Remarks on the Universal Approximation Property of Feedforward Neural Networks

Jiří Kupka[1,*,†], Zahra Alijani[1,†] and Petra Števuliáková[1,†]

[1]*Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, Centre of Excellence IT4Innovations, 30. dubna 22, 702 00 Ostrava, Czech Republic*

Abstract

This paper presents a structured overview and novel insights into the universal approximation property of feedforward neural networks. We categorize existing results based on the characteristics of activation functions — ranging from strictly monotonic to weakly monotonic and continuous almost everywhere — and examine their implications under architectural constraints such as bounded depth and width. Building on classical results by Cybenko [1], Hornik [2], and Maiorov [3], we introduce new activation functions that enable even simpler neural network architectures to retain universal approximation capabilities. Notably, we demonstrate that single-layer networks with only two neurons and fixed weights can approximate any continuous univariate function, and that two-layer networks can extend this capability to multivariate functions. These findings refine the known lower bounds of neural network complexity and offer constructive approaches that preserve strict monotonicity, improving upon prior work that relied on relaxed monotonicity conditions. Our results contribute to the theoretical foundation of neural networks and open pathways for designing minimal yet expressive architectures.

Keywords

Universal Approximation Theorem, Neural Network, Activation Function

1. Introduction

In this paper, we would like to contribute to mathematical theoretical backgrounds of neural networks, which are nowadays used in various parts of our lives, not only in industrial applications (e.g. through image and video processing tools) but also in many aspects of real life, like automated medical and psychological diagnosis [4, 5, 6], automated detection of mental conditions, etc. We want to discuss a bit one of the fundamental results in neural networks, namely the **universal approximation theorem**. This theorem, in its purest form, states that a feedforward neural network with a single hidden layer can approximate any continuous function on a compact subset of $\mathbb{R}^n$, provided it has a sufficient number of neurons in the hidden layer. However, these results generally do not specify how many neurons are required to achieve this approximation.

Most universal approximation results fall into one of two categories:

- **Arbitrary Width:** Networks with a limited number of hidden layers but an unrestricted number of neurons.
- **Arbitrary Depth:** Networks with an unrestricted number of layers, each with a limited number of neurons.

Some results do not fit neatly into either category, for instance, they can be addressing networks with both bounded width and bounded depth.

It is essential to note that to obtain an upper bound for width, one needs to construct approximations with arbitrary precision for each function within the specified target function class. Conversely,

ITAT'25: Information Technologies – Applications and Theory, September 26–30, 2025, Telgárt, Slovakia

*Corresponding author.

†These authors contributed equally.

Jiri.Kupka@osu.cz (J. Kupka); Zahra.Alijani@osu.cz (Z. Alijani); Petra.Stevuliakova@osu.cz (P. Števuliáková)

0000-0001-5940-0576 (J. Kupka); 000-0002-1448-9068 (Z. Alijani); 0000-0002-7879-1397 (P. Števuliáková)

establishing a lower bound for width only requires identifying one function in the class that the neural network architecture cannot approximate when its width is below the bound. Therefore, proving upper bounds for the minimum width is generally regarded as more difficult than demonstrating lower bounds.

While Section 2 provides some fundamental results and notation, Section 3 consists of a short survey of classical and recent universal approximation theorems, classified mainly due to the type of activation function, which is a kind of new approach, to our best knowledge. In the last section, we mentioned, without proofs, our recent results [7], which enrich some of the results mentioned in Section 3. In brief, our results provide new lower boundaries of the complexity of feedforward single-hidden-layer and two-hidden-layer neural networks, still providing the property of universal approximation.

2. Notation and a Fundamental Result

In this section, we briefly recall the basic concepts and fundamental facts about neural networks, activation functions, and universal approximation properties. Our objective is to understand and apply the universal approximation property using a wide range of activation functions, without restricting ourselves to specific structural assumptions. The following types of neural networks are the basic objects of our study.

Single-Layer Feedforward Neural Network (SLFN):

$$f_N(\bar{x}) := \sum_{i=1}^{N} a_i \sigma(\bar{w}^i \cdot \bar{x} - b_i)$$

where:

- $\bar{x} \in \mathbb{R}^d$ is the input vector.
- $\bar{w}^i \in \mathbb{R}^d$ are the weights for the i-th neuron in the hidden layer.
- $b_i \in \mathbb{R}$ are bias terms.
- $a_i \in \mathbb{R}$ are output weights.
- $\sigma(\cdot)$ is an univariate activation function.

