Neural Networks 176 (2024) 106333

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

On the Kolmogorov neural networks

Aysu Ismayilova [a], Vugar E. Ismailov [b,*]

[a] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
[b] Institute of Mathematics and Mechanics, Baku, Azerbaijan Center for Mathematics and its Applications, Khazar University, Baku, Azerbaijan

ARTICLE INFO

Keywords:
Kolmogorov's superposition theorem
Lipschitz function
Dual space
Conjugate operator
Indicator function
Linear functional

ABSTRACT

In this paper, we show that the Kolmogorov two hidden layer neural network model with a continuous, discontinuous bounded and unbounded activation function in the second hidden layer can precisely represent continuous, discontinuous bounded and all unbounded multivariate functions, respectively.

1. Introduction

There are a lot of papers in neural network literature on the capability of special neural networks, called the Kolmogorov neural networks or Kolmogorov's mapping neural networks, to precisely represent each continuous multivariate function. But precise representation in other function classes has not been considered there. In this paper, we show that the Kolmogorov networks have an extreme power of representing not only continuous, but also discontinuous bounded and all unbounded multivariate functions.

The idea of constructing of the above mentioned neural networks stems from the famous Kolmogorov superposition theorem. This theorem positively solves Hilbert's 13th problem. Hilbert in his address to the International Congress of Mathematicians held in Paris in 1900, outlined 23 outstanding mathematical problems, the 13th of which asked: Is the root of the equation

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

a superposition of continuous functions of two variables? Hilbert thought the answer should be negative — surely functions of three variables are more complex than those of two. It should be remarked that this problem resisted all efforts to prove it for more than 50 years. Almost all the mathematicians, interested in this problem, were attempting to prove the validity of Hilbert's conjecture. But in 1957, Kolmogorov (1957) refuted all expectations by proving that each continuous function of three and more variables can be represented by superpositions of continuous functions of one variable and the single function of two variables, namely the addition function $x + y$.

In literature, there are many versions of Kolmogorov's superposition theorem. From the perspective of applications in neural networks, we cite the following version, which is due to Sprecher (1996, 1997):

Theorem 1.1. *Let $d \geq 2$ and $\gamma \geq 2d + 2$ be given integers and $\mathbb{I} = [0, 1]$. There exists a universal monotonic increasing function φ of the class $Lip[\ln 2/\ln \gamma]$ such that every continuous d-variable function $f : \mathbb{I}^d \to \mathbb{R}$ has the representation*

$$f(x_1, \dots, x_d) = \sum_{q=0}^{2d} g_q \left(\sum_{p=1}^{d} \lambda_p \varphi(x_p + aq) \right), \tag{1.1}$$

where g_q is some continuous one-variable function depending on f. Here $a = [\gamma(\gamma - 1)]^{-1}$, $\lambda_1 = 1, \lambda_p = \sum_{r=1}^{\infty} \gamma^{-(p-1)(d^r-1)/(d-1)}$ for $p = 2, 3, \dots, d$.

Eq. (1.1) has the following interpretation as a feedforward neural network consisting of an input layer, two hidden layers and an output layer. The input layer having d neurons sends signals $x_1, \dots, x_d$ to the first hidden layer with $d(2d+1)$ neurons and the activation function φ. The (q, p)-th neuron $y_{q,p}$ $(0 \leq q \leq 2d,\ 1 \leq p \leq d)$ produces the signal $\varphi(x_p + aq)$. These signals are sent to the second hidden layer consisting of $2d + 1$ neurons and the activation functions g_q. The qth neuron z_q $(0 \leq q \leq 2d)$ produces the signal $g_q\left(\sum_{p=1}^{d} \lambda_p y_{q,p}\right)$. Finally, the output layer with the unique output neuron just sums up these last signals to produce the number $f(x_1, \dots, x_d)$. Feedforward neural networks with this structure are usually called the Kolmogorov neural networks. Fig. 1 displays Kolmogorov's neural network in case of $d = 2$.

The relevance of Kolmogorov's superposition theorem to neural networks was first observed by Hecht-Nielsen (1987). In Hecht-Nielsen's interpretation the Kolmogorov neural network (called in Hecht-Nielsen (1987) Kolmogorov's mapping neural network) had three layers, two hidden layers in Fig. 1 were identified as a single layer. Although

* Corresponding author.
E-mail addresses: a.ismayilova@uu.nl (A. Ismayilova), vugaris@mail.ru (V.E. Ismailov).

https://doi.org/10.1016/j.neunet.2024.106333
Received 28 October 2023; Received in revised form 25 February 2024; Accepted 21 April 2024
Available online 22 April 2024

