

Lower Bounds for Approximation by MLP Neural Networks

Vitaly Maiorov and Allan Pinkus

Abstract. The degree of approximation by a single hidden layer MLP model with n units in the hidden layer is bounded below by the degree of approximation by a linear combination of n ridge functions. We prove that there exists an analytic, strictly monotone, sigmoidal activation function for which this lower bound is essentially attained. We also prove, using this same activation function, that one can approximate arbitrarily well any continuous function on any compact domain by a two hidden layer MLP using a fixed finite number of units in each layer.

Key Words. multilayer feedforward perceptron model, degree of approximation, lower bounds, Kolmogorov Superposition Theorem.

§1. Introduction

This paper is concerned with the multilayer feedforward perceptron (MLP) model. A lower bound on the degree to which the single hidden layer MLP model with n units in the hidden layer and a single output can approximate any function is given by the extent to which a linear combination of n ridge functions can approximate this same function. We prove that this lower bound is essentially attainable by an MLP model whose activation function is sigmoidal, strictly increasing, and analytic. We also prove, using this same activation function, that there is no theoretical lower bound on the error of approximation if we permit two hidden layers.

Ridge functions are multivariate functions of the form

$$g(a_1x_1 + \cdots + a_dx_d) = g(\mathbf{a} \cdot \mathbf{x})$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$, $\mathbf{a} = (a_1, \dots, a_d) \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ is a fixed *direction*, and $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ are the variables. In other words, they are multivariate functions constant on the parallel hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$, $c \in \mathbb{R}$. Ridge functions have been considered in the study of hyperbolic partial differential equations (where they go under the name of *plane waves*), computerized tomography, projection pursuit, approximation theory, and neural networks (see Pinkus [19] for further details).

Ridge functions are relevant in the theory of neural networks as they appear in the single hidden layer MLP model. The mathematical expression for the output in this model with d inputs (d units in the input layer) given by $\mathbf{x} = (x_1, \dots, x_d)$, a single hidden layer

with n units (neurons), a continuous activation function σ , and a single output (with linear activation function and no threshold), is given by

$$\sum_{j=1}^n c_j \sigma(\mathbf{w}^j \cdot \mathbf{x} - \theta_j). \quad (1.1)$$

Here the $\mathbf{w}^j = (w_1^j, \dots, w_d^j)$ are the *weights* between the input and hidden layer, the θ_j the bias or threshold (shift), and the c_j the weights between the hidden and outer layer.

Each factor

$$\sigma(\mathbf{w}^j \cdot \mathbf{x} - \theta_j)$$

is a ridge function. As such a *lower bound* on the extent to which this MLP model with n units in the single hidden layer can approximate any function is given by the approximation order from the manifold

$$M_n = \left\{ \sum_{j=1}^n g_j(\mathbf{a}^j \cdot \mathbf{x}) : \mathbf{a}^j \in \mathbb{R}^d, g_j \in C(\mathbb{R}), j = 1, \dots, n \right\}.$$

In Maiorov [13] (see also Oskolkov [18] for the case $d = 2$) are determined upper and lower bounds on the degree of approximation from M_n to some Sobolev type spaces of functions with derivatives of all orders up to r in L^2 , defined on the unit ball in \mathbb{R}^d . Without going into the details, as they are not relevant in what follows, they prove that for all functions in this set one can approximate in the L^2 norm from M_n to within approximation error

$$c_1 n^{-r/(d-1)}$$

where c_1 is some constant independent of n . It is also proven that for each n there exists a function in the set for which one cannot approximate from M_n with approximation error less than

$$c_2 n^{-r/(d-1)}$$

for some other constant c_2 independent of n . In Maiorov, Meir, Ratsaby [14] it is shown that the set of functions for which this lower bound holds is of large measure. See Maiorov [13] and Maiorov, Meir, Ratsaby [14] for details. The point we wish to make here is twofold. Firstly, in the single hidden layer MLP model, ridge function approximation is a lower bound on the degree of approximation. Secondly, ridge function approximation itself is bounded below (away from zero) with some dependence on n (depending on the set to be approximated).