Two-Layer Feedforward Neural Network (TLFN):

$$f_{NM}(\bar{x}) := \sum_{i=1}^{M} c_i \sigma\left(\sum_{j=1}^{N} a_{ij} \sigma(\bar{w}^{ij} \cdot \bar{x} - b_{ij}) - d_i \right),$$

where $\bar{x}, \bar{w}^{ij} \in \mathbb{R}^d$; $a_{ij}, b_{ij}, c_i, d_i \in \mathbb{R}$ and $\sigma(\cdot)$ is the fixed univariate activation function.

The foundational results on universal approximation property of neural networks are from 1989 due to Cybenko [1], Funahashi [8] and Hornik et al. [2]. The authors independently proved that a neural network with a single hidden layer can approximate any continuous function on a compact domain. In addition, Cybenko [1] used a continuous sigmoidal activation function, Funahashi [8] worked with a continuous activation function that is nonconstant, bounded and monotone increasing, and similarly Hornik et al. [2] used a continuous nonconstant activation function. Two years later, Hornik (in [9]) proved that not the specific choice of the activation function but rather the feed-forward architecture ensures the required property. The results presented by Cybenko, Funahashi and Hornik et al. are not constructive in a simple way, constructive approximations were first presented in [10, 11].

Below we denote the space of continuous functions $f : [0,1]^d \to \mathbb{R}$ by $C([0,1]^d)$, and by C^∞ map we denote a map for which every derivative exists and is continuous.

Theorem 2.1 (Cybenko, 1989)**.** *For any continuous function $f \in C([0,1]^d)$ and any $\epsilon > 0$, there exists a neural network with a finite number of neurons such that:*

$$\sup_{\bar{x} \in [0,1]^d} |f(\bar{x}) - f_N(\bar{x})| < \epsilon.$$

This result relies on:

- The Stone-Weierstrass theorem, ensuring the density of polynomials in $C([0,1]^d)$.
- The discriminatory property of sigmoidal functions, which guarantees their ability to separate measures.

The case of *arbitrary depth* has been extensively studied by many authors, particularly in the context of neural networks with ReLU activation functions (see, e.g., [12], [13], [14], [15]). More recently, Kidger generalized these results to a broader class of activation functions in [16], extending the applicability of universal approximation theorems beyond ReLU-based architectures. Building on this, recent work [17] introduces fractional order derivatives into activation functions, offering tunable flexibility that helps networks better capture complex patterns and improve learning performance.

An intriguing and novel approach was introduced in [18], where the authors demonstrated that universal approximation can be achieved using a finite set of mappings. This vocabulary-based method opens new perspectives on the structure and design of neural networks for function approximation.

3. State-of-the-Art Knowledge - Structured According to Properties of Activation Functions

Quite a number of theorems on universal approximation can be found in the literature, and it seems that the ideas in the proofs vary a lot. In the following, we structure theorems on the universal approximation property according to the properties of activation functions. The idea is to focus mainly on the case of neural networks with bounded depth and width.

3.1. Strictly Monotone Continuous Activation Functions

In the bounded-width setting, neural networks are restricted to a fixed number of neurons per layer but can have arbitrary depth. Kidger and Lyons [16] showed that such deep networks, with width as small as $n + m + 2$ (where n is the input dimension and m the output dimension), are still universal approximators—provided the activation function is continuous, non-affine, and non-polynomial. Their result significantly generalizes earlier work, which primarily focused on ReLU-based architectures, by extending the universality to a broader class of activation functions.

In the bounded-depth setting, the depth of the neural network is fixed, while the width is allowed to grow. Classical results by Cybenko [1] and Hornik et al. [2] demonstrate that shallow networks—specifically, those with a single hidden layer—can approximate any continuous function on a compact domain, provided the width is sufficiently large. These foundational results highlight the expressive power of wide and shallow architectures.

More recently, Cai [18] introduced a novel constructive approach to universal approximation using a finite set of mappings, referred to as a ”vocabulary”. This method enables the approximation of continuous functions by composing a fixed set of nonlinear transformations, offering a compact and interpretable representation of the function space. While the vocabulary-based method is not limited to shallow networks, it provides new insights into how expressive power can be achieved even under architectural constraints.

Additionally, Kratsios [19] provided a general characterization of universal approximation under various architectural constraints, including bounded depth. His work shows that with appropriate modifications—such as sparse connectivity or shifted activation functions—universal approximation can still be achieved, even when depth is limited.