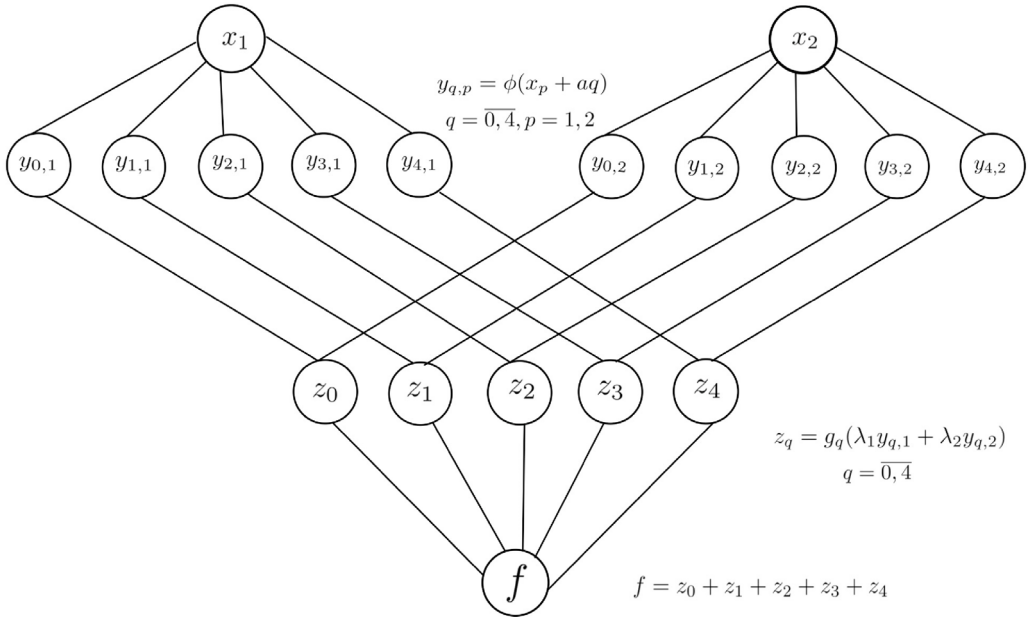

Fig. 1. The neural network interpretation of Kolmogorov's superposition theorem.

Kolmogorov's theorem reveals the profound character of feedforward neural networks to precisely represent each continuous function, it was considered by some authors as non-constructive. For example, Girosi and Poggio pointed out that Kolmogorov's network is not useful. In Girosi and Poggio (1989) they argued that for an implementation of a network that has good properties, the functions corresponding to the layers in the network have to be smooth, which is not the case for the functions φ and g_q in Kolmogorov's network. This criticism was addressed by Kůrkova (1991, 1992) pointing out that the relevance of Kolmogorov's superposition theorem to approximation by neural networks is different. Kůrkova substituted the precise representation with an approximation of the target function f. For this purpose she used sigmoidal functions σ (which is defined as a function with the property $\lim_{t\to-\infty}\sigma(t) = 0$ and $\lim_{t\to+\infty}\sigma(t) = 1$) and finite linear combinations $\sum_k c_k\sigma(w_k x-\theta_k)$ to approximate functions of one variable, in particular Kolmogorov's inner universal and outer functions. As a result, the number of terms in the outer sum was increased, but her method enabled an estimation of the number of hidden neurons depending on the approximation accuracy ε. Note that in Kůrkova (1992) the number of hidden neurons increases when the approximation accuracy $\varepsilon \to \infty$.

The above dependency of ε on the number of neurons was eliminated in Nakamura, Mines, and Kreinovich (1993). They developed algorithms that generate the activation functions with guaranteed accuracy and keep number of hidden neurons independent of ε. All operations and functions in Nakamura et al. (1993) were defined constructively, which means that they are implementable in a computer program. However, as noted by the authors of Nakamura et al. (1993), the algorithms that they constructed in the proofs are very complicated and not suitable for practical usage. For other approximative, but constructive approaches to function approximation by using Kolmogorov's superposition theorem, see de Figueiredo (1980), Igelnik and Parikh (2003) and Nees (1994).

Using Kůrkova's ideas, Katsura and Sprecher (1994) constructed a sequence of functions $\{\varphi_n\}_{n=1}^{\infty}$ that converges to the inner function φ and constructed $2d+1$ series of functions that converge to the outer functions g_q in (1.1).

Later Sprecher (1996, 1997) developed an algorithm for the computation of the inner and outer functions in (1.1). In these papers, the inner function φ is defined as an extension of a function which is explicitly defined on the set of so-called terminating rational numbers. Note that these numbers are dense in $\mathbb{R}$. He proved continuity and monotonicity of the resulting function φ.

The papers (Sprecher, 1996, 1997) were discussed in Köppen (2002), where it was pointed out that Sprecher's function φ does not possess the continuity and monotonicity properties. To fill this gap, Köppen suggested a modified inner function and stated its continuity. He defined φ recursively on the same set of terminating rational numbers D and claimed that this recursion terminates. Köppen assumed that there exists an extension from D to $\mathbb{R}$ as in Sprecher's construction and that this extended φ is monotone increasing and continuous, but he did not give a proof for it. Such a proof was given in Braun and Griebel (2009) and Braun (2009). That is, it was shown that Köppen's φ indeed exists, i.e., it is well defined and has the necessary continuity and monotonicity properties.

2. Main result

In this section we show that the above Theorem 1.1 can be generalized to discontinuous bounded and also to all unbounded functions. In addition, we prove that in all cases the outer functions g_q can be replaced by a single outer function g.

Obviously, all functions on a given compact set X can be divided into the following three nonintersecting classes. These are the classes of continuous, discontinuous bounded and unbounded functions. Note that if a function is continuous on X, then it is automatically bounded. The following theorem gives a precise representation formula for each of these classes.

Theorem 2.1. *Assume $d \geq 2$ and $\gamma \geq 2d+2$ are given integers and $\mathbb{I} = [0,1]$. Set $a = [\gamma(\gamma-1)]^{-1}$, $\lambda_1 = 1, \lambda_p = \sum_{r=1}^{\infty}\gamma^{-(p-1)(d^r-1)/(d-1)}$ for $p = 2,3,\dots,d$ and $b_q = (2d+1)q$ for $q = 0,1,\dots,2d$. Then there exists a universal monotonic increasing function φ of the class $Lip[\ln 2/\ln\gamma]$ with the property:*

Each d-variable function $f : \mathbb{I}^d \to \mathbb{R}$ can be precisely represented in the form

$$f(x_1,\dots,x_d) = \sum_{q=0}^{2d} g\left(\sum_{p=1}^{d}\lambda_p\varphi(x_p+aq)+b_q\right), \tag{2.1}$$

where g is a one-variable function depending on f. If f is continuous, then g can be chosen continuous as well. If f is discontinuous bounded, then g is discontinuous bounded; and if f is unbounded, then g is unbounded.

