Noncompact uniform universal approximation

van Nuland, Teun D.H.

DOI
10.1016/j.neunet.2024.106181

Publication date
2024

Document Version
Final published version

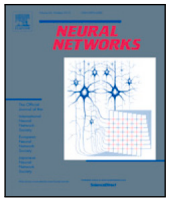
Published in
Neural Networks

Neural Networks 173 (2024) 106181

Contents lists available at ScienceDirect

Neural Networks

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



Full Length Article

Noncompact uniform universal approximation

Teun D.H. van Nuland

TU Delft, EWI/DIAM, P.O. Box 5031, 2600 GA Delft, The Netherlands

ARTICLE INFO

Keywords:
Universal approximation theorem
Feedforward ANN
Uniform convergence
Functional analysis
Ridge functions
Deep learning

ABSTRACT

The universal approximation theorem is generalised to uniform convergence on the (noncompact) input space $\mathbb{R}^n$. All continuous functions that vanish at infinity can be uniformly approximated by neural networks with one hidden layer, for all activation functions φ that are continuous, nonpolynomial, and asymptotically polynomial at $\pm\infty$. When φ is moreover bounded, we exactly determine which functions can be uniformly approximated by neural networks, with the following unexpected results. Let $\overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$ denote the vector space of functions that are uniformly approximable by neural networks with l hidden layers and n inputs. For all n and all $l \geq 2$, $\overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$ turns out to be an algebra under the pointwise product. If the left limit of φ differs from its right limit (for instance, when φ is sigmoidal) the algebra $\overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$ ($l \geq 2$) is independent of φ and l, and equals the closed span of products of sigmoids composed with one-dimensional projections. If the left limit of φ equals its right limit, $\overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$ ($l \geq 1$) equals the (real part of the) commutative resolvent algebra, a C*-algebra which is used in mathematical approaches to quantum theory. In the latter case, the algebra is independent of $l \geq 1$, whereas in the former case $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$ is strictly bigger than $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}$.

1. Introduction

Neural networks can uniformly approximate any continuous function only when the magnitude of the considered input values is bounded by a predetermined constant. Typical universal approximation theorems that use the entire noncompact input space $\mathbb{R}^n$ make use of convergence 'uniformly on compacts' (Barron, 1993; Cybenko, 1989; Hartman, Keeler, & Kowalski, 1990; Hornik, 1991; Hornik, Stinchcombe, & White, 1989; Leshno, Lin, Pinkus, & Schocken, 1993; Long, Wu, & Nan, 2007) or convergence with respect to an integral norm on $\mathbb{R}^n$ (Hornik, 1991; Kidger & Lyons, 2020). Such theorems do not rule out errors in the approximation growing exponentially (or worse) in the magnitude of the input values.

Noncompact and *uniform* approximation – which uses convergence with respect to the supremum norm over $\mathbb{R}^n$ – is a much stronger notion. In theory, it allows one to train a network up to a desired precision which is then respected by *all* input values. It also gives a more honest picture of the generalisation capability of neural networks, as we shall see later.

It is a common misconception that every continuous function on $\mathbb{R}^n$ can be uniformly approximated; in fact many commonplace continuous functions cannot.[1] The question remains: precisely *which* functions can be uniformly approximated?

Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ be continuous and nonlinear, with asymptotically linear behaviour near $\pm\infty$. One-layer neural networks are by definition linear combinations of functions of the form

$$x \mapsto \varphi(a \cdot x + b) \qquad (a \in \mathbb{R}^n, b \in \mathbb{R}), \tag{1}$$

where $\cdot$ is the standard inner product on $\mathbb{R}^n$. Such functions are constant in $n-1$ directions. If $n \geq 2$, a nonzero one-layer neural network will therefore never be in $C_0(\mathbb{R}^n)$, the space of continuous functions that vanish at infinity,[2] no matter the activation function or the amount of nodes.[3] Our first result is that, nonetheless, all functions in $C_0(\mathbb{R}^n)$ are uniformly approximable by one-layer neural networks (and therefore also by arbitrarily deep neural networks). This generalises the universal approximation theorem to a truly noncompact statement.

We also precisely characterise the space of (uniformly) approximable functions in the case that φ is moreover bounded. The above result then implies that the space of approximable functions is some vector space between $C_0(\mathbb{R}^n)$ and the space of bounded continuous functions, $C_b(\mathbb{R}^n)$.

By giving an explicit characterisation, we shall prove that this vector space is an *algebra* under the usual pointwise operations. Equivalently, products of neural networks are approximable by neural networks.

This uncovers a novel connection between neural networks and the theory of C*-algebras (J., 1990), as any norm-closed subalgebra of

E-mail address: teunvn@gmail.com.

[1] E.g., $\sin(x)$, e^x, unless the activation function is specially tailored for these.

[2] If a neural network vanishes (approximately) at infinity, it means that the network responds consistently to large inputs, like outliers.

[3] See Theorem 6.1.

https://doi.org/10.1016/j.neunet.2024.106181
Received 4 August 2023; Received in revised form 28 December 2023; Accepted 13 February 2024
Available online 15 February 2024

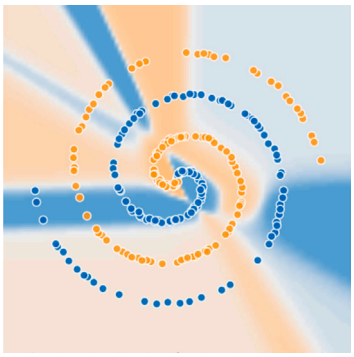


Fig. 1. Example of a neural network in which wedge functions (cf. Definition 5.7 and Fig. 3) are clearly visible in the contour plot. The network has been given insufficient nodes/layers/time to fit the data at all relevant scales, and has only succeeded on the small scale. At a slightly larger scale the wedge functions already become apparent, and this paper proves that this behaviour is in fact unavoidable at sufficiently large scale. This image was produced using https://www.matlabsolutions.com/visualize-neural-network/neural-network.html.

$C_b(\mathbb{R}^n)$ is a real C*-algebra. We do not rely on the theory of C*-algebras in this paper, but one should know that this theory initiated these findings, and might have merit for the machine learning community for reasons discussed in Hashimoto, Wang, and Matsui (2022).

Below, we discuss the explicit characterisation of the space of approximable functions, which notably does not depend greatly on the activation function φ, but only on the question whether $\varphi(-\infty) := \lim_{x\to-\infty}\varphi(x)$ equals $\varphi(\infty) := \lim_{x\to\infty}\varphi(x)$.

1.1. The case $\varphi(-\infty) = \varphi(\infty)$

We first consider the class of φ satisfying $\varphi(-\infty) = \varphi(\infty)$. We find that, for any amount of hidden layers, the vector space of approximable functions is equal to the real part of the *commutative resolvent algebra*, defined in van Nuland (2019).