Maiorov et al. in 1999 ([3]) proved that there exists an analytic, strictly increasing sigmoidal activation function for which the neural network with limited width and depth is a universal approximator. Firstly, they proved that there exists an activation function for which single hidden layer neural network approximation is essentially identical (same approximation order) to that of ridge function

approximation. The theoretical lower bound is given by the approximation order of the manifold M_n restricted to the unit ball B^d in $\mathbb{R}^d$ with a boundary S^{d-1} defined as follows:

$$M_n = \left\{ \sum_{i=1}^{n} g_i(\bar{a}^i \cdot \bar{x} \,|\, \bar{a}^i \in S^{d-1},\ g_i \in C([-1,1])) \right\}. \tag{1}$$

Theorem 3.1. *[3] There exists an activation function σ which is real analytic, strictly increasing, and sigmoidal satisfying the following. Given $f \in M_n$ and $\epsilon > 0$, there exist constants c_i, integers r_i and vectors $\bar{w}^i \in S^{d-1}$ such that*

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

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

Secondly in [3], they considered the neural network with two hidden layers. They showed that for their constructed activation function, any continuous function on the unit cube in $\mathbb{R}^d$ can be uniformly approximated with any error.

Theorem 3.2. *[3] 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 $\epsilon > 0$, there exist real constants $d_i, c_{ij}, \theta_{ij}, \gamma_i$, and vectors $\bar{w}^{ij} \in \mathbb{R}^d$ for which*

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

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

Proof. The proof is based on an improved version of Kolmogorov Superposition Theorem and the activation function constructed by polynomials with rational coefficients. □

Remark 1. *If there is a will to replace the demand of analyticity of σ by only C^∞, then Theorem 3.2 can be stated with $2d+1$ units in the first layer and $4d+3$ units in the second layer. Similarly, for the demand of strict monotonicity and sigmoidality, Theorem 3.2 can be proven with d units in the first layer and $2d+1$ units in the second layer. The restriction of Theorem 3.2 to the unit cube is for convenience only. The same result holds over any compact subset of $\mathbb{R}^d$.*

Remark 2. *The activation function used in the above results is pathological and this demonstrates that the properties of being analytic, strictly monotone, and sigmoidal may not be as significant as is often assumed. Essentially, these pathologies can be hidden even within functions that possess such desirable characteristics because powerful tools like translation and composition can still introduce them.*

Remark 3. *The activation function σ is wonderfully smooth, but unacceptably complex. Theoretical results such as these 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.*

Remark 4. *Proposition 1 in [3] provides a weaker result than Theorem 2 (in [3]), namely, the universal approximation property (for $f \in M_n$) is proved for an activation function ϕ being C^∞, strictly increasing, and sigmoidal. The proof of this proposition provides a construction, which was later, to a huge extent, used in [20]; however, the authors in [20] lost the strict monotonicity. The construction needs a fact that there exists a dense C^∞ family of functions in $C([-1,1])$.*

3.2. Weakly Monotone Continuous Activation Functions

In Guliyev's research [20, 21], the notion of *λ-monotonicity* is a relaxed version of traditional monotonicity, which is central to constructing the sigmoidal activation functions used in the approximation theorems. This concept allows for slight deviations from strict monotonicity while maintaining enough structure to support function approximation in neural networks.

Definition 3.3. *A function $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$, is said to be λ-monotone if there exists a strictly monotonic function $u : X \to \mathbb{R}$ such that:*

$$|f(x) - u(x)| \leq \lambda \quad \forall x \in X,$$

where $\lambda > 0$ is a small, positive real number that quantifies the allowable deviation from strict monotonicity.

In essence:

- For $\lambda = 0$, $f(x)$ coincides with $u(x)$ and is strictly monotonic.
- For $\lambda > 0$, $f(x)$ can exhibit small oscillations around $u(x)$, but the deviation is controlled and bounded by λ.

The Role of λ-Monotonicity in Guliyev's Papers

In Guliyev's construction of the sigmoidal activation function $\sigma(x)$, λ-monotonicity is a critical property that balances the following:

- The constructed $\sigma(x)$ is C^∞ (infinitely differentiable), which is crucial for neural network applications. Unlike strict monotonicity, λ-monotonicity allows the function to be flexible and computationally efficient.
- λ-monotonicity ensures that the activation function behaves like a monotonic function, with deviations limited by the parameter λ. This prevents erratic behavior while retaining flexibility.
- The sigmoidal function $\sigma(x)$ constructed with λ-monotonicity satisfies:

  $$h(x) < \sigma(x) < 1, \quad \forall x \in [s, +\infty),$$

  and

  $$|\sigma(x) - h(x)| \leq \lambda.$$

  Here, $h(x)$ is a strictly increasing auxiliary sigmoidal function defined as:

  $$h(x) = 1 - \frac{\min\{1/2, \lambda\}}{1 + \log(x - s + 1)}.$$

  This inequality ensures that $\sigma(x)$ closely follows $h(x)$, with λ-bounded deviations, making $\sigma(x)$ a suitable choice for neural network activation.