Theorem 2.1 gives rise to the following feedforward neural network model. This model contains four layers: the input layer with d neurons $x_1,\dots,x_d$, the first hidden layer with $d(2d+1)$ neurons $y_{q,p}$, $0\le q\le 2d$, $1\le p\le d$, the second hidden layer with $2d+1$ neurons z_q, $0\le q\le 2d$, and the output layer with a single neuron w. Activation functions of the first and second hidden layers are φ and g, respectively. The connecting rules between the layers are as follows:

$$\text{The input layer}: \ x_1,\dots,x_d;$$

$$\text{The first hidden layer}: \ y_{q,p}=\varphi(x_p+aq) \text{ for } q=0,1,\dots,2d;\ p=1,\dots,d;$$

$$\text{The second hidden layer}: \ z_q=g\left(\sum_{p=1}^{d}\lambda_p y_{q,p}+b_q\right) \text{ for } q=0,1,\dots,2d;$$

$$\text{The output layer}: \ w=\sum_{q=0}^{2d} z_q.$$

Theorem 2.1 means that any multivariate function $f(x_1,\dots,x_d)$ can be implemented by such a network. That is, $w=f(x_1,\dots,x_d)$. The only parameter depending on f is the activation function g of the second hidden layer. It carries continuity and boundedness properties of the given f. More precisely, g is continuous if f is continuous, g is discontinuous bounded if f is discontinuous bounded, and g is unbounded if f is unbounded.

Proof. Using Sprecher's result, it is easy to prove (2.1) for continuous $f:\mathbb{I}^d\to\mathbb{R}$. Note that for such a function representation (1.1) is valid. In the proof of (1.1) φ was constructed in such a way that $\varphi(\mathbb{I})=\mathbb{I}$ and $\varphi(x+1)=\varphi(x)+1$ for $x\in\mathbb{I}$ (see Sprecher (1996)). This means that $\varphi([0,2])=[0,2]$. Since $0\le x_p+aq<2$ for all indices p and q, we have $0\le\varphi(x_p+aq)<2$. On the other hand it is not difficult to check that $\max\{\lambda_1,\dots,\lambda_p\}=1$. Hence the ranges of all the functions $\sum_{p=1}^{d}\lambda_p\varphi(x_p+aq)$ in (2.1) fall into $[0,2d]$. It follows that the ranges of the functions

$$\Psi_q(x_1,\dots,x_d):=\sum_{p=1}^{d}\lambda_p\varphi(x_p+aq)+b_q,\quad q=0,1,\dots,2d,$$

are pairwise disjoint.

Let Y_q denote the range of $\Psi_q(x_1,\dots,x_d)$. That is,

$$Y_q:=\Psi_q(\mathbb{I}^d) \text{ for all } q=0,1,\dots,2d.$$

Set

$$Y:=\cup_{q=0}^{2d}Y_q.$$

Note that all Y_q and Y are compact sets. Construct the function g on Y by the following way:

$$g(y):=g_q(y-b_q) \text{ if } y\in Y_q,\ q=0,1,\dots,2d. \tag{2.2}$$

Since $Y_i\cap Y_j\neq\emptyset$, for all $0\le i,j\le 2d$, $i\neq j$, this formula makes g a well-defined function on Y. Clearly, g is continuous on Y. This is because g is continuous on each Y_q and the sets $Y_0,\dots,Y_{2d}$ are subsets of pairwise non-overlapping closed intervals $[0,2d]$, $[2d+1,2d+2d+1],\dots,[2d(2d+1),2d+2d(2d+1)]$, respectively. We can extend g by continuity to the whole $\mathbb{R}$. In fact, there are many ways of doing this. Now taking (2.2) into account in (1.1), we obtain (2.1).

Now let us prove (2.1) for bounded $f:\mathbb{I}^d\to\mathbb{R}$. We use the above proved fact that representation (2.1) holds for continuous multivariate functions defined on $\mathbb{I}^d$. Let for any compact set X, $C(X)$ and $B(X)$ stand for the spaces of continuous functions on X and bounded functions on X, respectively. Consider the operator

$$T:C(Y)\to C(\mathbb{I}^d),\ Tg=\sum_{q=0}^{2d}g\left(\sum_{p=1}^{d}\lambda_p\varphi(x_p+aq)+b_q\right). \tag{2.3}$$

Since (2.1) is valid for all $f\in C(\mathbb{I}^d)$, the operator T in (2.3) is a surjection. Consider also the dual spaces $C(Y)^*$ and $C(\mathbb{I}^d)^*$. These are the spaces of regular real-valued measures of finite total variation defined on Borel subsets of Y and $\mathbb{I}^d$, respectively. For solid information on dual spaces we refer the readers to the book by Rudin (1991).

The conjugate operator $T^*:C(\mathbb{I}^d)^*\to C(Y)^*$ has the form

$$T^*\mu=\sum_{q=0}^{2d}\Psi_q\circ\mu.$$

Here $\Psi_q\circ\mu$ is a measure in $C(Y_q)^*$, which is defined as follows

$$\left[\Psi_q\circ\mu\right](Q)=\mu(\Psi_q^{-1}(Q)),\ \text{for all } Q\in\mathcal{B}_q,$$

where $\mathcal{B}_q$ is the set of Borel subsets of Y_q.

