j \xi_k, \qquad a_{j,k} := (\psi_j, \psi_k).$$

We have $\xi_1^2 + \cdots + \xi_r^2 = 4r$, so $\|g_\theta - g_{\theta'}\|^2 \geqslant 4rm^{-2}\mu$, where μ is the minimum of the quadratic form $\sum_{j,k=1}^r a_{j,k} y_j y_k$ on the unit sphere $y_1^2 + \cdots + y_r^2 = 1$, which is equal to the smallest eigevalue of the Gramm matrix $A = [a_{j,k}]_{j,k=1}^r$. All the eigenvalues are contained (see, for example, [8]) in the Gerschgorin intervals $|\lambda - a_{j,j}| \leqslant \sum_{k: k \neq j} |a_{j,k}|$, $j=1, ..., r$, so that due to (13) $\mu \geqslant (1/2) \min \|\psi_i\|^2 \geqslant (1/2) \min \|\phi_i\|^2$. Therefore the g_θ are $O(\varepsilon)$-distinguishable:

$$\|g_\theta - g_{\theta'}\| \geqslant m^{-1} \sqrt{2r} \min \|\phi_i\| \geqslant Cm^{-1/2} \min \|\phi_i\|.$$

Adjusting the constants, we obtain the statement of the lemma. ∎

Proof of Theorem 4. We may assume that $D=[0,\pi]^d$, since every convex open $D\subset\mathbf{R}^d$ contains a cube. We suppose that $\rho_n^*:=\rho_n^*(V,\mathscr{A}_s,L_2(D))\leqslant Cn^{-\alpha}$ for some C, $\alpha>0$ and estimate $H_\delta(V)$ for $\delta:=Cn^{-\alpha}$. By Lemma 2, for some $l>0$ the set $\mathscr{A}_s$ has an δ-net $\mathscr{A}_s^\delta$ consisting of $O(|\delta|^{-l})$ elements. Likewise, the ball $|c|_1=\sum_{i=1}^n|c_i|\leqslant 1$ of the space l_1^n has a δ-net Λ^δ in l_1^n consisting of δ^{-n} elements. We obtain an $O(\delta)$-net for the set of all linear combinations

$$g=\sum_{i=1}^n c_i\phi_i,\qquad \phi_i\in\mathscr{A}_s\qquad \sum_{i=1}^n|c_i|=1, \tag{14}$$

by taking g with $\phi_i\in\mathscr{A}_s^\delta$, $c=(c_i)_1^n\in\Lambda^\delta$. These g form a set of cardinality $\leqslant C(\delta^{-n})(\delta^{-l})^n$. Since by assumption every $f\in V$ can be approximated by some g of (14) with an error $\leqslant\delta$, we have the inequality $H_\delta(V)\leqslant C_1 n\log n$, with some C_1 independent of n.

To estimate $H_\delta(V)$ from below, we use Lemma 3. As we have noted, there is a $\gamma>0$ for which all the functions $g_\omega(x):=\gamma\,|\omega|^{-1}\sin\omega x$, $\omega\neq 0$, belong to V. Let $R:=(1/\delta)^{1/(1+d/2)}$, $\delta=Cn^{-\alpha}$. The functions g_ω corresponding to the integer vectors ω with $|\omega|\leqslant R$ are pairwise orthogonal in $L_2(D)$ (hence satisfy (13)), and $\min\|g_\omega\|=O(1/R)$. The number of these g_ω is $m\sim C_dR^d$. For ε of Lemma 3 we have $\varepsilon\sim(R\sqrt{m})^{-1}\sim\delta$, hence $H_\delta(V)\geqslant Cm\geqslant Cn^{\alpha/(1/2+1/d)}$. Comparing this with the upper estimate for $H_\delta(V)$, we have $C_1 n\log n\geqslant Cn^{\alpha/(1/2+1/d)}$. If we now assume that $\alpha=1/2+1/d+\eta$, $\eta>0$, then for large n we come to a contradiction which implies the inequality (11).