3.2.1. Bounded Width of NN

In scenarios where the width of the neural network is bounded, the depth must increase to maintain approximation capabilities. Guliyev's activation functions are particularly useful here because their smoothness and controlled deviation from monotonicity allow for efficient composition across layers. This means that even with a fixed number of neurons per layer, the network can still approximate complex functions by increasing the depth of the neural network.

This is consistent with the findings of Ohn and Kim [22], who showed that deep networks with general smooth activation functions can achieve minimax optimal approximation rates for Hölder continuous functions. Their results emphasize the importance of smoothness over strict monotonicity in deep architectures.

3.2.2. Bounded Depth of NN

When the depth is fixed, the network must rely on increased width to achieve approximation. In this case, the flexibility of λ-monotonic activation functions becomes crucial. Their ability to approximate strictly increasing functions with bounded error allows for constructing wide, shallow networks that still achieve good approximation performance.

Recent work by Biswas et al. [23] introduced a smooth, non-monotonic activation function (Sqish) that performs well in both standard and adversarial settings. This supports the idea that relaxing monotonicity constraints can lead to practical and effective activation functions. Similarly, Sartor et al. [24] demonstrated that constrained monotonic neural networks with saturating activations can still achieve universal approximation, further validating the theoretical foundation of λ-monotonicity.

3.2.3. Bounded Width and Depth of NN

The first paper of Guliyev and Ismailov [20] addresses the problem of approximating continuous functions using single hidden layer feedforward neural networks (SLFNs) with fixed weights. The goal is to show that such networks, with only two neurons in the hidden layer and fixed weights set to 1, can approximate any continuous univariate function in a compact interval.

Theorem 3.4. *[20] For any continuous univariate function $f(x)$ and any $\epsilon > 0$, there exists an SLFN with fixed weights such that:*

$$\left| f(x) - \sum_{i=1}^{n} e_i \sigma(w_i \cdot x + b_i) \right| < \epsilon \quad \forall x \in [a, b].$$

This holds for univariate functions, but for multivariate functions, no such approximation is generally possible with fixed weights, e.g., see the arguments in the last part of [20].

Limitation. SLFNs cannot approximate multivariate functions due to a lack of sufficient capacity to model interactions between multiple variables.

According to another paper of Guliyev and Ismailov [21], the networks characterized by a two-layer architecture and fixed weights can approximate any continuous multivariate function over compact domains. This gives them substantially greater expressive power compared to single-layer networks. The extra layer allows for the modeling of more intricate non-linear combinations of input variables, which tend to be challenging for single-layer networks to handle.

Theorem 3.5. *[21] For any continuous multivariate function $f(x_1, \ldots, x_d)$ and any $\epsilon > 0$, there exist constants $e_p, c_{pq}, \theta_{pq}, \zeta_p$ such that:*

$$\left| f(x) - \sum_{p=1}^{2d+2} e_p \sigma\left(\sum_{q=1}^{d} c_{pq} \sigma(w_q \cdot x - \theta_{pq}) - \zeta_p \right) \right| < \epsilon$$

for all $x \in [a, b]^d$.

TLFNs with fixed weights can approximate any continuous multivariate function $f(x_1, x_2, \ldots, x_d)$ to arbitrary precision on a compact domain. The network uses fixed weights as unit coordinate vectors, and a sigmoidal activation function is constructed to achieve this approximation.

Advantage. Introducing a second hidden layer improves the understanding of the network of complex interactions between various variables, thereby allowing it to approximate multivariate functions more effectively. Unlike SLFNs, which have fixed weights and are restricted to approximating univariate functions, TLFNs with their two-layer architecture and fixed weights can approximate any continuous multivariate function on compact sets. This design significantly boosts their expressive capacity in comparison to single-layer networks. The added complexity stems from the additional layer, which enables more intricate non-linear combinations of input variables — a capability that single-layer networks find difficult to achieve.

3.3. Continuous Activation Functions

When there are no constraints on the architecture of the neural network, the classical universal approximation theorem applies. It states that a feedforward neural network with a continuous, non-polynomial activation function can approximate any continuous function on a compact subset of $\mathbb{R}^n$ to arbitrary precision.

This result holds for a wide class of continuous activation functions, including:

- Sigmoidal functions (e.g., logistic, tanh)
- Smooth approximations of ReLU (e.g., softplus)
- Weakly monotonic or λ-monotonic functions [20, 21]
- General smooth activations as studied in [22]

When the depth is fixed, the network must rely on increased width to achieve universal approximation. This setting is more restrictive, but still allows for the approximation of continuous functions under certain conditions. For example, Lu et al. [25] showed that ReLU networks with fixed depth can approximate any continuous function if the width is sufficiently large.

