Neural Networks, Vol. 5, pp. 501–506, 1992
Printed in the USA.

0893-6080/92 $5.00 + .00

ORIGINAL CONTRIBUTION

Kolmogorov's Theorem and Multilayer Neural Networks

VĚRA KŮRKOVÁ

Czechoslovak Academy of Sciences

(*Received* 1 *February* 1991; *revised and accepted* 20 *September* 1991)

Abstract—*Taking advantage of techniques developed by Kolmogorov, we give a direct proof of the universal approximation capabilities of perceptron type networks with two hidden layers. From our proof, we derive estimates of numbers of hidden units based on properties of the function being approximated and the accuracy of its approximation.*

Keywords—Feedforward neural networks, Multilayer perceptron type networks, Sigmoidal activation function, Approximations of continuous functions, Uniform approximation, Universal approximation capabilities, Estimates of number of hidden units, Modulus of continuity.

1. INTRODUCTION

The ability to, quite well, approximate various functions encountered in applications of sufficiently elaborate hierarchies of perceptrons deserves a theoretical justification. Recently, universal approximation capabilities of three-layered perceptron type networks, with more or less general types of activation functions, were confirmed by several authors (Carroll & Dickinson, 1989, Cybenko, 1989, Funahashi, 1989, Hecht-Nielsen, 1989, Hornik, 1991, Hornik, Stinchcombe, & White, 1989, Stinchcombe and White, 1989), who have taken elegant advantage of various advanced theorems from functional analysis. However, all of these theorems have focused only on existence of an approximation, having supposed that the number of hidden units is not bounded. Important questions that remain to be answered deal with feasibility: what properties of the function being implemented play a role in determining the number of hidden units, and how quickly does this number grow with increasing accuracy?

Several years ago, before the above mentioned results were known, Hecht-Nielsen (1987, 1990) called attention to the significance and potential applicability to neurocomputing of Kolmogorov's remarkable theorem concerning representation of continuous functions defined on an n-dimensional cube by sums and superpositions of continuous functions of one variable. However, as was pointed out by Girosi and Poggio (1989), the one-variable functions constructed by Kolmogorov (1957) as well as their later improvements by Lorentz (1966) and Sprecher (1965), are far from being any of the type of functions currently used in neurocomputing. In the present paper, we show that by sacrificing exactness of a representation, we can eliminate this difficulty. We give an approximation version of Kolmogorov's theorem, where all one-variable functions are finite linear combinations of affine transformations with an arbitrary sigmoidal function (i.e., a function $\sigma:\mathbb{R} \to [0,1]$ with $\lim_{t \to -\infty} \sigma(t) = 0$ and $\lim_{t \to \infty} \sigma(t) = 1$). We derive that any continuous function defined on an n-dimensional cube can be implemented by means of a perceptron type network with two hidden layers with any sigmoidal activation function.

Although above mentioned approximation theorems require only one hidden layer, an advantage of our approach is in the directness of our argument, requiring no use of advanced theorems from functional analysis, and so enabling to estimate numbers of hidden units as functions of the accuracy desired and the rate of increase of the function being approximated. Moreover, our construction provides a universal set of weights and biases for a network capable of approximating all functions from a certain set (characterized by a bounded norm and modulus of continuity) with a given accuracy so that only weights corresponding to the output units should be learned. This simplifies considerably a prob-

Acknowledgment: The author is grateful for helpful conversation with Professors Robert Hecht-Nielsen and Halbert White from the University of California, San Diego, CA.

Requests for reprints should be sent to Věra Kůrková, Institute of Computer Science, Czechoslovak Academy of Sciences, Pod vodárenskou věží 2, P.O. Box 5, 182 07 Prague 8, Czechoslovakia.

lem of learning by transforming it to a problem of linear regression.

The organization of our paper is as follows: In Section 2, we deal with preliminaries, in Section 3, we explain the context of the problem and present our main results, while all proofs are postponed to Section 4.