It is a well known fact in functional analysis that an operator $F:U\to V$ between Banach spaces U and V is a surjection if and only if the conjugate operator $F^*:V^*\to U^*$ is one-to-one and has a closed range (and vice versa). The last is equivalent to the inequality

$$\|x\|\le\epsilon\left\|F^*x\right\|,$$

for all $x\in V^*$ and some $\epsilon>0$ independent of x (see, e.g., Rudin (1991, Ch. 4, Theorem 4.14)). Applying this fact to our problem, we obtain that there exists a positive number ϵ such that the inequality

$$\|\mu\|\le\epsilon\left\|T^*\mu\right\|=\epsilon\left\|\sum_{q=0}^{2d}\Psi_q\circ\mu\right\|\le\epsilon\sum_{q=0}^{2d}\left\|\Psi_q\circ\mu\right\| \tag{2.4}$$

holds for all $\mu\in C(\mathbb{I}^d)^*$.

For any compact set X, consider a linear space $l_1(X)$ consisting of discrete measures in X. That is, $\nu\in l_1(X)$ means that

$$\nu=\sum_{i=1}^{\infty}a_i\delta_{t_i},\ \|\nu\|=\sum_{i=1}^{\infty}|a_i|<\infty$$

where $\{a_i\}_{i=1}^{\infty}$ is a sequence of real numbers such that $\sum_{i=1}^{\infty}|a_i|<\infty$, $\{t_i\}_{i=1}^{\infty}$ is a sequence in $\mathbb{I}^d$ and δ_{t_i} are point masses at t_i. Note that $l_1(X)\subset C(X)^*$. Thus (2.4) holds also for all $\mu\in l_1(\mathbb{I}^d)$.

Now construct the following Banach space

$$S=\left\{\nu=(\nu_0,\dots,\nu_{2d}):\nu_q\in l_1(Y_q),\ 0\le q\le 2d\right\},\ \|\nu\|=\sum_{q=0}^{2d}\left\|\nu_q\right\|<+\infty.$$

Consider its dual S^*. Since for any X, the dual of $l_1(X)$ is the space of bounded functions on X (see, e.g., Dunford and Schwartz (1959, Ch. 4)), the space S^* has the following structure:

$$S^*=\left\{g=(g_0,\dots,g_{2d}):g_q\in B(Y_q),\ 0\le q\le 2d\right\},$$
$$\|g\|=\max_{0\le q\le 2d}\left\|g_q\right\|<+\infty.$$

Note that $g\in S^*$ acts on $\nu\in S$ as follows

$$g[\nu]=\sum_{q=0}^{2d}\int_{Y_q}g_q d\nu_q.$$

Consider the operator

$$Z:l_1(\mathbb{I}^d)\to S;\ Z(\mu)=\left(\Psi_0\circ\mu,\dots,\Psi_{2d}\circ\mu\right)$$

and its conjugate

$$Z^*:S^*\to l_1(\mathbb{I}^d)^*.$$

Note that by the definition of conjugate operator

$$Z^*g[\mu]=g[Z\mu]=\sum_{q=0}^{2d}\int_{Y_q}g_q d(\Psi_q\circ\mu)=\sum_{q=0}^{2d}\int_{\mathbb{I}^d}g_q\circ\Psi_q d\mu$$
$$=\int_{\mathbb{I}^d}\left(\sum_{q=0}^{2d}g_q\circ\Psi_q\right)d\mu.$$

Therefore,

$$Z^*g=\sum_{q=0}^{2d}g_q\circ\Psi_q. \tag{2.5}$$

Again by the above mentioned fact of functional analysis the operator Z^* is surjective if and only if there exists $\epsilon > 0$ such that

$$\|\mu\| \le \epsilon \|Z\mu\| = \epsilon \sum_{q=0}^{2d} \left\| \Psi_q \circ \mu \right\|$$

holds for all $\mu \in l_1(\mathbb{I}^d)$. But we have seen above in (2.4) that this inequality indeed holds for all $\mu \in C(\mathbb{I}^d)^*$, hence for all $\mu \in l_1(\mathbb{I}^d) \subset C(\mathbb{I}^d)^*$. This means that Z is an injective operator with closed range. We obtain that Z^* is a surjective operator and hence its range is all of $l_1(\mathbb{I}^d)^* = B(\mathbb{I}^d)$. Comparing this assertion with (2.5) gives that for any bounded function $f \in B(\mathbb{I}^d)$ the representation

$$f = \sum_{q=0}^{2d} g_q \circ \Psi_q,$$

is valid for some bounded $g_q \in B(Y_q)$. Since the sets Y_q, $q = 0, 1, \dots, 2d$, are pairwise disjoint, we can construct a single function g as in (2.2), which is well defined and bounded on Y. One can extend g to the whole real line in such a way that g remains bounded also on $\mathbb{R}$. Thus for any $f \in B(\mathbb{I}^d)$ we have the precise representation

$$f(x_1, \dots, x_d) = \sum_{q=0}^{2d} g(\Psi_q(x_1, \dots, x_d)), \tag{2.6}$$

where $g \in B(\mathbb{R})$. If f is bounded and discontinuous, then g will be discontinuous as well, otherwise the right-hand side of (2.6) will be a continuous function whereas the left-hand side is discontinuous. This proves the second part of the theorem. It should be remarked that the method of using $l_1(X)$ measures in representation of bounded functions on X was introduced by Sternfeld (1978) and refined by Khavinson (1997, Ch. 1).

Now let us prove the theorem for an unbounded function $f : \mathbb{I}^d \to \mathbb{R}$. In fact, we will prove the validity of (2.1) for any d-variable function f defined on $\mathbb{I}^d$. Certainly, if f is unbounded, then g must be unbounded too, otherwise the right hand side of (2.1) would be bounded, contradicting the equality in (2.1). In Ismailov (2023), the second author proved a slightly different version of this part, where not a single but $2d+1$ outer functions g_q in the representation formula are needed. For completeness we repeat the main details of that proof here and show where and how all that g_q can be replaced by a single function g. Again we use the representation formula (2.1) for continuous functions, which has been proved above. In fact, we will use the following concise form of (2.1)

$$f(x_1, \dots, x_d) = \sum_{q=0}^{2d} g(\Psi_q(x_1, \dots, x_d)), \tag{2.7}$$

where $f \in C(\mathbb{I}^d)$, $g \in C(\mathbb{R})$ and Ψ_q are defined above. From (2.7) we can obtain the following important property of the family of functions $\{\Psi_q\}_{q=0}^{2d}$, which we formulate as a lemma.