Smooth activation functions can also be used in this setting. Zhang et al. [26] demonstrated that networks using a wide range of smooth activations (e.g., softplus, GELU, Swish) can approximate ReLU networks with only modest increases in width and depth. This implies that smooth activations retain expressive power even in shallow architectures, provided the network is wide enough.

The expressive power of deep neural networks with fixed width has been a subject of significant interest. In particular, Hanin [27] showed that ReLU networks with a width as small as $d+1$ are sufficient to arbitrarily approximate any continuous convex function on the unit cube $[0, 1]^d$. Furthermore, for general continuous functions, a width of $d + 3$ suffices.

Another specification of width bounds is based on the input and output dimensions of the networks. Johnson [28] derived the lower bound of width with uniformly continuous activations when the output dimension is one. Later, Cai [29] reached the optimal minimal width based on the input and output dimensions over all classes of activations. Recently, Rochau et al. in [30] generalized the universal approximation results of Johnson [28] to higher output dimensions and achieved the lower bound of width even tighter than the one stated by Cai [29].

3.4. Continuous Almost Everywhere Activation Functions

Leshno et al. [31] investigate the universal approximation capabilities of neural networks and analyze the criteria for single-layer networks (SLFNs) to approximate continuous functions. Their work demonstrates that single-layer feedforward neural networks (SLFNs) with a continuous activation function σ have the universal approximation property if and only if σ is not a polynomial. This finding extends previous results, such as the one by Cybenko [1], by explicitly outlining the necessary and sufficient conditions for universal approximation. Theorem 1 in [31] states:

Theorem 3.6. *Let $\sigma \in M$, where M is the set of locally bounded, piecewise continuous functions whose discontinuity points have a closure of zero Lebesgue measure. Define:*

$$E_n = span\{\sigma(w \cdot x + \theta) \,:\, w \in \mathbb{R}^n,\, \theta \in \mathbb{R}\}.$$

Then E_n is dense in $C(\mathbb{R}^n)$ if and only if σ is not an algebraic polynomial (almost everywhere).

Proof. Now we provide a sketch of the proof from [31].

1. If σ is a polynomial, E_n cannot be dense in $C(\mathbb{R}^n)$.
2. If σ is not a polynomial, E_n is dense in $C(\mathbb{R}^n)$.

The argument relies on:

- Density properties of function spaces.
- Weierstrass's theorem for polynomial approximation.

- Functional analysis to construct approximations.

STEP 1. If σ is a polynomial, E_n is not dense. Assume σ is a polynomial of degree k, i.e., $\sigma(x) = c_k x^k + \cdots + c_0$. Then $\sigma(w \cdot x + \theta)$ is also a polynomial of degree k, as it is a transformation of σ. The span of $\sigma(w \cdot x + \theta)$ will consist of polynomials of degree $\leq k$. This span cannot approximate functions that require higher complexity, such as non-polynomial or discontinuous functions. Thus, E_n cannot be dense in $C(\mathbb{R}^n)$.

STEP 2. If E_n is dense in $C(\mathbb{R})$, it is dense in $C(\mathbb{R}^n)$ Consider the space $V = \text{span}\{f(a \cdot x) : a \in \mathbb{R}^n, f \in C(\mathbb{R})\}$. Known results in functional analysis state that V is dense in $C(\mathbb{R}^n)$ (using ridge functions). Given $g \in C(\mathbb{R}^n)$, for any compact set $K \subset \mathbb{R}^n$, there exist $f \in C(\mathbb{R})$ and directions $a_1, \ldots, a_k$ such that:

$$g(x) \approx \sum_{i=1}^{k} f(a_i \cdot x).$$

If σ is dense in $C(\mathbb{R})$, it can approximate any $f(a_i \cdot x)$, making E_n dense in $C(\mathbb{R}^n)$.

STEP 3. If σ is smooth and not a polynomial, E_1 is dense in $C(\mathbb{R})$. Assume σ is C^∞, meaning it has derivatives of all orders, and it is not a polynomial. For any $g \in C(\mathbb{R})$, Weierstrass's theorem ensures that polynomials are dense in $C(\mathbb{R})$ on compact sets. The derivatives of $\sigma(w \cdot x + \theta)$ are:

$$\frac{d^k}{dw^k}\sigma(wx + \theta) = x^k \sigma^{(k)}(wx + \theta).$$

Since σ is not a polynomial, at least one derivative $\sigma^{(k)}$ is non-zero for some k. Hence, E_1 can generate polynomials of all degrees. Combining this with Weierstrass's theorem, E_1 is dense in $C(\mathbb{R})$.