2. PRELIMINARIES

By a **multilayer perceptron-type network** we mean a multilayer network where units in each hidden layer sum up weighted inputs from the preceding layer, add to this sum a constant (bias) and then apply a common activation function, while units in the output layer only sum weighted inputs. Since a multioutput network can be composed of one-output networks, we shall restrict ourselves only to one-output networks. Since in application, values of possible input vectors are bounded, we shall suppose that they are within a unit n-dimensional cube $I^n (I = [0,1])$. In this paper we shall consider only perceptron type networks with sigmoidal activation functions (i.e., functions $\sigma:\mathbb{R}\to I$ with $\lim_{t\to-\infty}\sigma(t) = 0$ and $\lim_{t\to\infty}\sigma(t) = 1$ (where $\mathbb{R}$ denotes the set of real numbers). So functions used in perceptron type networks are finite linear combinations of compositions of affine transformations of $\mathbb{R}$ with some sigmoidal function σ (i.e., functions of the form $\sum_{i=1}^{k} a_i\sigma(b_i x + c_i)$). We call them **staircase-like functions of a type** σ and denote set of all such functions $S(\sigma)$.

In the context of neural networks, we are interested only in approximate realizations of functions. Questions concerning only the existence of approximation can be formulated in topological terms, such as the closure of a set and a dense subset. We recall their definitions: A **closure** of a subset D of a topological space S, usually denoted by $\bar{D}$, is the set of points in S with the property that every neighborhood of such a point has a nonempty intersection with D. A subset D of a topological space S is called **dense** if $\bar{D} = S$. By $C(S)$ we denote the set of all continuous real-valued functions on S.

Topologies considered in this context are naturally defined by metrics, while several metrics can induce the same topology. The most often used metrics are the standard metrics from functional analysis defined by the supremum and the L^p norms. The supremum norm, defined by $\|f\| = \sup\{|f(x)|, x \in X\}$ with induced supremum metrics $\|f - g\| = \sup\{|f(x) - g(x)|, x \in X\}$ and a derived topology called the topology of uniform convergence, is suitable for applications demanding the same quality of performance for all input vectors from a set X. If some input environment measure μ, expressing the importance or likelihood of various input vectors, can be specified, then it is more convenient to use L^p norms ($p \geq 1$), defined for a measure μ on a set X of input vectors on the set $L^p_\mu(X)$ of all the real functions f on X for which the Lebesgue integral $\int |f|^p d\mu$ is finite by $\|f\|_{p\mu} = (\int |f|^p d\mu)^{1/p}$ with the induced psedometrics $\rho_{p\mu}(f, g) = (\int |f - g|^p d\mu)^{1/p}$.

In this paper, we shall formulate our results for simplicity only for the supremum norm. However, for reasonable measures on I^n the space $C(I^n)$ is dense in $L^p_\mu(I^n)$ with the topology induced by $\rho_{p\mu}$ for every $p \geq 1$, and moreover the topology of uniform convergence on the space $L^p_\mu(I^n)$ is finer (contains more open sets) than the topology induced by $\rho_{p\mu}$. So approximation capabilities with respect to supremum norm guarantee approximation capabilities also with respect to all reasonable input environment measures.

A function $\omega_f:(0,\infty) \to \mathbb{R}$ is called a **modulus of continuity** of a function $f: I^n \to \mathbb{R}$ if

$$\omega_f(\delta) = \sup\{|f(x_1, \cdots x_n) - f(y_1, \cdots y_n)|, (x_1, \cdots x_n), (y_1, \cdots y_n) \in I^n \text{ with } |x_p - y_p| < \delta \text{ for every } p = 1, \cdots n\}.$$

We call real numbers $w_1, \cdots w_n$ **integer independent** if $\sum_{i=1}^{k} w_i z_i \neq 0$ for any integers $z_1, \cdots z_k$.

For a positive real ϵ we call a set $X \subseteq \mathbb{R}$ **ϵ-distinguishable** if the distance between any two of its points exceeds ϵ.

By $\mathbb{N}$ we denote the set of all natural numbers.