In Bauer and Fulsche (2023), van Nuland (2019, 2022), van Nuland and Stienstra (2020), the commutative resolvent algebra is studied because it is the classical counterpart of the resolvent algebra, a quantum observable algebra that was introduced in Buchholz and Grundling (2007, 2008) for the purpose of (nonrelativistic) algebraic quantum field theory. This establishes a connection between machine learning and quantum algebra that seems unexplored so far, and for instance different from standard approaches to quantum neural networks (Schuld, Sinayskiy, & Petruccione, 2014). A useful application of the noncompact uniform approximation theorem to mathematical quantum physics will be demonstrated in a separate paper (Buchholz & van Nuland, 2023).

1.2. The case $\varphi(-\infty) \neq \varphi(\infty)$

Our final main theorem expresses the space of approximable functions in the case of $\varphi(-\infty) \neq \varphi(\infty)$ and gives novel insight into the approximation capability and limitations of neural networks.

When using two or more hidden layers, the space of approximable functions equals the closed span of products (with arbitrarily many factors) of sigmoids composed with one-dimensional projections. A way to visualise these products is as the wedge-shaped functions appearing in Fig. 3 and Definition 5.7, related to Voronoi diagrams (Montufar, Pascanu, Cho, & Bengio, 2014) and tropical geometry (Maragos, Charisopoulos, & Theodosis, 2021; Zhang, Naitzat, & Lim, 2018), and familiar to anyone who has ever visualised the approximation behaviour of neural networks in cases were there is a sufficiently complicated structure in the data. Indeed, when a neural network is prioritising the fitting of a small-scale structure, at a slightly larger scale one can often see the wedge functions of Definition 5.7 appearing. See, for example, Fig. 1. In fact, the rigidity of these wedge functions can prevent the neural network from converging locally if there are not enough nodes or there is not enough time. Thus, although the mathematical novelty of this paper resides at the 'infinitely large' scale, the proof in Section 5 offers an insightful perspective on the appearance of wedge shapes in general.

Opposite to the earlier case, in the present case ($\varphi(-\infty) \neq \varphi(\infty)$) there are two-layer neural networks which cannot be approximated by one-layer neural networks. We shall give a class of examples of such functions, including quite simple ones.

2. Notation and summary of main results

We let $\mathbb{N} = \{1, 2, \dots\}$. We work over the field $\mathbb{R}$. For any $n \in \mathbb{N}$, we denote by $C(\mathbb{R}^n)$, $C_b(\mathbb{R}^n)$, $C_0(\mathbb{R}^n)$, and $C_c(\mathbb{R}^n)$ respectively the continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$, the bounded ones, the ones vanishing at infinity (i.e., $\lim_{\|x\|\to\infty} f(x) = 0$), and the compactly supported ones. The support of a function f is denoted by $\operatorname{supp} f$. By $\overline{S}$ we denote the uniform closure of a set S of functions $\mathbb{R}^n \to \mathbb{R}$, i.e., the closure in the topology induced by the extended metric obtained from the supremum norm. We denote $\overline{\operatorname{span}}\, S := \overline{\operatorname{span} S}$, where $\operatorname{span} S$ is the $\mathbb{R}$-linear span of S. For $a, x \in \mathbb{R}^n$, we denote by $a \cdot x := \sum_{j=1}^n a_j x_j$ the Euclidean inner product, and define functions $p_a : \mathbb{R}^n \to \mathbb{R}$ by $p_a(x) := a \cdot x$. We let $P_V : \mathbb{R}^n \to V$ denote the orthogonal projection onto any linear subspace $V \subseteq \mathbb{R}^n$.

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function. We define the space of (feedforward) neural networks with n input nodes, one hidden layer, one output node, and activation function φ, as the following subspace of the vector space of all functions $\mathbb{R}^n \to \mathbb{R}$:

$$\mathcal{N}_\varphi^1(\mathbb{R}^n) := \operatorname{span}\left\{ x \mapsto \varphi(a \cdot x + b) \;\middle|\; a \in \mathbb{R}^n,\ b \in \mathbb{R} \right\}. \tag{2}$$

The space of networks with l hidden layers can then be defined recursively[4]:

$$\mathcal{N}_\varphi^l(\mathbb{R}^n) := \operatorname{span}\left\{ x \mapsto \varphi(f(x) + b) \middle| f \in \mathcal{N}_\varphi^{l-1}(\mathbb{R}^n),\ b \in \mathbb{R} \right\}. \tag{3}$$

The space of networks with an arbitrary amount of hidden layers is denoted by $\mathcal{N}_\varphi^\infty(\mathbb{R}^n) := \bigcup_{l=1}^\infty \mathcal{N}_\varphi^l(\mathbb{R}^n)$. Most results shall be stated for networks with only one output node, because extending these results to m output nodes, i.e., to the spaces $\mathcal{N}_\varphi^l(\mathbb{R}^n, \mathbb{R}^m) := \mathcal{N}_\varphi^l(\mathbb{R}^n)^{\oplus m}$ and $\mathcal{N}_\varphi^\infty(\mathbb{R}^n, \mathbb{R}^m) := \cup_l \mathcal{N}_\varphi^l(\mathbb{R}^n, \mathbb{R}^m)$, is straightforward.

The following theorem summarises our first main result, which is formulated for the largest possible class of activation functions in Theorem 3.7.

Theorem 2.1. *Let $n, l \in \mathbb{N}$, and let $\varphi \in C(\mathbb{R})$ be nonlinear with $\lim_{x\to\infty}(\varphi(x) - a_1 x - b_1) = 0$ and $\lim_{x\to-\infty}(\varphi(x) - a_2 x - b_2) = 0$ for certain $a_1, b_1, a_2, b_2 \in \mathbb{R}$. Then,*

$$C_0(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}.$$

The following theorem summarises our second and third main result, which are written in stronger form as Theorem 4.5 and Theorem 5.9, combined with Theorem 5.10.

Theorem 2.2. *Let $n \in \mathbb{N}$, and let $\varphi \in C(\mathbb{R})$ be nonconstant such that $\varphi(-\infty) = \lim_{x\to-\infty}\varphi(x)$ and $\varphi(\infty) = \lim_{x\to\infty}\varphi(x)$ exist and are finite.*

(1) If $\varphi(-\infty) = \varphi(\infty)$ then $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)} = \overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ and

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} = \overline{\operatorname{span}}\left\{ x \mapsto g(P(x)) \;\middle|\; \begin{array}{l} P : \mathbb{R}^n \to \mathbb{R}^k \text{ linear,} \\ g \in C_0(\mathbb{R}^k),\ k \in \mathbb{Z}_{\geq 0} \end{array} \right\}. \tag{4}$$

[4] The biases b are redundant for $l \geq 2$.

(2) If $\varphi(-\infty) \neq \varphi(\infty)$ then $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)} = \overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ and

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} = \overline{\operatorname{span}} \left\{ x \mapsto \prod_{j=1}^m \tanh(a_j \cdot x) \;\middle|\; m \in \mathbb{Z}_{\geq 0},\ a_j \in \mathbb{R}^n \right\}. \quad (5)$$

If, moreover, $n \geq 2$, then $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)} \neq \overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$.

The sigmoid tanh is used for explicitness, but can be replaced by any sigmoid of choice, as will be discussed in Section 5.