In Section 2 we prove that there exists an activation function σ for which single hidden layer MLP approximation (i.e., approximation by functions of the form (1.1)) is essentially identical (same approximation order) to that of ridge function approximation. In other words the theoretical lower bound given by ridge function approximation can be attained. This is not in the least surprising. What is somewhat unexpected is that we are able to do this with an activation function σ which is analytic (as smooth as is possible), strictly increasing, and sigmoidal (i.e., $\lim_{t \rightarrow -\infty} \sigma(t) = 0$ and $\lim_{t \rightarrow \infty} \sigma(t) = 1$).

In Section 3 we consider the MLP model with two hidden layers. (Two hidden layers are modeled by an iteration of (1.1).) We prove, using the above activation function, that in this situation there is no theoretical lower bound on the error of approximation. To be more precise, we prove that for our constructed activation function, any continuous function on the unit cube in \mathbb{R}^d can be uniformly approximated to within any error by a two hidden layer MLP with $3d$ units in the first hidden layer and $6d + 3$ units in the second hidden layer.

The activation function used in the above results is pathological, and this brings us to what we consider to be one of the implications of these results. It is that the properties of analyticity, strict monotonicity and sigmoidality are not truly significant in this context (although they may be for other purposes). That is, our pathologies may be hidden among these very *nice* properties, as translation and composition are very powerful tools.

MLP's with a single hidden layer have been much studied these last few years. There is now a fairly reasonable basic theoretical understanding of their approximation properties, although much remains to be done. This is not the case with this same model with more than one hidden layer. The results of Section 3 suggest that these multi-hidden layer models are well worth further study.

We do not, for one moment, suggest that one try to construct and use the above mentioned σ . This σ is wonderfully smooth but unacceptably complex. Theoretical results such as the above have a different purpose. They are meant to tell us what is possible and, sometimes more importantly, what is not. They are also meant to explain why certain things are or are not possible by highlighting their salient characteristics.

§2. Construction of σ

We prove two results in this section. The first result, which may also be of interest, will be used to prove our main result.

Let B^d denote the unit ball in \mathbb{R}^d , i.e., $B^d = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$, and S^{d-1} its boundary, i.e., $S^{d-1} = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$. If we consider M_n restricted to B^d , then it may be written as

$$M_n = \left\{ \sum_{j=1}^n g_j(\mathbf{a}^j \cdot \mathbf{x}) : \mathbf{a}^j \in S^{d-1}, g_j \in C[-1, 1] \right\}.$$

Proposition 1. *There exists a function ϕ which is C^∞ , strictly increasing, and sigmoidal satisfying the following. Given $f \in M_n$ and $\varepsilon > 0$, there exist real constants c_i , integers r_i and vectors $\mathbf{w}^i \in S^{d-1}$, $i = 1, \dots, n + d + 1$, such that*

$$\left| f(\mathbf{x}) - \sum_{i=1}^{n+d+1} c_i \phi(\mathbf{w}^i \cdot \mathbf{x} - r_i) \right| < \varepsilon$$

for all $\mathbf{x} \in B^d$.

Proof: The space $C[-1, 1]$ is separable. That is, it contains a countable dense subset. Let $\{u_k\}_{k=1}^\infty$ be such a subset. Thus to each $g \in C[-1, 1]$ and each $\varepsilon > 0$ there exists a m (dependent upon g and ε) for which

$$|g(t) - u_m(t)| < \varepsilon$$

for all $t \in [-1, 1]$. Assume each u_k is in $C^\infty[-1, 1]$. (We can, for example, choose the $\{u_k\}_{k=1}^\infty$ from among the set of all polynomials with rational coefficients.)