STEP 4. Approximation of Continuous Functions: The argument is extended to locally bounded, piecewise continuous σ (not necessarily smooth): For any $g \in C(\mathbb{R})$ and any compact interval $[a, b]$, convolve g with a smooth function ϕ of compact support:

$$(g * \phi)(x) = \int g(y)\phi(x - y)dy.$$

The convolution $g * \phi$ is smooth, and σ can approximate it because σ is dense in C^∞ class.

STEP 5. Non-polynomiality is Necessary: Assume σ is non-polynomial. For any $g \in C(\mathbb{R})$, construct an approximation using:

$$g(x) \approx \sum_{i=1}^{m} c_i \sigma(w_i \cdot x + \theta_i).$$

If σ were polynomial, this representation would restrict g to a finite-dimensional polynomial space, which contradicts g's generality in $C(\mathbb{R})$. Thus, σ must be non-polynomial for E_n to be dense.

□

This theorem implies that the multilayer feedforward network with a non-polynomial activation function and thresholds in each neuron can approximate any continuous function on a compact domain to arbitrary precision, given enough hidden units. Polynomial activation functions constrain the network to operate within a finite-dimensional space of polynomial functions, which is insufficient for approximating the infinite-dimensional space $C(\mathbb{R}^n)$ of continuous functions. This is why the universal approximation property requires non-polynomial activation functions.

4. Interesting Facts and New Results

In this section, we highlight interesting facts that are fundamental for understanding constrains in universal approximation theorems. In addition, we present here our results that provide new lower limits of the complexity of neural networks with universal approximation properties.

4.1. Fundamental Reasons Why a Polynomial Activation Function Fails to Guarantee the Universal Approximation Property

This is due to the restricted expressive ability of polynomials. Here is the explanation:

Polynomials Are Finite-Dimensional. A polynomial of degree k is a function of the form:

$$P(x) = c_k x^k + c_{k-1} x^{k-1} + \cdots + c_0,$$

where c_i are coefficients. The space of all polynomials of degree $\leq k$ is a finite-dimensional vector space. For example, in $\mathbb{R}$, it has dimension $k+1$. If the activation function $\sigma(x)$ is a polynomial, any function generated by a neural network with this activation is a linear combination of polynomial terms of σ. This means that the network output is constrained to a space of polynomials.

Polynomials Cannot Approximate Non-Polynomial Functions. Universal approximation requires the ability to represent any continuous function on a compact domain $K \subset \mathbb{R}^n$. By the Weierstrass approximation theorem, polynomials can approximate continuous functions on compact sets. However, this property applies only if the degree of the polynomial is unbounded. If $\sigma(x)$ is a fixed polynomial, the neural network is limited to the finite-dimensional space spanned by σ and its transformations. Thus, it cannot approximate functions requiring higher complexity or non-polynomial behavior.

Restrictive Ridge Function Combinations. Neural networks with a single hidden layer approximate functions using combinations of ridge functions:

$$f(x) \approx \sum_{i=1}^{m} c_i \sigma(w_i \cdot x + \theta_i).$$

If $\sigma(x)$ is polynomial, this combination is restricted to polynomial ridge functions, which cannot form a dense set in $C(\mathbb{R}^n)$.

4.2. Main Results

Within this subsection, we would like to emphasize our new results. As we could see from the previous part of this paper, although the first Cybenko's result and its proof are quite easy and straightforward, the universal approximability can be studied from many points of view, and some of the open directions are not simple at all. We looked to some classic ([3]) and also recent results ([20, 21]) in which the simplest possible structures of feedforward neural networks are considered. Naturally, the neural network possessing a simple structure must have a complex activation function in order to provide rich approximation ability.

Recently in [7], we provided new constructions of activation functions ϑ and Θ different from those mentioned in [3]. This allowed us to preserve the universal approximation property for even simpler neural networks that it has been previously known. And we achieved this not only for C^∞ functions, but also for analytical functions. In the following, we assume the manifold M_n defined by (1).

Theorem 4.1. *Let $f \in M_n$ and let $\vartheta : \mathbb{R} \to [-1, 1]$ be a C^∞, strictly increasing, sigmoidal function. Then for any $\epsilon > 0$ there exist constants c_i, integers k_i and vectors $\bar{w}^i \in S^{d-1}$, $i = 1, 2, \ldots, 2n$, such that*

$$\left| f(\bar{x}) - \sum_{i=1}^{2n} c_i \vartheta(\bar{w}^i \cdot \bar{x} - k_i) \right| < \epsilon$$

for all $\bar{x}$ from the unit ball B^d.