Lemma 2.1. *For any finite set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{I}^d$ the system of equations*

$$\sum_{j=1}^{n} \mu_j \delta_{\Psi_q(\mathbf{x}_j)} = 0, \ \ q = 0, \dots, 2d, \tag{2.8}$$

with respect to μ_j has only a zero solution.

In this lemma and in the sequel $\delta_{\mathcal{A}}$ stands for the indicator function of a set $\mathcal{A} \subset \mathbb{R}$. That is,

$$\delta_{\mathcal{A}}(t) = \begin{cases} 1, & \text{if } t \in \mathcal{A} \\ 0, & \text{if } t \notin \mathcal{A} \end{cases}$$

Note that in (2.8) $\delta_{\Psi_q(\mathbf{x}_j)}$ are indicator functions of the single point sets $\{\Psi_q(\mathbf{x}_j)\}$.

Let us explain Eq. (2.8) in detail. We will see that it stands for a system of certain linear equations. To show this, fix the subscript q. Let the set $\{\Psi_q(\mathbf{x}_j),\ j = 1, \dots, n\}$ have s_q different values, which we denote by $\gamma_1^q, \gamma_2^q, \dots, \gamma_{s_q}^q$. Then (2.8) implies that

$$\sum_j \mu_j = 0,$$

where the sum is taken over all j such that $\Psi_q(\mathbf{x}_j) = \gamma_k^q$, $k = 1, \dots, s_q$. Thus for fixed q we have s_q linear homogeneous equations in $\mu_1, \dots, \mu_n$. The coefficients of these equations are 0 and 1. By varying q, we will have $\sum_{q=0}^{2d} s_q$ such equations. Thus (2.8), in its expanded form, stands for the system of these equations.

Note that not only $\{\Psi_q\}_{q=0}^{2d}$, but a family of arbitrarily given multivariate functions $\{h_1, \dots, h_k\}$ on any set X can generate (2.8). It should be remarked that in this case finite sets $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ satisfying the system of Eqs. (2.8) when there is a nonzero solution $(\mu_1, \dots, \mu_n)$ were exploited under the name of "closed paths" in several works of the second author (see, e.g., Ismailov (2012, 2017, 2021)).

Let us now prove the lemma. Assume the contrary. Assume that there is a finite set $p = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ in $\mathbb{I}^d$ such that the system of Eqs. (2.8) has a nonzero solution $\mu = (\mu_1, \dots, \mu_n)$. Without loss of generality we may assume that all the numbers $\mu_j \neq 0$, $j = 1, \dots, n$. Otherwise we can remove all zero components μ_j from μ and the corresponding $\mathbf{x}_j$ (having the same index) from p and consider the resulting sets $\{\mu_j\}$ and $\{\mathbf{x}_j\}$. Consider the linear functional

$$G_p(f) = \sum_{j=1}^{n} \mu_j f(\mathbf{x}_j), \ \ G_p : C(\mathbb{I}^d) \to \mathbb{R}.$$

This functional annihilates all sums of the form $\sum_{q=0}^{2d} g(\Psi_q(x_1, \dots, x_d))$, and hence, according to (2.7), every function $f \in C(\mathbb{I}^d)$. That is, $G_p(f) = 0$ for any $f \in C(\mathbb{I}^d)$. On the other hand by Urysohn's lemma (see, e.g., Willard (1970, Ch. 5)) there exists a continuous function f_0 with the property: $f_0(\mathbf{x}_j) = 1$ for indices j such that $\mu_j > 0$; $f_0(\mathbf{x}_j) = -1$ for indices j such that $\mu_j < 0$; and $-1 < f_0(\mathbf{x}) < 1$ for $\mathbf{x} \in \mathbb{I}^d \setminus p$. For this function we have $G_p(f_0) = \sum_{j=1}^{n} \left|\mu_j\right| \neq 0$. The obtained contradiction means that Eq. (2.8) has only a zero solution $(\mu_1, \dots, \mu_n)$ for any finite subset $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{I}^d$. The lemma has been proved.

Now let us return to the proof of the third part of our theorem. We want to prove that for any function $f : \mathbb{I}^d \to \mathbb{R}$, the following representation holds

$$f(x_1, \dots, x_d) = \sum_{q=0}^{2d} g(\Psi_q(x_1, \dots, x_d)),$$

where g is a one-variable function depending on f.

Recall that we denoted ranges of Ψ_q by Y_q and Y is the union of Y_q. Consider the following set

$$\mathcal{L} = \{A = \{y_0, \dots, y_{2d}\} : \text{if there exists } \mathbf{x} \in \mathbb{I}^d \text{ s.t. } \Psi_q(\mathbf{x}) = y_q, \ q = 0, \dots, 2d\}. \tag{2.9}$$