We will now construct a sigmoidal function ϕ , i.e., for which $\lim_{t \rightarrow -\infty} \phi(t) = 0$ and $\lim_{t \rightarrow \infty} \phi(t) = 1$, which is strictly increasing and in C^∞ and is such that for each $g \in C[-1, 1]$ and $\varepsilon > 0$ there exists an integer m and real coefficients a_1^m , a_2^m , and a_3^m such that

$$|g(t) - (a_1^m \phi(t - 7) + a_2^m \phi(t - 3) + a_3^m \phi(t + 4m + 1))| < \varepsilon$$

for all $t \in [-1, 1]$. We do this by constructing ϕ so that $a_1^k \phi(t - 7) + a_2^k \phi(t - 3) + a_3^k \phi(t + 4k + 1) = u_k(t)$, for each k .

Let h be any C^∞ , strictly monotone (with $h'(x) > 0$ for all x), sigmoidal function. There are many, e.g., $h(t) = 1/(1 + e^{-t})$. We define ϕ on $[4k, 4k + 2]$ in the following way. Set $\phi(t + 4k + 1) = b_k + c_k t + d_k u_k(t)$ for $t \in [-1, 1]$ where we choose the constants b_k , c_k and d_k so that

a) $\phi(4k) = h(4k)$.

b) $0 < \phi'(t) \leq h'(t)$ on $[4k, 4k + 2]$.

This is easily done. We make one further assumption. On the intervals $[-4, -2]$ and $[-8, -6]$ we demand that ϕ again satisfy conditions (a) and (b), as above, and be linear, and that $\phi(t - 3)$ and $\phi(t - 7)$ be linearly independent on $[-1, 1]$. To finish the construction, simply fill in the gaps in the domain of definition of ϕ (including all of $(-\infty, -8)$) in such a way that $\lim_{t \rightarrow -\infty} \phi(t) = 0$. From the construction there exists, for each $k \geq 1$, reals a_1^k, a_2^k, a_3^k , for which

$$a_1^k \phi(t - 7) + a_2^k \phi(t - 3) + a_3^k \phi(t + 4k + 1) = u_k(t),$$

for all $t \in [-1, 1]$. With this construction we now complete the proof of the proposition.

Let $f \in M_n$. Thus

$$f(\mathbf{x}) = \sum_{j=1}^n g_j(\mathbf{a}^j \cdot \mathbf{x})$$

for some $g_j \in C[-1, 1]$ and $\mathbf{a}^j \in S^{d-1}$, $j = 1, \dots, n$. From the above construction of ϕ there exist constants b_1^j, b_2^j, b_3^j and an integer r_j such that

$$|g_j(t) - (b_1^j \phi(t - 7) + b_2^j \phi(t - 3) + b_3^j \phi(t + r_j))| < \varepsilon/n$$

for all $t \in [-1, 1]$ and $j = 1, \dots, n$.

Thus

$$|g_j(\mathbf{a}^j \cdot \mathbf{x}) - (b_1^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 7) + b_2^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 3) + b_3^j \phi(\mathbf{a}^j \cdot \mathbf{x} + r_j))| < \varepsilon/n$$

for all $\mathbf{x} \in B^d$, and hence

$$\left| f(\mathbf{x}) - \sum_{j=1}^n \left(b_1^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 7) + b_2^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 3) + b_3^j \phi(\mathbf{a}^j \cdot \mathbf{x} + r_j) \right) \right| < \varepsilon$$

for all $\mathbf{x} \in B^d$. Now each $\phi(\mathbf{a}^j \cdot \mathbf{x} - 7)$, $\phi(\mathbf{a}^j \cdot \mathbf{x} - 3)$, $j = 1, \dots, n$, is a linear function, i.e., a linear combination of $1, x_1, \dots, x_d$. As such the

$$\sum_{j=1}^n b_1^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 7) + b_2^j \phi(\mathbf{a}^j \cdot \mathbf{x} - 3)$$

may be rewritten using at most $d + 1$ terms from the sum. This proves the proposition. ■

We now generalize the above construction to show how we may replace the C^∞ function by an analytic function. We will prove:

Theorem 2. *There exists a function σ which is real analytic, strictly increasing, and sigmoidal satisfying the following. Given $f \in M_n$ and $\varepsilon > 0$, there exist constants c_i , integers r_i and vectors $\mathbf{w}^i \in S^{d-1}$, $i = 1, \dots, 3n$, such that*

$$\left| f(\mathbf{x}) - \sum_{i=1}^{3n} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - r_i) \right| < \varepsilon$$

for all $\mathbf{x} \in B^d$.