Corollary 2.3. *Let $n, m \in \mathbb{N}$ and let $\varphi \in C(\mathbb{R})$ be such that $\lim_{x\to-\infty} \varphi(x)$ and $\lim_{x\to\infty} \varphi(x)$ exist and are finite. Then the vector space $\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ is an algebra. Equivalently, pointwise products of neural networks are uniformly approximable by neural networks.*

Proof. If φ is constant, $\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ consists of constant functions, so the statement holds.

Otherwise, Theorem 2.2 allows us to consider two cases. For case 1 ($\varphi(-\infty) = \varphi(\infty)$), we will now show that the right-hand side of (4) is an algebra. For $i = 1, 2$ we fix $k_i \in \mathbb{Z}_{\geq 0}$, $g_i \in C_0(\mathbb{R}^{k_i})$ and linear maps $P_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$. We may always write

$$P_i(x) = (a_{i,1} \cdot x, \dots, a_{i,k_i} \cdot x) \qquad (x \in \mathbb{R}^n,\ i = 1, 2),$$

for vectors $a_{i,1}, \dots, a_{i,k_i} \in \mathbb{R}^n$. We define the number $k := k_1 + k_2$, define the linear function $P : \mathbb{R}^n \to \mathbb{R}^k$ by

$$P(x) := (a_{1,1} \cdot x, \dots, a_{1,k_1} \cdot x, a_{2,1} \cdot x, \dots, a_{2,k_2} \cdot x) \qquad (x \in \mathbb{R}^n),$$

and define the function $g : \mathbb{R}^k \to \mathbb{R}$ by

$$g(y_1, \dots, y_k) := g_1(y_1, \dots, y_{k_1}) g_2(y_{k_1+1}, \dots, y_k) \qquad (y \in \mathbb{R}^k).$$

It follows that $g \in C_0(\mathbb{R}^k)$, and the pointwise product of $g_1 \circ P_1$ and $g_2 \circ P_2$ evaluates to

$$(g_1 \circ P_1) \cdot (g_2 \circ P_2) = g \circ P.$$

Hence, by (4), $\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ is an algebra.

In case 2 ($\varphi(-\infty) \neq \varphi(\infty)$), the explicit characterisation of $\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ given by Theorem 2.2 is an algebra by construction. □

We remark that the number k used in the above proof is not necessarily minimal, in the sense that the representation $g \circ P$ in (4) is not unique. An in-depth analysis of the algebra in (4) is made in van Nuland (2019), including an alternative to the above proof in van Nuland (2019, Lemma 2.2(i)), cf. Section 4.

A particular aspect of Theorem 2.2 is that, for $\varphi(-\infty) \neq \varphi(\infty)$, there exist functions in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$ that are not in $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}$. In Section 5.2 we give a class of explicit examples (including a mollified AND function) and thus prove the following stronger statement.

Theorem 2.4. *Let $n \in \mathbb{N}$, $n \geq 2$, and let $\varphi \in C(\mathbb{R})$ be such that $\lim_{x\to-\infty} \varphi(x)$ and $\lim_{x\to\infty} \varphi(x)$ exist and are finite and distinct. For every $d > 0$, there are two-layer networks $f \in \mathcal{N}_\varphi^2(\mathbb{R}^n)$ with a fixed uniform distance d from the whole collection of one-layer networks, $\mathcal{N}_\varphi^1(\mathbb{R}^n)$.*

That is, no matter how hard you train the one-layer network g to approximate f, no matter the amount of nodes, there will exist an input value x such that $|f(x) - g(x)| > d$.

3. Approximation of continuous functions vanishing at infinity

The purpose of this section is to prove Theorem 3.7, which states that $C_0(\mathbb{R}^n)$ can be approximated by neural networks with one hidden layer. This result holds for a slightly larger class of activation functions than mentioned in Theorem 2.1 (in fact, the optimal class), allowing discontinuities and polynomial growth.

However, we shall first prove this approximation theorem in the simpler case that $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$. Furthermore, it will be useful to first consider $n = 1$ and $n = 2$.

Lemma 3.1. *Let $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$. We have $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} = C_0(\mathbb{R}) \oplus \mathbb{R}1$.*

Proof. We observe that any function

$$x \mapsto \varphi(ax + b) \qquad (a, b \in \mathbb{R})$$

is in $C_0(\mathbb{R})$ when $a \neq 0$. If $a = 0$, then the above map is constant, i.e., in $\mathbb{R}1$. By using (2) we find that $\mathcal{N}_\varphi^1(\mathbb{R}) \subseteq C_0(\mathbb{R}) \oplus \mathbb{R}1$, and since $C_0(\mathbb{R}) \oplus \mathbb{R}1$ is closed with respect to the supremum norm, we conclude that $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} \subseteq C_0(\mathbb{R}) \oplus \mathbb{R}1$.

The rest of the proof proceeds exactly as in the proof of Cybenko (1989, Theorem 1), replacing $C([0,1]^n)$ with $C_0(\mathbb{R})$ and replacing (Cybenko, 1989, Lemma 1) with (Hornik, 1991, Theorem 5) (quite similar to the proof of Proposition 3.5). It is however good to note that this strategy naively fails for $C_0(\mathbb{R}^n)$ when $n > 1$, as the functions $x \mapsto \varphi(a \cdot x + b)$ are not in $C_0(\mathbb{R}^n)$ when $n > 1$. □

For $n > 1$, a noncompact uniform approximation theorem requires new ideas not considered by, e.g., Cybenko (1989), Hornik (1991), Kidger and Lyons (2020).

The core idea in the case $n = 2$ is to give meaning to the formal expression

$$\int_0^{2\pi} g \circ p_{(\cos\theta, \sin\theta)}\, d\theta, \quad (6)$$

for $p_{(\cos\theta,\sin\theta)}(x, y) = x\cos\theta + y\sin\theta$, and a suitable function g such that (6) is in $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ and in a way generates $C_0(\mathbb{R}^2)$. What complicates the proof is that, whatever (6) means, it is not a Bochner integral with respect to the supremum norm. Worse yet, the integrand both has inseparable range and is discontinuous, because $\|g \circ p_{(\cos\theta,\sin\theta)} - g \circ p_{(\cos\theta',\sin\theta')}\|_\infty \geq \|g\|_\infty$ for every $\theta \neq \theta' \in [0, \pi)$. The following lemma shows that at least its pointwise interpretation is in $C_0(\mathbb{R}^2)$.