3. MAIN RESULTS

First, let us briefly recall the history of Kolmogorov's representation theorem. Hilbert, in his famous lecture "Mathematische Probleme" at the 2nd International Congress of Mathematics held in Paris in 1900, gave a list of 23 open problems, solutions of which he supposed to be the most important for further development of mathematics. The 13th problem, although formulated as a concrete minor hypothesis, concerned solutions of polynomial equations (Kůrková, 1991). Could roots of a general algebraic equation of higher degree be expressed, analogously to the solution by radicals, by sums and compositions of one-variable functions of some suitable type? Hilbert conjectured that the roots of the equation $x^7 + ax^3 + bx^2 + cx + 1 = 0$ as functions of the three coefficients a, b, c are not representable by sums and superpositions even of functions of two variables. This was disproved by Arnold (1957). Kolmogorov (1957) even proved a general representation theorem stating that any real-valued continuous function f defined on an n-dimensional cube I^n ($n \geq 2$) can be represented as

$$f(x_1, \cdots x_n) = \sum_{q=1}^{2n+1} \phi_q\left(\sum_{p=1}^{n} \psi_{pq}(x_p)\right),$$

where ϕ_q ($q = 1, \cdots 2n + 1$) and ψ_{pq} ($p = 1, \cdots n$, $q = 1, \cdots 2n + 1$) are continuous functions of one variable. Moreover, the functions ψ_{pq} are universal for the given dimension n; they are independent of f. Only the functions ϕ_q are specific for the given function f. Several authors have improved on Kolmogorov's representation. Lorentz (1966) showed that the functions ϕ_q can be replaced by only one function ϕ and Sprecher (1965) replaced the functions ψ_{pq} by $\lambda^{pq}\psi_q$, where λ is a constant and ψ_q are monotonic increasing functions belonging to the class $\mathrm{Lip}[\ln 2/\ln(2n + 2)]$.

Hecht-Nielsen (1987) reformulated Sprecher's version of the representation theorem in the language of neural networks as follows: Any continuous function defined on an n-dimensional cube can be implemented exactly by a three-layered network having $2n + 1$ units in the hidden layer with transfer functions $\lambda^{pq}\psi_q$ ($p = 1, \cdots n, q = 1, \cdots 2n + 1$) from the input to the hidden layer and ϕ from all of the hidden units to the output one. As Hecht-Nielsen pointed out, the universality of the transfer functions ψ_{pq} can be exploited for representations of functions with values in spaces of higher dimensions using Kolmogorov's theorem for compositions of the function f with all of the projections π_i: $\mathbb{R}^m \to \mathbb{R}$, $i = 1, \cdots m$.

However, possessing even fractal graphs, the functions ψ_q and ϕ are highly nonsmooth (this explains the failure of Hilbert's intuition—functions with fractal graphs had been supposed to be pathological then). Nevertheless, staircase-like functions of any sigmoidal type have a pleasant property, that they can approximate any continuous function on any closed interval with an arbitrary accuracy. Taking advantage of this fact, we can derive from Kolmogorov's representation theorem the following approximation theorem.

THEOREM 1. *Let $n \in \mathbb{N}$ with $n \geq 2$, $\sigma:\mathbb{R} \to I$ be a sigmoidal function, $f \in C(I^n)$, and ϵ be a positive real number. Then there exist $k \in \mathbb{N}$ and functions ϕ_i, $\psi_{pi} \in S(\sigma)$ such that*

$$|f(x_1, \cdots x_n) - \sum_{i=1}^{k} \phi_i\Big(\sum_{p=1}^{n} \psi_{pj}x_p\Big)| < \epsilon$$

$$\text{for every} \quad (x_1, \cdots x_n) \in I^n.$$

The same argument was used by Funahashi (1989) for increasing continuous sigmoidal functions. Of course, such nondirect arguments do not provide any estimates of numbers of hidden units. Being complicated and tricky, Kolmogorov's construction of the functions ϕ_q *anf* ψ_{pq} contains a lot of unnecessary assumptions. The only really relevant property of the functions used in the induction construction of the functions ϕ_q *anf* ψ_{pq} is that they have prescribed values on finitely many closed intervals, elsewhere they can be arbitrary, provided they are sufficiently bounded. However, such functions can be approximated arbitrarily well by staircase-like functions of any sigmoidal type. Moreover, if our goal is only approximation, we can even considerably simplify the induction construction in such a way that the staircase-like functions involved have much less steps than in the case of the Kolmogorov's original construction.