Theorem 4.2. *Let $f \in M_n$ and let $\Theta : \mathbb{R} \to [-1, 1]$ be an analytic, strictly increasing, sigmoidal function. Then for any $\epsilon > 0$, there are constants c_i, integers r_i and vectors $\bar{w}^i \in S^{d-1}$, $i = 1, 2, \ldots, 2n$, satisfying*

$$\left| f(\bar{x}) - \sum_{i=1}^{2n} (c_i \Theta(\bar{w}^i \cdot \bar{x} - r_i)) \right| < \varepsilon$$

for all $\bar{x}$ from the unit ball B^d.

Those results provide an approximation property for functions of just one variable. Based on these results, we are able to provide improved results for multivariate functions as well, with the help of the Kolmogorov Representation Theorem.

Theorem 4.3. *Let $f \in C([0, 1]^d)$ and let $\Theta : \mathbb{R} \to [-1, 1]$ be an analytic, strictly increasing, sigmoidal function. Then for any $\epsilon > 0$, there exist constants d_i, c_{ij}, r_{ij}, s_i and vectors $\bar{w}^j \in \mathbb{R}^d$, $i = 1, 2, \ldots, 4d + 2$, $j = 1, 2, \ldots, 2d$, satisfying*

$$\left| f(\bar{x}) - \sum_{i=1}^{4d+2} d_i \Theta \left(\sum_{j=1}^{2d} (c_{ij} \Theta(\bar{w}^j \cdot \bar{x} + r_i) + s_i)) \right) \right| < \epsilon$$

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

It should be noted that there are not only positive news coming from our results. Maiorov et al in [3] relied on proofs which are valid in a more general setting of normed linear spaces. We could simplify the structure of neural network, but we have lost the validity for normed linear spaces.

However, there are some positive consequences. In the above, we discuss the role of λ-monotonicity in two papers by Guliyev and Ismailov ([20, 21]). They provided a nice constructive approach, with the help of monic polynomials, which led to universal approximators of both uni- and multi-variate continuous functions. The price of this was that the strict monotonicity was replaced by a weaker monotonicity, namely λ-monotonicity. With the help of our newly constructed activation function κ [7], we could preserve the nice constructive approach motivated by Guliyev and Ismailov ([20, 21]) and still have a strictly increasing activation function.

Using this function, we prove the following results.

Theorem 4.4. *Let $f \in C([-1, 1])$ and $\varepsilon > 0$. Then there exist constants $c_1, c_2, d \in \mathbb{R}$ such that*

$$|f(x) - c_1 \kappa(-x - d) - c_2 \kappa(x + d)| < \varepsilon, \quad \forall x \in [-1, 1].$$

This result implies that a neural network with a single hidden layer and only two neurons, using the activation function κ, can approximate any continuous function on a compact interval arbitrarily well. An analogous result is proved for multivariate functions.

Theorem 4.5. *Let $f \in C([-1, 1]^d)$ and $\epsilon > 0$. Then there exist constants d_i, c_{ij}, r_{ij}, s_i and vectors $\bar{w}^j \in \mathbb{R}^d$, $i = 1, 2, \ldots, 4d + 2$, $j = 1, 2, \ldots, 2d$, such that*

$$\left| f(\bar{x}) - \sum_{i=1}^{4d+2} d_i \kappa \left(\sum_{j=1}^{2d} (c_{ij} \kappa(\bar{w}^j \cdot \bar{x} + r_{ij}) + s_i)) \right) \right| < \epsilon$$

for all $\bar{x} \in [-1, 1]^d$, where the weights $\bar{w}^j$, $j = 1, \ldots, 2d$, are fixed as follows:

$$\bar{w}^1 = (1, 0, \ldots, 0), \quad \ldots, \quad \bar{w}^d = (0, 0, \ldots, 1),$$

$$\bar{w}^{d+1} = (-1, 0, \ldots, 0), \quad \ldots, \quad \bar{w}^{2d} = (0, 0, \ldots, -1).$$

The results obtained by Maiorov et al. in [3] are quite old and describe the situation for quite simple function spaces, they were further extended and transferred to more complex function spaces. It will be a part of our future research to study how our new results affect current knowledge in such spaces.

Moreover, it should be highlighted that our results are purely theoretical and, as for other purely theoretical results, we did not consider any application so far, neither our results were motivated by any application.

As for possible implementations, any reader can follow the part presented in the last section of our manuscript, where we are motivated by Guliyev and Ismailov's constructive approach, to which they provide a code as well. We do not provide any source code; any reader can follow our recommendations to prepare its own one.