Lemma 3.2. *Let $g \in C_c(\mathbb{R})$, and define*

$$f(x, y) := \int_0^{2\pi} (g \circ p_{(\cos\theta,\sin\theta)})(x, y)\, d\theta \quad (7)$$

for all $(x, y) \in \mathbb{R}^2$. Then $f \in C_0(\mathbb{R}^2)$. If moreover $\int_{\mathbb{R}} g(x)dx = 0$, then there exists $C \in \mathbb{R}$ such that

$$|f(x, y)| \leq \frac{C}{1 + \|(x, y)\|^3} \qquad ((x, y) \in \mathbb{R}^2). \quad (8)$$

Proof. We rewrite (7) by noting that, for $(x, y) = R(\cos\varphi, \sin\varphi)$, we have $p_{(\cos\theta,\sin\theta)}(x, y) = R\sin(\theta - \varphi + \pi)$. By a substitution $\theta \mapsto \theta + \varphi - \pi$, it follows that

$$f(x, y) = \int_0^{2\pi} g(R\sin\theta)\, d\theta, \quad (9)$$

where $R = \sqrt{x^2 + y^2}$. Uniform continuity of g implies that $g(R_n \sin\theta)$ converges to $g(R\sin\theta)$ uniformly in θ whenever $R_n \to R$, hence proving continuity of f.

Let $a > 0$ be such that $\operatorname{supp} g \subseteq [-a, a]$. If $R > a$ then there exists a $\delta = \delta(R) \in [0, \pi/2)$ such that $\sin(\delta) = a/R$, implying that $R\sin(\theta) \notin \operatorname{supp} g$ whenever $|\sin\theta| > \sin\delta$. Using this δ and subsequently making the substitution $u = \sin\theta$, we find

$$\begin{aligned} &f(x, y) \\ &= \int_0^\delta g(R\sin\theta)d\theta + \int_{\pi-\delta}^{\pi+\delta} g(R\sin\theta)d\theta + \int_{2\pi-\delta}^{2\pi} g(R\sin\theta)d\theta \\ &= \int_{-\delta}^{\delta} (g(R\sin\theta)) + g(R\sin(\theta+\pi))d\theta \\ &= \int_{-a/R}^{a/R} (g(Ru) + g(-Ru)) \frac{1}{\sqrt{1-u^2}} du \\ &= \int_{-a/R}^{a/R} g(Ru) \frac{2}{\sqrt{1-u^2}} du. \end{aligned}$$

Fig. 2. First three elements of a sequence of 1-layer neural networks uniformly approximating a function in $C_0(\mathbb{R}^2)$. Cf. Lemma 3.3. To increase the locality of the limit function, the ridge functions $g\circ p_a$ need to satisfy $\int g(x)dx = 0$, unlike what is shown in the picture. Note that L^p convergence is out of the question, as each element of the sequence has infinite L^p norm, cf. Pinkus (1999, Section 7) and Theorem 6.1(2).

As $\frac{2}{\sqrt{1-u^2}} = 2 + \mathcal{O}(u^2)$ for $u \to 0$, there exists a $C > 0$ such that, for large enough R, we have $|\frac{2}{\sqrt{1-u^2}} - 2| \leq Cu^2$ for all $u \in [-a/R, a/R]$. We obtain

$$\left| f(x,y) - 2\int_{-a/R}^{a/R} g(Ru)du \right| \leq \int_{-a/R}^{a/R} |g(Ru)|Cu^2 du,$$

which by substitution $u \mapsto R^{-1}u$ becomes

$$\left| f(x,y) - 2R^{-1}\int_{-a}^{a} g(u)du \right| \leq R^{-3}\int_{-a}^{a} |g(u)|Cu^2 du.$$

Hence $f(x,y) = \mathcal{O}(R^{-1}) = \mathcal{O}(\|(x,y)\|^{-1})$, in particular $f \in C_0(\mathbb{R}^2)$. If moreover $\int g(u)du = 0$, then for large enough $R = \|(x,y)\|$ we obtain

$$|f(x,y)| \leq R^{-3}\int_{-a}^{a} |g(u)|Cu^2\, du, \tag{10}$$

which implies the lemma. □

The following result shows that, although (6) is not a Bochner integral, its pointwise interpretation is a limit of a particular sequence of Riemann sums as depicted in Fig. 2. Hence, besides being in $C_0(\mathbb{R}^2)$, the function f of Lemma 3.2 is also an element of $\overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ for any $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$.

Lemma 3.3. *Let $g \in C_c(\mathbb{R})$ be Lipschitz continuous. We define*

$$f(x,y) := \int_0^{2\pi} (g\circ p_{(\cos\theta,\sin\theta)})(x,y)\, d\theta \tag{11}$$

for all $(x,y) \in \mathbb{R}^2$ and

$$f_N := \frac{2\pi}{N}\sum_{k=0}^{N-1} g\circ p_{(\cos\frac{2\pi k}{N},\sin\frac{2\pi k}{N})},$$

for all $N \in \mathbb{N}$. We have $f_N \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ for all $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$. Moreover, the sequence $(f_{2^m})_{m\geq 1}$ converges uniformly to f.

Proof. The fact that $f_N \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ follows by noting that Lemma 3.1 implies that $g \in C_0(\mathbb{R}) \subseteq \overline{\mathcal{N}_\varphi^1(\mathbb{R})}$ and that the map $f \mapsto f\circ p_a$ is linear and bounded (with respect to the uniform norm) and sends $\mathcal{N}_\varphi^1(\mathbb{R})$ into $\mathcal{N}_\varphi^1(\mathbb{R}^2)$, and hence sends $g \in \overline{\mathcal{N}_\varphi^1(\mathbb{R})}$ to $g\circ p_a \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$.

For all $(x,y) \in \mathbb{R}^2$, we define the number

$$\Phi_N(x,y) := \#\left\{ k \in \{0,\dots,N-1\} \;\middle|\; p_{(\cos\frac{2\pi k}{N},\sin\frac{2\pi k}{N})}(x,y) \in \operatorname{supp} g \right\},$$

which leads to the bound

$$|f_N(x,y)| \leq \frac{2\pi\,\|g\|_\infty}{N}\Phi_N(x,y). \tag{12}$$

If $R[\theta] : \mathbb{R}^2 \to \mathbb{R}^2$ denotes the rotation by an angle θ, we have

$$\Phi_N(x,y) = \sum_{k=0}^{N-1} \Phi_1\big(R[\tfrac{2\pi k}{N}](x,y)\big). \tag{13}$$

By using (13) twice, we obtain, for every $M, p \in \mathbb{N}$,

$$\begin{aligned}\Phi_{2^{M+p}}(x,y) &= \sum_{k=0}^{2^{M+p}-1} \Phi_1\big(R[\tfrac{2\pi k}{2^{M+p}}](x,y)\big)\\ &= \sum_{j=0}^{2^M-1}\sum_{r=0}^{2^p-1} \Phi_1\big(R[\tfrac{2\pi(2^p j+r)}{2^{M+p}}](x,y)\big)\\ &= \sum_{r=0}^{2^p-1}\sum_{j=0}^{2^M-1} \Phi_1\big(R[\tfrac{2\pi j}{2^M}]\big(R[\tfrac{2\pi r}{2^{M+p}}](x,y)\big)\big)\\ &= \sum_{r=0}^{2^p-1} \Phi_{2^M}\big(R[\tfrac{2\pi r}{2^{M+p}}](x,y)\big).\end{aligned} \tag{14}$$