THEOREM 2. *Let $n \in \mathbb{N}$ with $n \geq 2$, $\sigma:\mathbb{R} \to I$ be a sigmoidal function, $f \in C(I^n)$ and ϵ a positive real number. Then for every $m \in \mathbb{N}$ such that $m \geq 2n + 1$ and $n/(m - n) + v < \epsilon/\|f\|$ and $\omega_f(1/m) < v(m - n)/(2m - 3n)$ for some positive real v, f can be approximated with an accuracy ϵ by a perceptron type network with two hidden layers, containing $nm(m + 1)$ units in the first hidden layer and $m^2(m + 1)^n$ units in the second one, with an activation function σ in such a way that all weights and biases, with the exception of weights corresponding to the transfer from the second hidden layer to the output unit, are universal for all functions g with $\|g\| \leq \|f\|$ and $\omega_g \leq \omega_f$.*

It is not surprising that upper estimates of number of hidden units needed for good approximations of general continuous functions are very large. Perhaps, for special types of functions, better estimates could be obtained. The above theorem guarantees possibility of constructing perceptron type networks with two hidden layers with universal set of weights for approximations of functions within a certain class so that only weights corresponding to transfer from the second hidden layer to the output unit are specific for the function being approximated. Since these specific weights appear linearly in the parametrized expression, the problem of learning is in such networks transformed to the problem of a linear regression.

4. MATHEMATICAL PROOFS

LEMMA 1. *Let $\sigma:\mathbb{R} \to I$ be a sigmoidal function and $[a, b] \subset \mathbb{R}$ be a closed interval. Then the set of all functions $f:[a, b] \to \mathbb{R}$ of the form $f(x) = \sum_{i=1}^{k} w_i\sigma(v_i x + u_i)$, where w_i, v_i, u_i ($i = 1, \cdots k$) are any real numbers, is dense in $C([a, b])$ with the topology of uniform convergence.*

Proof of Theorem 1. By Kolmogorov's theorem

$$f(x_1, \cdots x_n) = \sum_{q=1}^{2n+1} \phi_q\Big(\sum_{p=1}^{n} \psi_{pq}(x_p)\Big).$$

Take $[a, b] \subset \mathbb{R}$ such that for every $p = 1, \cdots n$, $q = 1, \cdots 2n + 1$ $\psi_{pq}(I^n) \subseteq [a, b]$. By Lemma 1 for every $q = 1, \cdots 2n + 1$ there exists $g_q \in S(\sigma)$ such that $|g_q(x) - \phi_q(x)| < \epsilon/(2n(2n + 1))$ for every $x \in [a, b]$. Since g_q are uniformly continuous, there exists δ such that $|g_q(x) - g_q(y)| < \epsilon/(2n(2n + 1))$ for every $x, y \in [a, b]$ with $|x - y| < \delta$. For every $p = 1,$

$\cdots n, q = 1, \cdots 2n+1$ there exists $h_{pq} \in S(\sigma)$ such that for every $x \in I$ $|h_{pq}(x) - \psi_{pq}(x)| < \delta$. Hence for every $(x_1, \cdots x_n) \in I^n$

$$|\sum_{q=1}^{2n+1} g_q(\sum_{p=1}^{n} h_{pq}(x_p)) - f(x_1, \cdots x_n)| < \epsilon.$$

The following lemma, however simple, is essential for our proof of Theorem 2. It states that any finite family of steps can be approximated arbitrarily well by a function belonging to $S(\sigma)$.