Proof: Let ϕ be as above. We first modify ϕ in two ways. We choose the sequence $\{u_k\}_{k=1}^\infty$ such that in addition to the above it have the further property that to each $g \in C[-1, 1]$ there exists a sequence of positive integers $\{n_k\}_{k=1}^\infty$ satisfying $\lim_{k \rightarrow \infty} n_k = \infty$ for which

$$\lim_{k \rightarrow \infty} \|g - u_{n_k}\|_{C[-1,1]} = 0.$$

(If the original sequence $\{u_k\}_{k=1}^\infty$ does not have this property then we can obtain such a sequence by a diagonalization process.) We also assume that to each k there exist reals $a_{1,k}$, $a_{2,k}$ and $a_{3,k}$ such that

$$a_{1,k} \phi(t - 8k + 1) + a_{2,k} \phi(t - 8k + 5) + a_{3,k} \phi(t + 4k + 1) = u_k(t)$$

for $t \in [-1, 1]$. This is a simple modification of the previous construction where we let ϕ be linear and linearly independent on each of $[-8k, -8k + 2]$ and $[-8k + 4, -8k + 6]$.

Thus to each $g \in C[-1, 1]$ there exists a sequence of integers $\{n_k\}$ and sequences of real numbers $(a_{1,n_k}, a_{2,n_k}, a_{3,n_k})$ which depend on n_k (but not on g) such that

$$\lim_{k \rightarrow \infty} \|g - a_{1,n_k} \phi(\cdot - 8n_k + 1) - a_{2,n_k} \phi(\cdot - 8n_k + 5) - a_{3,n_k} \phi(\cdot + 4n_k + 1)\|_{C[-1,1]} = 0.$$

Let $\mathcal{A}_k = [-8k, -8k + 2] \cup [-8k + 4, -8k + 6] \cup [4k, 4k + 2]$, $k = 1, 2, \dots$. Let $h \in C(\mathbb{R})$ be any positive function satisfying

- $\lim_{|x| \rightarrow \infty} h(x) = 0$,
- $h(x) \leq \phi'(x)/2$ for all $x \in \mathbb{R}$,
- $h(x) \leq [k(|a_{1,k}| + |a_{2,k}| + |a_{3,k}|)]^{-1}$ for all $x \in \mathcal{A}_k$.

We now apply an approximation theorem due to Whitney [22]. This result may be found in Narasimhan [15, p. 34]. The particular case we will use states that given ϕ and h as above there exists a real analytic σ for which

$$|\phi(x) - \sigma(x)| < h(x), \quad x \in \mathbb{R} \quad (2.1)$$

and

$$|\phi'(x) - \sigma'(x)| < h(x), \quad x \in \mathbb{R}. \quad (2.2)$$

(This result is also contained in a generalization of Carleman's Theorem (Carleman [2]) due to Kaplan [6].)

Now, from property (a) and the fact that $\lim_{x \rightarrow -\infty} \phi(x) = 0$ and $\lim_{x \rightarrow \infty} \phi(x) = 1$ it follows that $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ and $\lim_{x \rightarrow \infty} \sigma(x) = 1$. From property (b) and (2.2) it easily follows that

$$\frac{\phi'(x)}{2} \leq \sigma'(x), \quad x \in \mathbb{R}.$$

As $\phi' > 0$, we have that σ is strictly increasing. Thus σ is analytic, strictly increasing and sigmoidal.