Moreover, for all $M \in \mathbb{N}$, the vectors $a_k := (\cos\frac{2\pi k}{2^M}, \sin\frac{2\pi k}{2^M})$ are pairwise linearly independent for $k = 0,\dots,2^{M-1}-1$ (and likewise for $k = 2^{M-1},\dots,2^M-1$). Any such linearly independent pair $a_k, a_{k'}$ forms a basis of $\mathbb{R}^2$ inducing a norm that is equivalent to the Euclidean norm, hence inducing a number $C_{kk'} > 0$ such that

$$\|(x,y)\| \leq C_{kk'}(|a_k\cdot(x,y)| + |a_{k'}\cdot(x,y)|) \qquad (x,y \in \mathbb{R}).$$

We obtain discs $B_{r_{kk'}}(0) \subseteq \mathbb{R}^2$ around 0 with radii $r_{kk'}$ such that $(x,y) \notin B_{r_{kk'}}(0)$ implies that $p_{a_k}(x,y) \notin \operatorname{supp} g$ or $p_{a_{k'}}(x,y) \notin \operatorname{supp} g$, and hence we obtain a radius $r = r(M) > 0$ such that for all $(x,y) \notin B_r(0)$ there is at most one $k \in \{0,\dots,2^{M-1}-1\}$ such that $p_{a_k}(x,y) \in \operatorname{supp} g$, and similarly for $k \in \{2^{M-1},\dots,2^M-1\}$. It follows that

$$\Phi_{2^M}(x,y) \leq 2 \qquad ((x,y) \in \mathbb{R}^2 \setminus B_r(0)).$$

Because the disc $B_r(0)$ is rotation invariant, it follows from (14) that $\Phi_{2^{M+p}}(x,y) \leq 2^p \cdot 2$ for all $p \in \mathbb{N}$ and (x,y) outside $B_r(0)$. By (12), we conclude that for all $M \in \mathbb{N}$ there exists an $r > 0$ such that

$$|f_{2^{M+p}}(x,y)| \leq \frac{4\pi\,\|g\|_\infty}{2^M} \qquad (p \in \mathbb{N}, (x,y) \in \mathbb{R}^2 \setminus B_r(0)).$$

As $f \in C_0(\mathbb{R}^2)$ by Lemma 3.2, it follows that for every $\epsilon > 0$ there exists an $M \in \mathbb{N}$ and an $r > 0$ such that, for all $m \geq M$ we have

$$\sup_{(x,y)\in\mathbb{R}^2\setminus B_r(0)} |f_{2^m}(x,y) - f(x,y)| < \epsilon. \tag{15}$$

Restricting to the compact $\overline{B_r(0)}$, the function

$$[0,2\pi) \to C_b(\overline{B_r(0)}), \qquad \theta \mapsto (g\circ p_{(\cos\theta,\sin\theta)})\restriction_{\overline{B_r(0)}}$$

is $\|\cdot\|_\infty$-continuous and separable valued, and hence Bochner integrable. We define simple functions $s_N : [0,2\pi) \to C_b(\overline{B_r(0)})$ by

$$s_N := \sum_{k=0}^{N-1} (g\circ p_{(\cos\frac{2\pi k}{N},\sin\frac{2\pi k}{N})})\restriction_{\overline{B_r(0)}} \cdot 1_{[\frac{2\pi k}{N},\frac{2\pi(k+1)}{N})},$$

where 1 is the indicator function. We obtain $\int_{[0,2\pi)} s_N = f_N\restriction_{\overline{B_r(0)}}$ and

$$\begin{aligned}&\|s_N(\theta) - (g\circ p_{(\cos\theta,\sin\theta)})\restriction_{\overline{B_r(0)}}\|_\infty\\ &\leq cr\Big(\Big|\cos\tfrac{2\pi k_{N,\theta}}{N} - \cos\theta\Big| + \Big|\sin\tfrac{2\pi k_{N,\theta}}{N} - \sin\theta\Big|\Big),\end{aligned} \tag{16}$$

where c is the Lipschitz constant of g and $k_{N,\theta}$ is the unique number such that $\theta \in [\frac{2\pi k_{N,\theta}}{N}, \frac{2\pi(k_{N,\theta}+1)}{N})$. As the latter interval has length $\frac{2\pi}{N}$ and cos has a Lipschitz constant of 1, we can bound (16) by $2cr\frac{2\pi}{N}$, which is independent of θ. Hence (16) converges to 0 uniformly in $\theta \in [0,2\pi)$. Therefore, by using that the Bochner integral commutes with the bounded linear map $f \mapsto f\restriction_{\overline{B_r(0)}}$, and subsequently applying the definition of the Bochner integral, we obtain

$$f\restriction_{\overline{B_r(0)}} = \int_0^{2\pi} (g\circ p_{(\cos\theta,\sin\theta)})\restriction_{\overline{B_r(0)}}\, d\theta = \lim_{N\to\infty} f_N\restriction_{\overline{B_r(0)}}$$

uniformly. Combining this with (15), we obtain the lemma. □

The importance of the following lemma can be appreciated by noting that, for all $f \in \mathcal{N}_\varphi^1(\mathbb{R}^2)$, the corresponding ψ is either constant or undefined.

Lemma 3.4. *Let $\varphi \in C_0(\mathbb{R})$, and let $g \in C_c(\mathbb{R})$ be Lipschitz continuous and satisfy $\int g(x) = 0$ and $\int g(x)x^2 \neq 0$. Let f be defined by (11). Then*

$f \in C_0(\mathbb{R}^2) \cap \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ and $f(x,y) = \mathcal{O}(\|(x,y)\|^{-3})$ for $\|(x,y)\| \to \infty$. Furthermore, the integral

$$\psi(x) = \int_{\mathbb{R}} f(x,y)\,dy$$

defines a function $\psi \in L^1(\mathbb{R}) \cap C_0(\mathbb{R})$, in fact, $\psi(x) = \mathcal{O}(|x|^{-2})$ for $|x| \to \infty$. Lastly, ψ is nonzero.

Proof. The statements $f \in C_0(\mathbb{R}^2) \cap \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$ and $f(x,y) = \mathcal{O}(\|(x,y)\|^{-3})$ follow from Lemmas 3.2 and 3.3. We moreover deduce that

$$\begin{aligned}|\psi(x)| &\leq \int_{\mathbb{R}} |f(x,y)|\,dy \leq C' \int_{\mathbb{R}} \frac{1}{(1+|x|+|y|)^3}\,dy \\ &= 2C' \int_0^\infty \frac{1}{(1+|x|+y)^3}\,dy = 2C' \int_{1+|x|}^\infty z^{-3}\,dz = C'(1+|x|)^{-2}.\end{aligned}$$

Continuity of ψ follows from (8) and the dominated convergence theorem, and hence $\psi \in L^1(\mathbb{R}) \cap C_0(\mathbb{R})$. For the last statement, we note that the proof of (10) can be sharpened by using $\frac{2}{\sqrt{1-u^2}} = 2 + u^2 + \mathcal{O}(u^4)$. Again denoting $R = \sqrt{x^2+y^2}$, we obtain

$$f(x,y) = R^{-3} \int_{-a}^{a} g(u)u^2\,du + \mathcal{O}(R^{-5}).$$

Without loss of generality, $\int g(u)u^2 > 0$. We have $R \geq |x|$, so if $|x|$ is large enough, $f(x,y) > 0$ for all $y \in \mathbb{R}$. Hence $\psi(x) > 0$ for such x. □