LEMMA 2. *Let $\sigma:\mathbb{R} \to I$ be a sigmoidal function, ϵ be a positive real number, $k \in \mathbb{N}$ and $x_1 < y_1 < x_2 < y_2 < \cdots < x_k < y_k$ be real numbers and $g:\{1, \cdots k\} \to \mathbb{R}$ be any mapping. Then there exists $\phi \in S(\sigma)$ of the form $\phi(x) = \sum_{i=1}^{k} a_i \sigma(b_i x + c_i)$ such that $|\phi(x) - g(j)| < \epsilon$ for every $x \in [x_j, y_j]$ and for every $j = 1, \cdots k$, and*

$$\|\phi\| \leq max\{|g(j)|, j = 1, \cdots k\} + \epsilon.$$

Proof. Choose some real number y_0 with $y_0 < x_1$ and set $g(0) = 0$. Denote $M = \max\{|g(x_j)|, j = 1, \cdots k\}$. Since σ is a sigmoidal function, there exists such a real number z that $0 < \sigma(x) < \epsilon/4Mk$ for every $x < z$ and $1-\epsilon/4Mk < \sigma(x) < 1$ for every $x > z$. For each $i = 1, \cdots k$, let $b_i x + c_i$ be the unique affine transformation of $\mathbb{R}$ mapping the interval $\langle y_{i-1}, x_i \rangle$ onto $\langle -z, z \rangle$, and let $a_i = g(i) - g(i-1)$. Then for every $x \in [x_j, y_j]$ and for every $j = 1, \cdots k$, we have

$$|\sum_{i=1}^{k} a_i \sigma(b_i x + c_i) - g(j)| \leq |\sum_{i=1}^{j} a_i \sigma(b_i x + c_i) - g(j)| + |\sum_{i=j+1}^{k} a_i \sigma(b_i x + c_i)|$$

$$\leq \sum_{i=1}^{j-1} |g(i)||\sigma(b_i x + c_i) - \sigma(b_{i+1} x + c_{i+1})| + |g(j)||\sigma(b_j x + c_j) - 1| + \sum_{i=j+1}^{k} |a_i||\sigma(b_i x + c_i)| \leq Mj\epsilon/2Mk + M(k-j)\epsilon/2Mk \leq \epsilon.$$

Consider staircase-like functions $\psi_p \in S(\sigma)$ $(p = 1, \cdots n)$ and a derived function Ψ defined on I^n by $\Psi(x_1, \cdots x_n) = \sum_{p=1}^{n} \psi_p(x_p)$. The function Ψ divides the cube I^n into small boxes, resembling Rubik's cube, with the edges corresponding to the steps of ψ, and the gaps between them to the slopes of ψ. If we guarantee that Ψ-images of these boxes are contained within closed mutually disjoint intervals, we can use Lemma 1 to obtain a function $\phi \in S(\sigma)$ approximating values of any function $f \in C(I^n)$ in some chosen points of those boxes that are Ψ-preimages of these intervals. The smaller the steps of ψ_p, the smaller the little boxes and hence the better approximation of f. The only problem is with the points within the gaps. However, this can be overcome by taking sufficiently many staircases with slopes over different intervals. This is, roughly speaking, the main idea of the following proof.

Proof of Theorem 2. First, using Lemma 2 we shall construct m sequences $\{\chi_i^q, i \in \mathbb{N}\}$, $1, \cdots m$ of staircase-like functions belonging to $S(\sigma)$. For every $i \in \mathbb{N}$ and $q = 1, \cdots m$ define the family $\mathcal{A}_i^q$ of those subintervals of I on which the prescribed values will be approximated by functions χ_i^q by

$$\mathcal{A}_i^q = \{[(j-1)/m^i + q/m^{i+1}, j/m^i + (q-1)/m^{i+1}] \cap I, j = 0, \cdots m^i\}.$$