Note that $\mathcal{L}$ is not a subset of Y. It is a set of some special subsets of Y. Each element of $\mathcal{L}$ is a set $A = \{y_0, \dots, y_{2d}\} \subset Y$ with the property that there exists at least one point $\mathbf{x} \in \mathbb{I}^d$ such that $\Psi_q(\mathbf{x}) = y_q$, $q = 0, \dots, 2d$. These $\mathbf{x}$ will be called generating points for A. It follows from Lemma 2.1 that in (2.9) for each element A there exists only one generating point $\mathbf{x} \in \mathbb{I}^d$. Since each $A \in \mathcal{L}$ corresponds to only one generating point $\mathbf{x} \in \mathbb{I}^d$, we can define the following set function

$$\tau : \mathcal{L} \to \mathbb{R}, \ \ \tau(A) = f(\mathbf{x}).$$

Consider now a class S of functions of the form $\sum_{j=1}^{m} r_j \delta_{D_j}$, where m is a positive integer, r_j are real numbers and D_j are elements of $\mathcal{L}$, $j = 1, \dots, m$. We fix neither the numbers m, r_j, nor the sets D_j. Clearly, S is a linear space. On elements of S, we define the linear functional

$$F : S \to \mathbb{R}, \ \ F\left(\sum_{j=1}^{m} r_j \delta_{D_j}\right) = \sum_{j=1}^{m} r_j \tau(D_j).$$

Introduce the linear space:

$$S' = \left\{ \sum_{j=1}^{m} r_j \delta_{\omega_j} \right\},$$

where $m \in \mathbb{N}$, $r_j \in \mathbb{R}$, $\omega_j \subset Y$. As above, we do not fix the parameters m, r_j and ω_j. Note that now we use not only the special subsets D_j of Y, but all possible subsets $\omega_j \subset Y$. Obviously, the space S' is larger than S. Consider the linear extension of F to the space S', which we denote by F'. That is, $F' : S' \to \mathbb{R}$ and $F'(z) = F(z)$ for all $z \in S$. The existence of such an extension is based on Zorn's lemma.

Define the following functions:

$$g_q : Y_q \to \mathbb{R},\ g_q(y_q) := F'(\delta_{y_q}),\ q = 0, \dots, 2d.$$

Since the sets Y_q are pairwise disjoint we can also define the single function

$$g : Y \to \mathbb{R},\ g(y) = g_q(y),\ \text{if } y \in Y_q. \tag{2.10}$$

Clearly, (2.10) correctly defines g on the whole Y, the union of ranges of Ψ_q.

Let now $\mathbf{x} = (x_1, \dots, x_d)$ be an arbitrary point in $\mathbb{I}^d$. Obviously, $\mathbf{x}$ is a generating point for some set $A = \{y_0, \dots, y_{2d}\} \in \mathcal{L}$. Thus we can write that

$$f(\mathbf{x}) = \tau(A) = F(\delta_A) = F\left(\sum_{q=0}^{2d} \delta_{y_q} \right) = F'\left(\sum_{q=0}^{2d} \delta_{y_q} \right) =$$
$$\sum_{q=0}^{2d} F'(\delta_{y_q}) = \sum_{q=0}^{2d} g_q(y_q) = \sum_{q=0}^{2d} g(y_q) = \sum_{q=0}^{2d} g(\Psi_q(\mathbf{x})).$$

This proves the theorem for all functions $f : \mathbb{I}^d \to \mathbb{R}$, hence for unbounded f.

Remark 1. The outer function g in (2.1) can be computed via g_q by using (2.2). The functions g_q, $q = 0, \dots, 2d$, in turn, can be found by applying Sprecher's iterative method (see Sprecher (1997)). This method constructs one-variable functions $g_q(y)$ with an algorithm which produces for each q a sequence of functions $\left\{ g_q^n(y) \right\}_{n=1}^{\infty}$ such that $\lim_{r\to\infty} \sum_{n=1}^{r} g_q^n(y) = g_q(y)$. Let us describe this algorithm. Starting with $f_0 = f$, for $n = 1, 2, \dots$ iterate the following steps.

Step 1. For $f_{n-1}(\mathbf{x})$ determine an integer k_n such that for any pairs $\mathbf{x}, \mathbf{x}' \in \mathbb{I}$ satisfying $|\mathbf{x} - \mathbf{x}'| \le \gamma^{-k_n}$ it holds that

$$|f_{n-1}(\mathbf{x}) - f_{n-1}(\mathbf{x}')| \le \varepsilon \|f_{n-1}\|,$$

where $0 < \varepsilon < 1/(d+1)$. This k_n determines rational coordinate points

$$\mathbf{c}_{k_n}^q = (c_{k_n,1}^q, \dots, c_{k_n,d}^q) \in D_{k_n}^d,$$

where

$$D_k \overset{def}{=} \left\{ c_k \in \mathbb{Q} : c_k = \sum_{r=1}^{k} i_r \gamma^{-r},\ i_r \in \{0, 1, \dots, \gamma - 1\} \right\},\ k \in \mathbb{N}.$$

Note that the union $D = \cup_{k\in\mathbb{N}} D_k$, which is called the set of terminating rational numbers, is dense in $\mathbb{I}$.

Step 2. For all $\mathbf{c}_{k_n}^q = (c_{k_n,1}^q, \dots, c_{k_n,d}^q) \in D_{k_n}^d$ compute the values

$$\xi(\mathbf{c}_{k_n}^q) = \sum_{p=1}^{d} \lambda_p \varphi\left(c_{k_n,p}^q + q \sum_{r=2}^{k_n} \gamma^{-r} \right)$$

and the one-variable functions

$$\omega(\mathbf{c}_{k_n}^q; y) = \sigma\left(\gamma^{\beta(k_n+1)} (y - \xi(\mathbf{c}_{k_n}^q)) + 1 \right)$$
$$- \sigma\left(\gamma^{\beta(k_n+1)} (y - \xi(\mathbf{c}_{k_n}^q) - (\gamma - 2) b_{k_n}) \right).$$

Here $\sigma : \mathbb{R} \to \mathbb{R}$ is an arbitrary continuous function with $\sigma(t) = 0$ when $t \le 0$ and $\sigma(t) = 1$ when $t \ge 1$; β is a function defined as $\beta(r) = (d^r - 1)/(d-1)$ for $r = 1, 2, \dots$, and $b_{k_n} = \sum_{r=k_n+1}^{\infty} \gamma^{-\beta(r)} \sum_{p=1}^{d} \lambda_p$.