Proposition 3.5. *For all $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$ we have $C_0(\mathbb{R}^2) \subseteq \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$.*

Proof. With the intention of finding a contradiction, we assume that $\mathcal{N}_\varphi^1(\mathbb{R}^2) \cap C_0(\mathbb{R}^2) \neq C_0(\mathbb{R}^2)$. By the Hahn–Banach theorem, we obtain a continuous linear map $L : C_0(\mathbb{R}^2) \to \mathbb{R}$ such that $L\left(\overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)} \cap C_0(\mathbb{R}^2)\right) = \{0\}$ and $L \neq 0$. By the Riesz–Markov–Kakutani theorem, there exists a finite signed Borel measure μ on $\mathbb{R}^2$ such that

$$L(f) = \int_{\mathbb{R}^2} f(x,y)d\mu(x,y) \qquad (f \in C_0(\mathbb{R}^2)).$$

As $L \neq 0$, we obtain $\mu \neq 0$. Let g, f, ψ be as in Lemma 3.4.

For all $a \in \mathbb{R}^2$ and $b, y \in \mathbb{R}$, let $a_\perp \in \mathbb{R}^2$ be a unit vector such that $a \cdot a_\perp = 0$. Define $\tilde{f}_{a,a_\perp,b,y}(v) := f(a \cdot v + b, a_\perp \cdot v + y)$ for $v \in \mathbb{R}^2$. If $h \in \mathcal{N}_\varphi^1(\mathbb{R}^n)$, then $x \mapsto h(Ax+b)$ belongs to $\mathcal{N}_\varphi^1(\mathbb{R}^n)$ as well, for any matrix $A \in \mathbb{R}^{n \times n}$ and any $b \in \mathbb{R}^n$; this follows directly from the definition. This easily extends to the closure, and hence $\tilde{f}_{a,a_\perp,b,y} \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)} \cap C_0(\mathbb{R}^2)$. By a substitution, and (8), we find

$$\begin{aligned}\int_{\mathbb{R}^2} \psi(a \cdot v + b)\,d\mu(v) &= \int_{\mathbb{R}^2} \int_{\mathbb{R}} f(a \cdot v + b, y)\,dy\,d\mu(v) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} f(a \cdot v + b, a_\perp \cdot v + y)\,d\mu(v)\,dy \\ &= \int_{\mathbb{R}} L(\tilde{f}_{a,a_\perp,b,y})\,dy = 0. \qquad (17)\end{aligned}$$

But, since ψ is bounded and nonconstant, Hornik (1991, Theorem 5) implies that there exists no nonzero finite measure μ such that (17) holds for all $a \in \mathbb{R}^2$ and $b \in \mathbb{R}$. We obtain a contradiction, which implies the lemma. □

We now move to higher dimensions, and obtain a noncompact uniform approximation theorem in the case that $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$.

Proposition 3.6. *Let $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$, and let $n \in \mathbb{N}$. Any function in $C_0(\mathbb{R}^n)$ is uniformly approximable on $\mathbb{R}^n$ by functions of the form*

$$x \mapsto \sum_{j=1}^k c_j\,\varphi(a_j \cdot x + b_j),$$

for $k \in \mathbb{N}$, $b_j, c_j \in \mathbb{R}$, and $a_j \in \mathbb{R}^n$. In fact, for any $l \in \mathbb{N}$,

$$C_0(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}.$$

Proof. We prove the first statement of the proposition by induction on n, and notice that $n = 1$ is Lemma 3.1 and $n = 2$ is Proposition 3.5. For $n > 2$, we use $C_0(\mathbb{R}^n) = \overline{C_0(\mathbb{R}) \otimes C_0(\mathbb{R}^{n-1})}$, which follows since $C_0(\mathbb{R}) \otimes C_0(\mathbb{R}^{n-1})$ is a subalgebra of $C_0(\mathbb{R}^n)$ that vanishes nowhere and separates points, and hence its closure equals $C_0(\mathbb{R}^n)$ by the locally compact version of the Stone–Weierstrass theorem. Therefore, in order to prove the first statement of the proposition it suffices to show that any function

$$(x_1, \ldots, x_n) \mapsto f_1(x_1)f_2(x_2, \ldots, x_n)$$

can be uniformly approximated by one-layer networks, for $f_1 \in C_0(\mathbb{R})$ and $f_2 \in C_0(\mathbb{R}^{n-1})$. By induction hypothesis, f_2 can be uniformly approximated by linear combinations of functions of the form $\varphi \circ p$, for affine maps $p : \mathbb{R}^{n-1} \to \mathbb{R}$, so we may assume without loss of generality that $f_2 = \varphi \circ p$ for a fixed affine map $p : \mathbb{R}^{n-1} \to \mathbb{R}$. It then suffices to approximate the function

$$(x_1, \ldots, x_n) \mapsto f_1(x_1)(\varphi \circ p)(x_2, \ldots, x_n),$$

which equals $(f_1 \otimes \varphi) \circ P$ for the affine map $P : \mathbb{R}^n \to \mathbb{R}^2$ given by $P(x_1, \ldots, x_n) = (x_1, p(x_2, \ldots, x_n))$. The function $f_1 \otimes \varphi \in C_0(\mathbb{R}^2)$ can be approximated by one-layer networks by using Proposition 3.5, and one-layer networks composed with an affine map P are still one-layer networks, yielding the first statement of the proposition.

Let $f \in \overline{\mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)}$ and let $g \in C_0(\mathbb{R})$ equal the identity on the range of f. By Lemma 3.1, g is uniformly approximated by one-layer networks $(g_m)_{m \geq 1} \subseteq \mathcal{N}_\varphi^1(\mathbb{R})$. We write $g \circ f - g_m \circ f_m = (g \circ f - g \circ f_m) + (g \circ f_m - g_m \circ f_m)$ for a sequence $(f_m)_{m \geq 1} \subseteq \mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)$ converging uniformly to f. By the uniform continuity of g and an $\epsilon/2$ argument, we obtain $g \circ f = \lim_m g_m \circ f_m \in \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$. Hence, $f = g \circ f \in \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$. A straightforward induction argument yields the proposition. □

As a consequence, we obtain a noncompact uniform approximation theorem for all activation functions in the optimal class, which – as argued below – has Theorem 2.1 as a special case.

Theorem 3.7. *A function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfies*

$$\overline{\mathcal{N}_\varphi^1(\mathbb{R})} \cap C_0(\mathbb{R}) \neq \{0\} \qquad (18)$$

if and only if for all $n, l \in \mathbb{N}$ we have

$$C_0(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}.$$

Proof. The 'if' direction follows by taking $l = n = 1$.