Define $g_i^q:\{0, \cdots m^i\} \to \mathbb{R}$ by $g_i^q(j) = j/m^i$. It remains to set accuracies v_i, within which will be values ascribed to intervals $A_{ij}^q \in \mathcal{A}_i^q$ by $g_i^q(j)$, approximated by functions from $S(\sigma)$. To do this, choose some integer independent numbers w_{pq}, $p = 1, \cdots n$, $q = 1, \cdots m$, and by means of them define mappings $\xi^q:I^n \to \mathbb{R}$ by

$$\xi^q(x_1, \cdots x_n) = \sum_{p=1}^{n} w_{pq} x_p.$$

For every $i \in \mathbb{N}$ put $D_i = \{j/m^i, j = 0, \cdots m^i\}$. Since for every $q = 1, \cdots m$ and for every $i \in \mathbb{N}$ $\xi^q(D_i)$ is finite, for every $i \in \mathbb{N}$ there exists a positive real number η_i with $\xi^q(D_i)$ being $2\eta_i$-distinguishable for every $q = 1, \cdots m$. Since all of the functions ξ^q are uniformly continuous, there exists such a positive real number v_i that whenever $(x_1, \cdots x_n), (y_1, \cdots y_n) \in I^n$ with $|x_p - y_p| < v_i$ for every $p = 1, \cdots n$, then $|\xi^q(x_1, \cdots x_n) - \xi^q(y_1, \cdots y_n)| < \eta_i$. By Lemma 2, for every $q = 1, \cdots m$ there exists a function $\chi_i^q \in S(\sigma)$ with $|\chi_i^q(x) - j/m^i| < v_i$ for every $x \in A_{ij}^q$ for all $j = 0, \cdots m^i$.

Construct m sequences of functions $\{\xi_i^q: I^n \to \mathbb{R}, i \in \mathbb{N}\}$ by setting $\xi_i^q(x_1, \cdots x_n) = \sum_{p=1}^{n} w_{pq} \chi_i^q(x_p)$ for every $(x_1, \cdots x_n) \in I^n$. Denote by $\mathcal{B}_i^q$ the family of all those n-dimensional boxes contained in I^n having all of their edges in $\mathcal{A}_i^q$. For a member of $\mathcal{B}_i^q$ call i its order and q its type. It is easy to verify that for every $i \in \mathbb{N}$, for every $q = 1, \cdots m$ and for every $B \in \mathcal{B}_i^q$, $B \cap D_i^n$ is a one-point set. Denote this point $\beta(B)$. Then

$$\xi_i^q(B) \subseteq [\xi_i^q(\beta(B)) - \eta_i, \xi_i^q(\beta(B)) + \eta_i].$$

Since numbers w_{pq}, $p = 1, \cdots n$, $q = 1, \cdots m$ are integer independent, images $\xi_i^q(B)$ and $\xi_i^q(B')$ of any two different members B, B' of $\mathcal{B}_i^q$ are disjoint. Moreover, even for boxes B and B' of different types q and q', but of the same order i, $\xi_i^q(B)$ and $\xi_i^{q'}(B')$ are disjoint with the only exception being those boxes B of any type q which have $\beta(B) = (0, \cdots 0)$. For each order $i \in \mathbb{N}$ and for each type $q = 2, \cdots m$, there is exactly one such box.

Let $f \in C(I^n)$. Since $m \geq 2n+1$, there exists $\delta > 0$ with $n/(m-n) + \delta(1 + n/2(n-m)) < 1$. Put

$$\alpha = n/(m-n) + \delta(1 + n/2(n-m)).$$

We shall construct by inducton using Lemma 2 a sequence of functions $\{\phi_i, i \in N\}$ belonging to $S(\sigma)$ and an increasing sequence of natural numbers $\{k_i, i \in N\}$ such that for every $i \in \mathbb{N}$

$$\|\phi_i\| \leq \alpha^{i-1}\|f\| \tag{1}$$

and

$$\|f - \sum_{q=1}^{m} \sum_{j=1}^{i} \phi_j \cdot \xi_{k_j}^q\| \leq \alpha^i \|f\|. \tag{2}$$

Put $\phi_0 = \text{const } 0$ and $k_0 = 0$. Suppose that for every $j < i$ ϕ_j and k_j are already defined. Put $h_i = f - \sum_{q=1}^{m} \sum_{j=1}^{i-1} \phi_j \cdot \xi_{k_j}^q$.

Since I^n is compact, h_i is uniformly continuous and hence there exists $k_i \in \mathbb{N}$ with $k_i > k_{i-1}$ such that the diameters of $\xi_{k_i}^q$-images of all boxes of the order k_j and of any type $q = 1, \cdots m$ are smaller than $\delta\|h_i\|/2$.