Finally, given $g \in C[-1, 1]$ let the sequence of integers $\{n_k\}$ and sequence of real numbers $(a_{1,n_k}, a_{2,n_k}, a_{3,n_k})$ satisfy

$$\lim_{k \rightarrow \infty} \|g - a_{1,n_k} \phi(\cdot - 8n_k + 1) - a_{2,n_k} \phi(\cdot - 8n_k + 5) - a_{3,n_k} \phi(\cdot + 4n_k + 1)\|_{C[-1,1]} = 0.$$

Then

$$\begin{aligned} & \|g - a_{1,n_k} \sigma(\cdot - 8n_k + 1) - a_{2,n_k} \sigma(\cdot - 8n_k + 5) - a_{3,n_k} \sigma(\cdot + 4n_k + 1)\|_{C[-1,1]} \\ & \leq \|g - a_{1,n_k} \phi(\cdot - 8n_k + 1) - a_{2,n_k} \phi(\cdot - 8n_k + 5) - a_{3,n_k} \phi(\cdot + 4n_k + 1)\|_{C[-1,1]} \\ & + \|a_{1,n_k} (\phi - \sigma)(\cdot - 8n_k + 1) + a_{2,n_k} (\phi - \sigma)(\cdot - 8n_k + 5) + a_{3,n_k} (\phi - \sigma)(\cdot + 4n_k + 1)\|_{C[-1,1]} \\ & \leq \|g - a_{1,n_k} \phi(\cdot - 8n_k + 1) - a_{2,n_k} \phi(\cdot - 8n_k + 5) - a_{3,n_k} \phi(\cdot + 4n_k + 1)\|_{C[-1,1]} \\ & \quad + \left(\sum_{i=1}^3 |a_{i,n_k}| \right) \|\phi - \sigma\|_{C(\mathcal{A}_{n_k})}. \end{aligned}$$

From property (c) and (2.1) we obtain

$$\begin{aligned} & \|g - a_{1,n_k} \sigma(\cdot - 8n_k + 1) - a_{2,n_k} \sigma(\cdot - 8n_k + 5) - a_{3,n_k} \sigma(\cdot + 4n_k + 1)\|_{C[-1,1]} \\ & \leq \|g - a_{1,n_k} \phi(\cdot - 8n_k + 1) - a_{2,n_k} \phi(\cdot - 8n_k + 5) - a_{3,n_k} \phi(\cdot + 4n_k + 1)\|_{C[-1,1]} + \frac{1}{n_k}. \end{aligned}$$

Thus

$$\lim_{k \rightarrow \infty} \|g - a_{1,n_k} \sigma(\cdot - 8n_k + 1) - a_{2,n_k} \sigma(\cdot - 8n_k + 5) - a_{3,n_k} \sigma(\cdot + 4n_k + 1)\|_{C[-1,1]} = 0.$$

This, together with the latter section of the proof of Proposition 1 proves Theorem 2. \blacksquare

Remark. If we are only interested in an analytic function whose (integer) shifts are dense in $C[-1, 1]$, without the above demands of sigmoidality and strict monotonicity, then we could also appeal directly to another result, namely Birkhoff's Theorem [1]. An explicit function which has this property is given by the Riemann-zeta function restricted to the line $\operatorname{Re}(z) = 3/4$ (this follows from work of Voronin [21]).

We have thus proven the following result which holds in many different normed linear spaces, but which for simplicity we state only for $C(B^d)$.

Theorem 3. *For every $f \in C(B^d)$, and any function $\phi \in C(\mathbb{R})$*

$$\inf_{g \in M_n} \|f - g\|_{C(B^d)} \leq \inf_{c_i, \theta_i, \mathbf{w}^i} \|f - \sum_{i=1}^n c_i \phi(\mathbf{w}^i \cdot -\theta_i)\|_{C(B^d)}.$$

Moreover there exists a function σ which is real analytic, strictly increasing, and sigmoidal for which

$$\inf_{g \in M_n} \|f - g\|_{C(B^d)} = \inf_{c_i, \theta_i, \mathbf{w}^i} \|f - \sum_{i=1}^{3n} c_i \sigma(\mathbf{w}^i \cdot -\theta_i)\|_{C(B^d)},$$

for every $f \in C(B^d)$.