Step 3. Compute the one-variable functions

$$g_q^n(y) = \frac{1}{2d+1} \sum_{\mathbf{c}_{k_n}^q} f_{n-1}(\mathbf{c}_{k_n}^q) \omega(\mathbf{c}_{k_n}^q; y),$$

where for each $q \in \{0, 1, \dots, 2d\}$ the above sum is taken over all values $\mathbf{c}_{k_n}^q \in D_{k_n}^d$.

Step 4. Substitute the transfer functions $\Phi_q(\mathbf{x}) = \sum_{p=1}^{d} \lambda_p \varphi(x_p + aq)$ and compute the d-variable functions

$$g_q^n(\Phi_q(\mathbf{x})) = \frac{1}{2d+1} \sum_{\mathbf{c}_{k_n}^q} f_{n-1}(\mathbf{c}_{k_n}^q) \omega(\mathbf{c}_{k_n}^q; \Phi_q(\mathbf{x})),\ q = 0, 1, \dots, 2d.$$

Step 5. Compute the function

$$f_n(x) = f(x) - \sum_{q=0}^{2d} \sum_{j=1}^{n} g_q^j(\Phi_q(\mathbf{x})).$$

This completes the nth iteration loop. Now replace n by $n+1$ and go to Step 1.

Remark 2. Theorem 2.1 is about precise representation of continuous and discontinuous multivariate functions by the Kolmogorov neural network model. The above interpretation of this model is not new and was given in many papers (see, e.g, Braun (2009) and related references therein). Here we do not discuss problems on approximation of the inner and outer functions in (2.1) by practical feedforward neural networks with popular activation functions. We think that this is an interesting topic for future research, which requires a specific approach. The difficulty here is in finding a suitable Kolmogorov type representation formula guaranteeing good neural network approximation scheme. It should be remarked that one such approach, involving deep ReLU networks, was described in Schmidt-Hieber (2021). Namely, a modified version of the Kolmogorov representation theorem was derived and then it was constructively proven that the inner and outer functions in such representation can be well approximated by deep ReLU networks (see Schmidt-Hieber (2021)).

However, the above Kolmogorov neural network model is linked with traditional multilayer neural networks. Specifically, this model can be approximated with arbitrary precision by two-hidden-layer feedforward networks with a sigmoidal activation function and limited neurons in each hidden layer. More precisely, one can construct a sigmoidal, almost monotone, infinitely smooth activation function σ such that for any $f \in C(\mathbb{I}^d)$ and any $\varepsilon > 0$, there exist constants d_p, c_{pq}, θ_{pq}, γ_p, and vectors $w^{pq} \in \mathbb{R}^d$ for which

$$\left| f(\mathbf{x}) - \sum_{p=1}^{2d+2} d_p \sigma\left(\sum_{q=1}^{d} c_{pq} \sigma(\mathbf{w}^{pq} \cdot \mathbf{x} - \theta_{pq}) - \gamma_p \right) \right| < \varepsilon$$

for all $\mathbf{x} \in \mathbb{I}^d$. Note that the above σ is a practically computable function. The paper (Guliyev & Ismailov, 2018) provides a programming algorithm that allows one to compute σ at any point of the real axis instantly.

3. Conclusions and remarks

The topic on the role of Kolmogorov superposition theorem in neural network theory is still active today (see, e.g., Jorgensen and Tian (2020), Montanelli and Yang (2020), Schmidt-Hieber (2021), Shen, Yang, and Zhang (2021)). The research in this area was developed mainly in two directions. In the first direction, the analysis was concentrated on approximative versions of Kolmogorov's superposition theorem and obtaining corresponding results on neural network approximation (see, e.g., Guliyev and Ismailov (2018), Ismailov (2014, 2021), Kůrkova (1991, 1992), Maiorov and Pinkus (1999)). In the second direction, representation power of Kolmogorov superposition-based neural networks were studied (see, e.g., Brattka (2007), Ismailov

(2023), Katsura and Sprecher (1994), Sprecher (1993, 1996, 1997, 2014), Sprecher and Draghici (2002)). Due to these second type works, there is a perspective for a practical usage of the precise representation of multivariate functions by Kolmogorov type neural networks.

This paper studies the Kolmogorov two-hidden-layer neural network model with one-variable activation functions φ and g in the first and second hidden layers, respectively. It shows that each multivariate function f can be precisely represented by this model. All parameters of the network except the second activation g are fixed and do not depend on f. The main result proves that if f is continuous, then g can be chosen continuous as well. Further if f is discontinuous bounded, then g is discontinuous bounded; and if f is unbounded, then g is unbounded.

It should be remarked that existence results and construction methods for the universal inner function φ were given in the papers (Braun, 2009; Braun & Griebel, 2009; Köppen, 2002; Sprecher, 1996) (see Introduction). Using these methods one can easily construct the functions Ψ_q, since all the numbers in their definitions are explicitly known. A numerical algorithm for the parallel computations g_q in the representation formula (1.1) was developed in Sprecher (1997). Taking into account (2.2) one can apply Sprecher's method for computation of the single outer function g (see Remark 1). These tips refer to computational aspects of the case when in Theorem 2.1 only continuous functions are involved. A practical construction of g in cases with discontinuous bounded and unbounded functions is not yet known. For such cases Theorem 2.1 gives only a theoretical understanding of the representation problem. This is because for the representation of discontinuous bounded functions we have derived (2.1) from the fact that the range of the operator Z^* is the whole space of bounded functions $B(\mathbb{I}^d)$. This fact directly gives us a formula (2.1) but does not tell how the bounded one-variable function g is attained. For the representation of unbounded functions we have used a linear extension of the functional F, existence of which is based on Zorn's lemma (see, e.g., Kolmogorov and Fomin (1989, Ch. 3)). Application of Zorn's lemma provides no mechanism for practical construction of such an extension. Zorn's lemma helps to assert only its existence.