By Lemma 2 there exists such $\phi_i \in S(\sigma)$ that for every $q = 1, \cdots m$, for every $B \in \mathcal{B}_{k_i}^q$ and for every

$$x \in [\xi_{k_i}^q(\beta(B)) - \eta_{k_i}, \xi_{k_i}^q(\beta(B)) + \eta_{k_i}]$$

we have

$$|\phi_i(x) - h_i(\beta(B))/(m-n)| < \delta\|h_i\|/2(m-n)$$

and

$$\|\phi_i\| < \|h_i\|/(m-n) + \delta\|h_i\|/2(m-n).$$

So (1) is fulfilled, since according to our induction assumption $\|h_i\| \leq \alpha^{i-1}\|f\|$.

To verify (2) it is sufficient to show that

$$\|h_i - \sum_{q=1}^{m} \phi_i \cdot \xi_{k_i}^q\| \leq \alpha\|h_i\|,$$

since

$$f - \sum_{q=1}^{m} \sum_{j=1}^{i} \phi_j \cdot \xi_{k_j}^q = h_i - \sum_{q=1}^{m} \phi_i \cdot \xi_{k_i}^q$$

and our induction assumption guarantees that $\|h_i\| \leq \alpha^{i-1}\|f\|$.

For every $(x_1, \cdots x_n) \in I^n$ there exist at least $m-n$ different values of q for which there exists a box $B^q \in \mathcal{B}_{k_i}^q$ with $(x_1, \cdots x_n) \in B^q$ (since in the worst case there is for each component x_p of $(x_1, \cdots x_n)$ a different value q_p with x_p being contained in no interval from $\mathcal{A}_{k_i}^q$). Suppose that for $q = 1, \cdots m-n$ $(x_1, \cdots x_n \in B^q$ for some $B^q \in \mathcal{B}_{k_i}^q$, and so $|\phi_i \cdot \xi_{k_i}^q(x_1, \cdots x_n) - h_i(\beta(B^q))/(m-n)| < \delta\|h_i\|/2(m-n)$ and $|h_i(x_1, \cdots x_n) - h_i(\beta(B^q))| < \delta\|h_i\|/2$. Hence

$$|h_i(x_1, \cdots x_n) - \sum_{q=1}^{m-n} \phi_i \cdot \xi_{k_i}^q(x_1, \cdots x_n)|$$

$$= |\sum_{q=1}^{m-n} ((h_i(x_1, \cdots x_n) - h(\beta(B^q))/(m-n)|$$

$$\leq \delta\|h_i\|.$$

For $q = m-n+1, \cdots m$ we only know that

$$|\phi_i \cdot \xi_{k_i}^q(x_1, \cdots x_n)| \leq \|\phi_i\|$$

$$\leq \|h_i\|/(m-n) + \delta\|h_i\|/2(m-n).$$

So

$$|h(x_1, \cdots x_n) - \sum_{q=1}^{m} \phi_i \xi_{k_i}^q(x_1, \cdots x_n)|$$

$$\leq \|h_i - \sum_{i=1}^{m-n} \phi_i \cdot \xi_{k_i}^q\| + \|\sum_{q=m-n+1}^{m} \phi_i \cdot \xi_{k_i}^q\|$$

$$\leq (\delta + n/(m-n) + \delta n/2(m-n))\|h_i\| = \alpha\|h_i\|.$$

For a given $\epsilon > 0$ take $i \in \mathbb{N}$ with $\alpha^i\|f\| < \epsilon$. For every $j = 1, \cdots i$ put $\psi_{pqj} = w_{pq} \cdot \chi_{k_i}^q$. Since $\chi_{k_i}^q \in S(\sigma)$, $\psi_{pqj} \in S(\sigma)$, too. So we have

$$|f(x_1, \ldots x_n) - \sum_{q=1}^{m} \sum_{j=1}^{i} \phi_j(\sum_{p=1}^{n} \psi_{pqj}(x_p))| < \epsilon$$

for every $(x_1, \ldots x_n) \in I^n$.