The restriction of the above results to the unit ball B^d is purely for convenience. The exact same results also hold on any compact subset of \mathbb{R}^d .

§3. Two Hidden Layers

In this section we prove the following result:

Theorem 4. *There exists an activation function σ which is real analytic, strictly increasing, and sigmoidal, and has the following property. For any $f \in C[0, 1]^d$ and $\varepsilon > 0$, there exist real constants $d_i, c_{ij}, \theta_{ij}, \gamma_i$, and vectors $\mathbf{w}^{ij} \in \mathbb{R}^d$ for which*

$$\left| f(\mathbf{x}) - \sum_{i=1}^{6d+3} d_i \sigma \left(\sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \right) \right| < \varepsilon,$$

for all $\mathbf{x} \in [0, 1]^d$.

In the proof of Theorem 4 we use the Kolmogorov Superposition Theorem. This theorem has been much quoted and discussed in the neural network literature, see Hecht-Nielsen [5], Girosi, Poggio [4], Kurkova [8], [9], [10], Lin, Unbehauen [11]. In fact Kurkova [9] uses the Kolmogorov Superposition Theorem to construct approximations in the two hidden layer MLP model with an arbitrary sigmoidal function, where the number of units needed is dependent on the smoothness properties of the function approximated and the desired error of approximation (this number grows to infinity as the error decreases). Kurkova [9] and others (see Frisch, Borzi, Ord, Percus, Williams [3], Sprecher [20], Katsuura, Sprecher [7], Nees [16], [17]) are interested in using the Kolmogorov Superposition Theorem to find good algorithms for approximation. This is not our aim.

We will show, using the activation function of Section 2, that a finite number of units in both hidden layers is sufficient to approximate arbitrarily well any continuous function.

The Kolmogorov Superposition Theorem answers (in the negative) Hilbert's 13th problem. It was proven by Kolmogorov in a series of papers in the late 1950's. We quote below an improved version of this theorem as this is, for us, a slightly more convenient form of this theorem (see Lorentz, v. Golitschek, Makovoz [12, p. 553] for a more detailed discussion).

Theorem 5. *There exist d constants $\lambda_j > 0$, $j = 1, \dots, d$, $\sum_{j=1}^d \lambda_j \leq 1$, and $2d + 1$ continuous strictly increasing functions φ_i , $i = 1, \dots, 2d + 1$, which map $[0, 1]$ to itself, such that every continuous function f of d variables on $[0, 1]^d$ can be represented in the form*

$$f(x_1, \dots, x_d) = \sum_{i=1}^{2d+1} g\left(\sum_{j=1}^d \lambda_j \varphi_i(x_j)\right) \quad (3.1)$$

for some $g \in C[0, 1]$ depending on f .

Note that this is a theorem about representing (and not approximating) functions. There have been numerous generalizations of this theorem in various directions. Attempts to understand the nature of this theorem have also led to interesting concepts related to the complexity of functions. Nonetheless the theorem itself has had few, if any, direct applications.

Proof of Theorem 4. We are given $f \in C[0, 1]^d$ and $\varepsilon > 0$. Let g and the φ_i be as in (3.1). We will use the σ constructed in Section 2. Recall that to any $h \in C[-1, 1]$ and $\eta > 0$ we can find constants a_1, a_2, a_3 and integers m_1, m_2, m_3 for which

$$|h(t) - (a_1\sigma(t + m_1) + a_2\sigma(t + m_2) + a_3\sigma(t + m_3))| < \eta$$

for all $t \in [-1, 1]$. This result is certainly valid when we restrict ourselves to the interval $[0, 1]$ and functions continuous thereon. As such for the above g there exist constants a_1, a_2, a_3 and integers m_1, m_2, m_3 such that

$$|g(t) - (a_1\sigma(t + m_1) + a_2\sigma(t + m_2) + a_3\sigma(t + m_3))| < \frac{\varepsilon}{2(2d + 1)} \quad (3.2)$$

for all $t \in [0, 1]$.