5. Conclusion

In the presented study, we provided a short survey on universal approximation theorems, structured according to properties of activation functions. We also mentioned our new results using strictly increasing activation functions and improving previous constructive results that used λ-monotonicity. Those results are, in some sense, consequences of ideas improving older classic results by Maiorov et al. ([3]), in a sense that we showed that even simpler neural networks can still have the property of universal approximation.

The natural continuation can be that we will try to extend our results to other, more complicated space, since so far we dealt with compact domains in $\mathbb{R}^d$.

Acknowledgments

J. Kupka and Z. Alijani have been supported by the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583", which is co-financed by the European Union.

The work of P. Števuliáková was financially supported by the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of control, signals and systems 2 (1989) 303–314.

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

[3] V. Maiorov, A. Pinkus, Lower bounds for approximation by mlp neural networks, Neurocomputing 25 (1999) 81–91.

[4] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Current Opinion in Behavioral Sciences 18 (2017) 43–49.

[5] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing wesad, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM international conference on multimodal interaction, 2018, pp. 400–408.

[6] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, M. De Vos, Joint classification and prediction cnn framework for automatic sleep stage classification, IEEE Transactions on Biomedical Engineering 66 (2018) 1285–1296.
[7] J. Kupka, Z. Alijani, P. Števuliáková, Simple neural networks do have universal approximation property, 2025. Manuscript, submitted to Neural Networks.
[8] K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural networks 2 (1989) 183–192.
[9] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (1991) 251–257.
[10] E. K. Blum, L. K. Li, Approximation theory and feedforward networks, Neural Networks 4 (1991) 511–515.
[11] V. Kůrková, Kolmogorov's theorem and multilayer neural networks, Neural Networks 5 (1992) 501–506.
[12] G. Gripenberg, Approximation by neural networks with a bounded number of nodes at each level, Journal of approximation theory 122 (2003) 260–266.
[13] D. Yarotsky, Universal approximations of invariant maps by neural networks, Constructive Approximation 55 (2022) 407–474.
[14] Z. Lu, H. Pu, F. Wang, Z. Hu, L. Wang, The expressive power of neural networks: A view from the width, Advances in neural information processing systems 30 (2017).
[15] B. Hanin, M. Sellke, Approximating continuous functions by relu nets of minimal width, arXiv preprint arXiv:1710.11278 (2017).
[16] P. Kidger, T. Lyons, Universal approximation with deep narrow networks, in: Conference on learning theory, PMLR, 2020, pp. 2306–2327.
[17] V. Molek, Z. Alijani, Fractional concepts in neural networks: Enhancing activation functions, Pattern Recognition Letters 174 (2025) 151–158.
[18] Y. Cai, Vocabulary for universal approximation: A linguistic perspective of mapping compositions, arXiv preprint arXiv:2305.12205 (2023).
[19] A. Kratsios, The universal approximation property: Characterization, construction, representation, and existence, Annals of Mathematics and Artificial Intelligence 89 (2021) 435–469.
[20] N. J. Guliyev, V. E. Ismailov, On the approximation by single hidden layer feedforward neural networks with fixed weights, Neural Networks 98 (2018) 296–304.
[21] N. J. Guliyev, V. E. Ismailov, Approximation capability of two hidden layer feedforward neural networks with fixed weights, Neurocomputing 316 (2018) 262–269.
[22] I. Ohn, Y. Kim, Smooth function approximation by deep neural networks with general activation functions, Entropy 21 (2019) 627.
[23] K. Biswas, M. Karri, U. Bağcı, A non-monotonic smooth activation function, ArXiv (2023).
[24] D. Sartor, A. Sinigaglia, G. A. Susto, Advancing constrained monotonic neural networks: Achieving universal approximation beyond bounded activations, arXiv preprint arXiv:2505.02537 (2025).
[25] Z. Lu, H. Pu, F. Wang, Z. Hu, L. Wang, The expressive power of neural networks: A view from the width, Advances in neural information processing systems 30 (2017).
[26] S. Zhang, J. Lu, H. Zhao, Deep network approximation: Beyond relu to diverse activation functions, Journal of Machine Learning Research 25 (2024) 1–39.
[27] B. Hanin, Universal function approximation by deep neural nets with bounded width and relu activations, Mathematics 7 (2019) 992.
[28] J. Johnson, Deep, skinny neural networks are not universal approximators, ArXiv abs/1810.00393 (2018).
[29] Y. Cai, Achieve the minimum width of neural networks for universal approximation, ICLR2023 camera ready arXiv:2209.11395 (2023).
[30] D. Rochau, R. Chan, H. Gottschalk, New advances in universal approximation with neural networks of minimal width, arXiv preprint arXiv:2411.08735 (2024).
[31] M. Leshno, V. Y. Lin, A. Pinkus, S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural networks 6 (1993) 861–867.