After a suitable change of indexes, we obtain an approximation of f in the form stated in Theorem 1.

Analyzing constructions used in this proof, we shall derive upper estimates of the number of hidden units. Take functions χ_1^q and families of boxes $\mathcal{B}_1^q$, $q = 1, \cdots m$, defined above. Each of these functions has $m+1$ steps and hence χ_1^q is of the form $\sum_{i=1}^{m+1} a_{qi}\sigma(b_{qi}x + c_{qi})$. Like above construct a function ϕ_1 with prescribed values on intervals containing ξ_1^q-images of boxes of all families $\mathcal{B}_1^q$. Since there is m families, and each family contains $(m+1)^n$ boxes, ϕ_1 is of the form

$$\sum_{j=1}^{m(m+1)^n} d_j\sigma(v_j y + u_j).$$

Because of our assumptions on m, f is approximated within the desired accuracy ϵ by

$$\sum_{q=1}^{m} \phi_1(\sum_{p=1}^{n} w_{pq}\chi_1^q(x_p))$$

$$= \sum_{q=1}^{m} (\sum_{j=1}^{m(m+1)^n} (d_j\sigma(\sum_{p=1}^{n} \sum_{i=1}^{m+1} v_j w_{pq} a_{qi}\sigma(b_{qi}x_p + c_{qi}))$$

$$+ u_j)).$$

REFERENCES

Arnold, V. I. (1957). On functions of three variables. *Doklady Akademii Nauk USSR*, **114**, 679–681.

Carroll, S. M., & Dickinson, B. W. (1989). Construction of neural nets using the Radon transform. In *Proceedings of the International Joint Conference on Neural Networks* (I, 607–611). New York: IEEE Press.

Cybenko, G. (1989). Approximation by superpositions of a single function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183–192.

Girosi, F., & Poggio, T. (1989). Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, **1**, 465–469.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In *Proceedings of the International Conference on Neural Networks*, (III, 11–14). New York: IEEE Press.

Hecht-Nielsen, R. (1989). Theory of the back-propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*, (I, 593–608). New York: IEEE Press.

Hecht-Nielsen, R. (1990). *Neurocomputing*. New York: Addison-Wesley Publishing Company.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**(2), 251–257.

Kolmogorov, A. N. (1957). On the representations of continuous functions of many variables by superpositions of continuous functions of one variable and addition. *Doklady Akademii Nauk USSR*, **114**(5), 953–956.

Kůrková, V. (1991). 13th Hilbert's problem and neural networks. In *Theoretical Aspects of Neurocomputing* (213–216). Singapore: World Scientific.

Kůrková, V. (1991). Kolmogorov's Theorem is relevant. *Neural Computation* **3**, 617–622.

Lorentz, G. G. (1966). *Approximation of functions*. New York: Holt, Reinhart and Winston.

Sprecher, D. A. (1965). On the structure of continuous functions of several variables, *Transactions of the American Mathematical Society* **115**, 340–355.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks* (I, 613–617). New York: IEEE Press.

NOMENCLATURE

$\mathbb{R}$	real line
$\mathbb{R}^m$	m-dimensional euclidean space
I	interval [0,1]
I^n	n-dimensional unit cube
π_i	projection mapping $\mathbb{R}^m \to \mathbb{R}$
$\mathbb{N}$	natural numbers
$\lim_{t \to \infty} \sigma(t)$	limit of $\sigma(t)$
$S(\sigma)$	set of all functions of the form $\sum_{i=1}^{k} a_i \sigma(b_i x + c_i)$
$C(X)$	set of all continuous functions on a topological space X
$\bar{D}$	the closure of a set D
$\|f\|$	supremum norm
$\int f d\mu$	Lebesgue integral
$L^p_\mu(X)$	set of all real functions on X for which $\int \|f\|^p d\mu$ is finite
$\|f\|_{p\mu}$	L^p norm
ω_f	modulus of continuity of f
$\mathrm{Lip}[\alpha]$	class of functions