Substituting (3.2) in (3.1) we obtain

$$\begin{aligned} \left| f(x_1, \dots, x_d) - \sum_{i=1}^{2d+1} \left[a_1\sigma\left(\sum_{j=1}^d \lambda_j \varphi_i(x_j) + m_1\right) + a_2\sigma\left(\sum_{j=1}^d \lambda_j \varphi_i(x_j) + m_2\right) \right. \right. \\ \left. \left. + a_3\sigma\left(\sum_{j=1}^d \lambda_j \varphi_i(x_j) + m_3\right) \right] \right| < \frac{\varepsilon}{2}. \end{aligned} \quad (3.3)$$

for all $(x_1, \dots, x_d) \in [0, 1]^d$.

We may rewrite (3.3) as

$$\left| f(x_1, \dots, x_d) - \sum_{i=1}^{6d+3} d_i \sigma \left(\sum_{j=1}^d \lambda_j \varphi_i(x_j) + \gamma_i \right) \right| < \frac{\varepsilon}{2} \quad (3.4)$$

for all $(x_1, \dots, x_d) \in [0, 1]^d$, where φ_i is, for each $i \geq 2d + 2$, equal to one of the φ_k for some $k \in \{1, \dots, 2d + 1\}$.

For each $i \in \{1, \dots, 6d + 3\}$ and $\delta > 0$ there exist constants b_{i1}, b_{i2}, b_{i3} and integers r_{i1}, r_{i2}, r_{i3} such that

$$|\varphi_i(x_j) - (b_{i1}\sigma(x_j + r_{i1}) + b_{i2}\sigma(x_j + r_{i2}) + b_{i3}\sigma(x_j + r_{i3}))| < \delta$$

for all $x_j \in [0, 1]$. Thus, since $\lambda_j > 0$, $\sum_{j=1}^d \lambda_j \leq 1$,

$$\left| \sum_{j=1}^d \lambda_j \varphi_i(x_j) - \sum_{j=1}^d \lambda_j (b_{i1}\sigma(x_j + r_{i1}) + b_{i2}\sigma(x_j + r_{i2}) + b_{i3}\sigma(x_j + r_{i3})) \right| < \delta$$

for all $(x_1, \dots, x_d) \in [0, 1]^d$.

This we can rewrite as

$$\left| \sum_{j=1}^d \lambda_j \varphi_i(x_j) - \sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) \right| < \delta \quad (3.5)$$

for all $(x_1, \dots, x_d) \in [0, 1]^d$, for some constants c_{ij} and θ_{ij} and vectors \mathbf{w}^{ij} (in fact the \mathbf{w}^{ij} are all unit vectors). As σ is uniformly continuous on every closed interval, we can choose $\delta > 0$ sufficiently small so that

$$\left| \sum_{i=1}^{6d+3} d_i \sigma \left(\sum_{j=1}^d \lambda_j \varphi_i(x_j) + \gamma_i \right) - \sum_{i=1}^{6d+3} d_i \sigma \left(\sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \right) \right| < \frac{\varepsilon}{2}. \quad (3.6)$$

From (3.4), (3.6), renumbering and renaming, the theorem follows. ■

Remark. If we are willing to replace the demand of analyticity of σ by only C^∞ (as in Proposition 1), then we can make do in Theorem 4 with $2d + 1$ units in the first layer and $4d + 3$ units in the second layer. If we forego the demand of strict monotonicity and sigmoidality then we can make do in Theorem 4 with d units in the first layer and $2d + 1$ units in the second layer. The restriction of Theorem 4 to the unit cube is for convenience only. The same result holds over any compact subset of \mathbb{R}^d .

Acknowledgements. The authors wish to express their appreciation to the reviewers and to Wolfgang Luh, Ronny Meir and Vilmos Totik for their help and suggestions. The research of V. M. was supported by The Center for Absorption in Science, Ministry of Immigrant Absorption, State of Israel. The research of A. P. was supported by the Fund for the Promotion of Research at the Technion.