For the converse direction, fix any $\chi \in \overline{\mathcal{N}_\varphi^1(\mathbb{R})} \cap C_0(\mathbb{R}) \setminus \{0\}$. By Proposition 3.6, we obtain $C_0(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\chi^l(\mathbb{R}^n)}$. It will therefore suffice to show that $\overline{\mathcal{N}_\chi^l(\mathbb{R}^n)} \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$, which we will do by induction on l. By defining the space of 0-layer neural networks as

$$\mathcal{N}_\phi^0(\mathbb{R}^n) = \{p : \mathbb{R}^n \to \mathbb{R} \ : \ p \text{ linear}\},$$

independently of ϕ, the induction basis $\overline{\mathcal{N}_\chi^0(\mathbb{R}^n)} \subseteq \overline{\mathcal{N}_\varphi^0(\mathbb{R}^n)}$ follows trivially, and, (3) holds also for $l = 1$ and can hence be applied in the induction step. Let $l \in \mathbb{N}$ and assume as an induction hypothesis that $\overline{\mathcal{N}_\chi^{l-1}(\mathbb{R}^n)} \subseteq \overline{\mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)}$. We will show that $\mathcal{N}_\chi^l(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$ by taking $g \in \mathcal{N}_\chi^l(\mathbb{R}^n)$ arbitrary. There exist $J \in \mathbb{N}$, $f_j \in \overline{\mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)}$, and $b_j, c_j \in \mathbb{R}$ $(j = 1, \ldots, J)$ such that

$$g(x) = \sum_{j=1}^J c_j\,\chi(f_j(x) + b_j).$$

We let $(\chi_m)_{m \geq 1} \subseteq \mathcal{N}_\varphi^1(\mathbb{R})$ be such that $\|\chi_m - \chi\|_\infty \to 0$, and we let $(f_{j,m})_{m \geq 1} \subseteq \mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)$ be such that $\|f_{j,m} - f_j\|_\infty \to 0$. We define the functions

$$g_m(x) := \sum_{j=1}^J c_j\,\chi_m(f_{j,m}(x) + b_j),$$

and conclude – from the uniform continuity of $\chi \in C_0(\mathbb{R})$ and an $\epsilon/2$-argument – that $\|g_m - g\|_\infty \to 0$. We may assume that χ_m is of the form

$$\chi_m(x) = \sum_{k=1}^{K_m} \tilde{c}_{k,m} \varphi(\tilde{a}_{k,m} x + \tilde{b}_{k,m}),$$

for $K_m \in \mathbb{N}$, $\tilde{a}_{k,m}, \tilde{b}_{k,m}, \tilde{c}_{k,m} \in \mathbb{R}$, which implies

$$g_m(x) = \sum_{j=1}^{J} \sum_{k=1}^{K_m} c_j \tilde{c}_{k,m} \varphi(\tilde{a}_{k,m} f_{j,m}(x) + \tilde{a}_{k,m} b_j + \tilde{b}_{k,m}),$$

and therefore $g_m \in \mathcal{N}_\varphi^l(\mathbb{R}^n)$ (conceptually, we have inserted the hidden layer of χ_m inside each node of the last hidden layer of g_m). We deduce that $g \in \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$, hence, $\mathcal{N}_\chi^l(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$, which immediately implies $\overline{\mathcal{N}_\chi^l(\mathbb{R}^n)} \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}$, the required induction step. We conclude that for all $n, l \in \mathbb{N}$ we have

$$C_0(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\chi^l(\mathbb{R}^n)} \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)},$$

completing the proof. □

We make the case that the assumption (18) is satisfied by most activation functions. First of all, we show that this assumption is implied by the assumptions of Theorem 2.1, which hold for activation functions like for instance ReLU, Leaky ReLU, ELU, GELU, Swish, Softplus, and all sigmoidal or C_0 functions.

Proof of Theorem 2.1. We define $\chi := \varphi_1 - \varphi \in C_b(\mathbb{R})$, where $\varphi_m(x) := \varphi(x - m)$. It follows that χ has finite limits at $\pm\infty$.

We now claim that χ is nonconstant, which we shall prove by contradiction.

Suppose that χ is constant, and define

$$\psi(x) := \varphi(x) - a_2 x - b_2. \tag{19}$$

Then $\lim_{x\to-\infty} \psi(x) = 0$. Furthermore, $\xi(x) := \psi(x-1) - \psi(x)$ satisfies

$$\xi(x) = \varphi(x-1) - a_2(x-1) - \varphi(x) + a_2 x = \chi(x) + a_2.$$

So ξ is constant, and moreover $\xi(x) = \lim_{x\to-\infty} \xi(x) = 0$, which implies $\psi(x) = \psi(x-1)$ for all $x \in \mathbb{R}$, i.e. ψ is periodic. Because ψ has a defined limit at $-\infty$, it must be constant. It follows from (19) that φ is linear, which contradicts our assumptions.

As χ is nonconstant, we obtain a nontrivial function

$$\chi_1 - \chi = \varphi_2 - 2\varphi_1 + \varphi \in \mathcal{N}_\varphi^1(\mathbb{R}) \cap C_0(\mathbb{R}) \setminus \{0\},$$

so the assumption of Theorem 3.7 is satisfied. □

The same argument repeated shows that continuous and nonpolynomial but asymptotically polynomial activation functions (converging to potentially different polynomials at $-\infty$ and ∞) satisfy $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} \cap C_0(\mathbb{R}) \neq \{0\}$, i.e. (18). This offers an alternative perspective on (Leshno et al., 1993) (cf. Pinkus (1999, Theorem 3.1)). The noncompact case is more subtle (e.g., there are activation functions that are 'universal' but which are not asymptotically polynomial), as exemplified by the fact that the sum of a sigmoid and *any* even function satisfies (18). Another standard argument shows that (18) is satisfied by discontinuous activation functions like the binary step function as well.

4. Bounded activation functions with identical left and right limits

In this section we precisely characterise the space of uniformly approximable functions in the easiest situation, namely in the case that $\varphi(-\infty)$ and $\varphi(\infty)$ are finite and equal.

We recall the definition of the commutative resolvent algebra from (van Nuland, 2019):

Definition 4.1. For $n \in \mathbb{N}$, the (real-valued part of the) commutative resolvent algebra on $\mathbb{R}^n$ is given by

$$C_{\mathcal{R}}(\mathbb{R}^n) = \overline{\text{span}} \left\{ g \circ P_V \;\middle|\; \begin{array}{l} V \subseteq \mathbb{R}^n \text{ linear subspace,} \\ g \in C_0(V) \end{array} \right\},$$

where P_V denotes the orthogonal projection onto the linear subspace $V \subseteq \mathbb{R}^n$.

There exists a slightly different characterisation of $C_{\mathcal{R}}(\mathbb{R}^n)$.