For $d=2$ we can use another construction. This time we take D to be the disk $|x|^2=x_1^2+x_2^2\leqslant 1$. Assuming that $\rho_n^*\leqslant\delta=Cn^{-\alpha}$ we get as before $H_\delta(V)\leqslant C_1 n\log n$. To obtain a lower estimate for $H_\delta(V)$, we choose an integer N from the condition $N^{3/2}(\log N)^{1/2}\sim 1/\delta$, and set $h:=a/(N\log N)$, with $a>0$ to be chosen later. We define the function $g\colon\mathbf{R}^2\to\mathbf{R}$, by setting $g(x)=g(x_1,x_2):=\operatorname{sign}x_1$ for $|x_1|\leqslant h/2$, $g(x):=0$ otherwise. Clearly, $g\in(1/4)\,V_D$. Let $g_{k,l}(x):=g(v_kx+b_l)$, with $v_k:=(\cos 2\pi k/N,\ \sin 2\pi k/N)$, $b_l:=l/N$, and let G be the set of all $g_{k,l}$ with $k=1,\dots,N$, $l=0,\pm1,\dots,\pm[N/2]$. The cardinality of G is $m\sim N^2$, and for $g_{k,l}\in G$, $\min\|g_{k,l}\|\sim\sqrt{h}$.

Most $g_{k,l}$ are pairwise orthogonal in $L_2(D)$. Indeed, let k, l be fixed. We have $g_{k,l}\perp g_{k,l'}$ if $l\neq l'$. If $k\neq k'$, then $g_{k,l}\perp g_{k',l'}$ for all those l' for which the support of the product $g_{k,l}(x)\,g_{k,l'}(x)$ is either completely inside or completely outside of D. It is not hard to see that for each $k'\neq k$ the scalar product $(g_{k,l},g_{k',l'})$ is $\neq 0$ for at most three values of l'; for these l',

$$|(g_{k,l},g_{k',l'})|\leqslant\frac{h^2}{\sin(|k-k'|/N)}\leqslant\frac{Ch^2N}{|k-k'|},$$

hence for fixed (k, l),

$$\sum_{(k', l') \neq (k, l)} |(g_{k,l}, g_{k',l'})| \leqslant Ch^2 N \sum_{j=1}^{N} j^{-1} \leqslant Ch^2 N \log N = Cah.$$

It follows that condition (13) for the functions $g_{k,l} \in G$ is fulfilled if a is sufficiently small, and we can use Lemma 3, with $\varepsilon = \sqrt{h}/\sqrt{N^2} \sim N^{-3/2}(\log N)^{-1/2} \sim \delta$. We have $H_\delta(V) \geqslant Cm \geqslant CN^2$, so that for $\delta = Cn^{-\alpha}$ must be $CN^2 \leqslant n \log n$, which is possible only if $\alpha < 3/4$. Thus for $d \geqslant 2$ we have $\rho_n^* \geqslant Cn^{-3/4-\eta}$, with arbitrarily small η. ∎

It is unclear whether the above construction can be modified for $d \geqslant 3$. Another open question is the lower estimate for $\rho_n(V)$, rather than for $\rho_n^*(V)$. The estimates (11) and (12) remain valid if in the definition of ρ_n^* one requires $\sum |c_i| \leqslant M$, with arbitrarily large $M > 0$ (C in (11) and (12) may depend on M), but it is not known if it is valid for unrestricted c_i. It would be interesting to exhibit an individual function $f \in V$ that is poorly approximable in the sense of (11) and (12).

References

1. A. R. Barron, Neural net approximation, *in* "Yale Workshop on Adaptive and Learning Systems," Yale University, 1992.
2. A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory* **39** (1993), 930–945.
3. E. S. Belinskii, Approximation by a "floating" system of exponentials on classes of smooth periodic functions, *Mat. Sb.* **132**, No. 1 (1987); *Mat. USSR Sb.* **60**, No. 1 (1988), 19–27.
4. G. Cybenko, Approximations by superpositions of a sigmoidal function, *Math. Control, Signal, Systems* **2** (1989), 303–314.
5. C. Darken, M. Donahue, L. Gurvits, and E. Sontag, Rate of approximation results motivated by robust neural network learning, extended abstract of Siemens Technical Report LS93-07, Siemens, 1993.
6. L. K. Jones, Constructive approximations for neural networks by sigmoidal functions, *Proc. IEEE* **78**, No. 10 (1990), 1586–1589.
7. L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural networking training, *Ann. of Statist.* **20** (1992), 608–613.
8. P. Lankaster, "Theory of Matrices," Academic Press, New York, 1969.
9. G. G. Lorentz, Metric entropy and approximation, *Bull. Amer. Math. Soc.* **72** (1966), 903–937
10. Y. Makovoz, On trigonometric *n*-widths and their generalization, *J. Approx. Theory* **41**, No. 4 (1984), 361–366.
11. G. Pisier, Remarques sur un résultat non publié de B. Maurey, *in* "Séminaire d'analyse fonctionelle 1980–1981, École Polytechnique, Centre de Mathématiques, Palaiseau."
12. A. Zygmund, "Trigonometric Series," Cambridge Univ. Press, Cambridge, 1959.