References

1. G. D. Birkhoff, “Démonstration d’une théorème élémentaire sur les fonctions entières”, *C. R. Acad. Sci. Paris*, **189** (1929), 473–475.
2. T. Carleman, “Sur une théorème de Weierstrass”, *Ark. Mat. Astronom. Fys.* **20B** (1927), 1–5.
3. H. L. Frisch, C. Borzi, D. Ord, J. K. Percus and G. O. Williams, “Approximate representation of functions of several variables in terms of functions of one variable”, *Phys. Review Letters* **63** (1989), 927–929.
4. F. Girosi and T. Poggio, “Representation properties of networks: Kolmogorov’s theorem is irrelevant”, *Neural Computation* **1** (1989), 465–469.
5. R. Hecht-Nielsen, “Kolmogorov’s mapping neural network existence theorem” in *Proceedings of the International Conference on Neural Networks*, (M. Caudill and C. Butler, eds), IEEE (San Diego), III, 1987, 11–14.
6. W. Kaplan, “Approximation by entire functions”, *Michigan Math. j.* **3** (1955/56), 43–52.
7. H. Katsuura and D. A. Sprecher, “Computational aspects of Kolmogorov’s superposition theorem”, *Neural Networks* **7** (1994), 455–461.
8. V. Kurkova, “Kolmogorov’s theorem is relevant”, *Neural Computation* **3** (1991), 617–622.
9. V. Kurkova, “Kolmogorov’s theorem and multilayer neural networks”, *Neural Networks* **5** (1992), 501–506.
10. V. Kurkova, “Kolmogorov’s theorem”, preprint.
11. J. N. Lin and R. Unbehauen, “On realization of a Kolmogorov network”, *Neural Computation* **5** (1993), 18–20.
12. G. G. Lorentz, M. v. Golitschek and Y. Makovoz, *Constructive Approximation, Advanced Problems*, Springer Verlag, New York, 1996.
13. V. E. Maiorov, “On best approximation by ridge functions”, to appear in *J. Approx. Theory*.
14. V. E. Maiorov, R. Meir and J. Ratsaby, “On the approximation of functional classes equipped with a uniform measure using ridge functions”, to appear in *J. Approx. Theory*.
15. R. Narasimhan, *Analysis on Real and Complex Manifolds*, North-Holland, Amsterdam, 1968.
16. M. Nees, “Approximative versions of Kolmogorov’s superposition theorem, proved constructively”, *J. Comput. Appl. Anal.* **54** (1994), 239–250.
17. M. Nees, “Chebyshev approximation by discrete superposition. Application to neural networks”, *Adv. Comp. Math.* **5** (1996), 137–151.
18. K. I. Oskolkov, “Ridge approximation, Chebyshev-Fourier analysis and optimal quadrature formulas”, preprint.
19. A. Pinkus, “Approximating by Ridge Functions”, in *Surface Fitting and Multiresolution Methods*, (A. Le Méhauté, C. Rabut and L. L. Schumaker, eds), Vanderbilt Univ. Press (Nashville), 1997, 279–292.
20. D. A. Sprecher, “A universal mapping for Kolmogorov’s superposition theorem”, *Neural Networks* **6** (1993), 1089–1094.

21. S.M. Voronin, “A theorem on the “universality” of the Riemann-zeta function”, *Izv. Akad. Nauk SSSR* **39** (1975), 475–486.
22. H. Whitney, “Analytic extensions of differentiable functions defined in closed sets”, *Trans. Amer. Math. Soc.* **36** (1934), 63–89.

Vitaly Maiorov
Department of Mathematics
Technion, I. I. T.
Haifa, 32000
Israel
maiorov@tx.technion.ac.il

Allan Pinkus
Department of Mathematics
Technion, I. I. T.
Haifa, 32000
Israel
pinkus@tx.technion.ac.il