CRediT authorship contribution statement

Aysu Ismayilova: Formal analysis, Investigation, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Vugar E. Ismailov:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

We declare that we have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Brattka, V. (2007). From Hilbert's 13th problem to the theory of neural networks: Constructive aspects of Kolmogorov's superposition theorem. In *Kolmogorov's heritage in mathematics* (pp. 253–280). Berlin: Springer.

Braun, J. (2009). *An application of Kolmogorov's superposition theorem to function reconstruction in higher dimensions* (Ph.D. dissertation), Universitat Bonn.

Braun, J., & Griebel, M. (2009). On a constructive proof of Kolmogorov's superposition theorem. *Constructive Approximation*, *30*(3), 653–675.

de Figueiredo, R. J. P. (1980). Implications and applications of Kolmogorov's superposition theorem. *IEEE Transactions on Automatic Control, AC*, *25*, 1227–1231.

Dunford, N., & Schwartz, J. T. (1959). *Linear operators. art I*. New York: Interscience.

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

Guliyev, N. J., & Ismailov, V. E. (2018). Approximation capability of two hidden layer feedforward neural networks with fixed weights. *Neurocomputing*, *316*, 262–269.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In *Proc. I987 IEEE int. conf. on neural networks, vol. 3* (pp. 11–14). New York: IEEE Press.

Igelnik, B., & Parikh, N. (2003). Kolmogorov's spline network. *IEEE Transactions on Neural Networks*, *14*, 725–733.

Ismailov, V. E. (2012). A note on the representation of continuous functions by linear superpositions. *Expositiones Mathematicae*, *30*, 96–101.

Ismailov, V. E. (2014). On the approximation by neural networks with bounded number of neurons in hidden layers. *Journal of Mathematical Analysis and Applications*, *417*(2), 963–969.

Ismailov, V. E. (2017). On the uniqueness of representation by linear superpositions. *Ukrainian Mathematical Journal*, *68*(12), 1874–1883.

Ismailov, V. E. (2021). *Mathematical surveys and monographs*: *vol. 263*, *Ridge functions and applications in neural networks* (p. 186). American Mathematical Society.

Ismailov, V. E. (2023). A three layer neural network can represent any multivariate function. *Journal of Mathematical Analysis and Applications*, *523*(1), 8, Paper(127096).

Jorgensen, P. E. T., & Tian, J. F. (2020). Superposition, reduction of multivariable problems, and approximation. *Analysis and Applications*, *18*(5), 771–801.

Katsura, H., & Sprecher, D. A. (1994). Computational aspects of Kolmogorov's superposition theorem. *Neural Networks*, *7*, 455–461.

Khavinson, S. Ya (1997). *Translations of mathematical monographs*: *vol. 159*, *Best approximation by linear superpositions (approximate nomography), translated from the Russian manuscript by D. Khavinson* (p. 175). Providence, RI: American Mathematical Society.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. (Russian). *Doklady Akademii Nauk SSSR*, *114*, 953–956.

Kolmogorov, A. N., & Fomin, S. V. (1989). *Elements of the theory of functions and functional analysis* (6th ed.). (p. 624). Moscow: Nauka.

Köppen, M. (2002). On the training of a Kolmogorov network. In *Lecture notes in computer science*: *vol. 2415*, *ICANN 2002* (pp. 474–479).

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

Kůrkova, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, *5*, 501–506.

Maiorov, V., & Pinkus, A. (1999). Lower bounds for approximation by MLP neural networks. *Neurocomputing*, *25*, 81–91.

Montanelli, H., & Yang, H. (2020). Error bounds for deep ReLU networks using the Kolmogorov-Arnold superposition theorem. *Neural Networks*, *129*, 1–6.

Nakamura, M., Mines, R., & Kreinovich, V. (1993). Guaranteed intervals for Kolmogorov's theorem (and their possible relation to neural networks). *Interval Computations*, *3*, 183–199.

Nees, M. (1994). Approximative versions of Kolmogorov's superposition theorem, proved constructively. *Journal of Computational and Applied Mathematics*, *54*, 239–250.

Rudin, W. (1991). *International series in pure and applied mathematics*, *Functional analysis* (2nd ed.). (p. 424). New York: McGraw-Hill, Inc..

Schmidt-Hieber, J. (2021). The Kolmogorov-Arnold representation theorem revisited. *Neural Networks*, *137*, 119–126.

Shen, Z., Yang, H., & Zhang, S. (2021). Neural network approximation: Three hidden layers are enough. *Neural Networks*, *141*, 160–173.

Sprecher, D. A. (1993). A universal mapping for Kolmogorov's superposition theorem. *Neural Networks*, *6*, 1089–1094.

Sprecher, D. A. (1996). A numerical implementation of Kolmogorov's superpositions. *Neural Networks*, *9*, 765–772.

Sprecher, D. A. (1997). A numerical implementation of Kolmogorov's superpositions II. *Neural Networks*, *10*, 447–457.

Sprecher, D. A. (2014). On computational algorithms for real-valued continuous functions of several variables. *Neural Networks*, *59*, 16–22.

Sprecher, D. A., & Draghici, S. (2002). Space-filling curves and Kolmogorov superposition-based neural networks. *Neural Networks*, *15*, 57–67.

Sternfeld, Y. (1978). Uniformly separating families of functions. *Israel Journal of Mathematics*, *29*, 61–91.

Willard, S. (1970). *General topology* (p. 369). Mass.-London-Don Mills, Ont: Addison-Wesley Publishing Co. Reading.