Lemma 4.2. *For all $n \in \mathbb{N}$ we have*

$$C_{\mathcal{R}}(\mathbb{R}^n) = \overline{\text{span}} \left\{ g \circ P \;\middle|\; \begin{array}{l} P : \mathbb{R}^n \to \mathbb{R}^k \text{ linear,} \\ g \in C_0(\mathbb{R}^k),\ k \in \mathbb{Z}_{\geq 0} \end{array} \right\}. \tag{20}$$

Proof. Each function $g \circ P$, for $P : \mathbb{R}^n \to \mathbb{R}^k$ linear and $g \in C_0(\mathbb{R}^k)$, can be written in the form

$$g \circ P = \tilde{g} \circ P_V \tag{21}$$

for $V := (\ker P)^\perp$, $P_V : \mathbb{R}^n \to V$ the corresponding orthogonal projection, and $\tilde{g} := g \circ P|_V$. As $P|_V : V \to \text{ran}\, P \subseteq \mathbb{R}^k$ is a linear isomorphism, and $g|_{\text{ran}\, P} \in C_0(\text{ran}\, P)$, it follows that $\tilde{g} \in C_0(V)$. This implies $\supseteq$ of (20) and the converse inclusion follows similarly (if not slightly easier). □

The (real part of the) commutative resolvent algebra is a closed subalgebra of $C_b(\mathbb{R}^n)$, as shown in the proof of Corollary 2.3, or, alternatively, in van Nuland (2019, Lemma 2.2). Although (van Nuland, 2019) works over the complex numbers, the above remark (over the real numbers) follows immediately by taking the real part.

Closed subalgebras of $C_b(\mathbb{R}^n)$ relate to deep neural networks in the following way.

Lemma 4.3. *Let $\varphi \in C_b(\mathbb{R})$ and $n \in \mathbb{N}$. If A is a (uniformly) closed subalgebra of $C_b(\mathbb{R}^n)$ with $\mathcal{N}_\varphi^1(\mathbb{R}^n) \subseteq A$, then $\mathcal{N}_\varphi^l(\mathbb{R}^n) \subseteq A$ for any $l \in \mathbb{N}$.*

Proof. The claim is trivial if $\varphi = 0$. Thus, assume that $\varphi \neq 0$ and note that then $\mathcal{N}_\varphi^1(\mathbb{R}^n) \subseteq A$ contains the constant functions. For $l \geq 2$, assume that $\mathcal{N}_\varphi^{l-1}(\mathbb{R}^n) \subseteq A$ and let $f \in \mathcal{N}_\varphi^{l-1}(\mathbb{R}^n)$ and $b \in \mathbb{R}$. We are to prove that $\varphi \circ (f + b) \in A$. As $f + b$ is a bounded function, the Stone–Weierstrass theorem supplies a sequence of polynomials $(p_k)_{k\geq 1}$ converging to φ uniformly on the range of $f + b$. Hence $p_k \circ (f + b)$ converges uniformly to $\varphi \circ (f + b)$. Because A is a unital algebra, $f + b \in A$ implies that

$$p_k \circ (f + b) = p_k(f + b) \in A.$$

As A is closed, we obtain $\varphi \circ (f + b) \in A$, which by induction implies that $\mathcal{N}_\varphi^l(\mathbb{R}^n) \subseteq A$ for every $l \geq 1$. □

We now turn to the case $\varphi(-\infty) = \varphi(\infty)$. Without loss of generality (because we are considering neural networks with biases b) we can assume that $\varphi(-\infty) = \varphi(\infty) = 0$, i.e., $\varphi \in C_0(\mathbb{R})$.

Lemma 4.4. *For any $\varphi \in C_0(\mathbb{R})$ and $n \in \mathbb{N}$ we have*

$$\mathcal{N}_\varphi^\infty(\mathbb{R}^n) \subseteq C_{\mathcal{R}}(\mathbb{R}^n).$$

Proof. Any network in $\mathcal{N}_\varphi^1(\mathbb{R}^n)$ is a linear combination of functions of the form

$$x \mapsto \varphi(a \cdot x + b), \tag{22}$$

for $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. By taking $P(x) := a \cdot x$ and $g(y) := \varphi(y + b)$, we find that the function (22) equals $g \circ P \in C_{\mathcal{R}}(\mathbb{R}^n)$. Hence, $\mathcal{N}_\varphi^1(\mathbb{R}^n) \subseteq C_{\mathcal{R}}(\mathbb{R}^n)$. By Lemma 4.3, this completes the proof. □

An immediate corollary of the above lemma is

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} \subseteq C_{\mathcal{R}}(\mathbb{R}^n).$$

The following theorem is the main result of this section, and states that the above inclusion is an equality. In fact, equality is already obtained with one hidden layer.

Theorem 4.5. *For any $n \in \mathbb{N}$ and any $\varphi \in C_0(\mathbb{R}) \setminus \{0\}$ we have*

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} = \overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)} = C_{\mathcal{R}}(\mathbb{R}^n).$$

Regarding systems with m output nodes, we therefore have

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n, \mathbb{R}^m)} = C_{\mathcal{R}}(\mathbb{R}^n)^{\oplus m}.$$

Proof. Let $g \circ P \in C_{\mathcal{R}}(\mathbb{R}^n)$ be a function of the form of (20), namely with $P : \mathbb{R}^n \to \mathbb{R}^k$ a linear map and $g \in C_0(\mathbb{R}^k)$. By Proposition 3.6, $g \in \overline{\operatorname{span}}\{x \mapsto \varphi(a \cdot x + b) \mid a \in \mathbb{R}^k, b \in \mathbb{R}\}$. Hence, by the continuity of the map $g \mapsto g \circ P$,

$$\begin{aligned} g \circ P &\in \overline{\operatorname{span}}\left\{x \mapsto \varphi(a \cdot P(x) + b) \mid a \in \mathbb{R}^k, b \in \mathbb{R}\right\} \\ &\subseteq \overline{\operatorname{span}}\left\{x \mapsto \varphi(\tilde{a} \cdot x + b) \mid \tilde{a} \in \mathbb{R}^n, b \in \mathbb{R}\right\}, \end{aligned}$$

where the inclusion is obtained by noting that $\mathbb{R}^n \to \mathbb{R}, x \mapsto a \cdot P(x)$ is linear, and hence given by $x \mapsto \tilde{a} \cdot x$ for an $\tilde{a} \in \mathbb{R}^n$. □

The following corollary will be used in Section 5.

Corollary 4.6. *For any $n, l \in \mathbb{N}$ and any nonconstant $\varphi \in C(\mathbb{R})$ with finite left and right limits we have*

$$C_{\mathcal{R}}(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^l(\mathbb{R}^n)}.$$

Proof. If $\varphi_1(x) := \varphi(x - 1)$ denotes the shift of φ, we have $\varphi - \varphi_1 \in C_0(\mathbb{R}) \setminus \{0\}$. Furthermore, we see that $\mathcal{N}_{\varphi-\varphi_1}^l(\mathbb{R}^n) \subseteq \mathcal{N}_\varphi^l(\mathbb{R}^n)$ for any number of hidden layers l. Hence, the result follows from Theorem 4.5. □

5. Bounded activation functions with distinct left and right limits

For this section, we let φ be continuous with finite and distinct left and right limits. In this case, describing the set of uniformly approximable functions is slightly more involved.

Definition 5.1. We define

$$\mathfrak{S}(\mathbb{R}) := \left\{ f \in C(\mathbb{R}) \;\middle|\; \lim_{x \to -\infty} f(x) \text{ and } \lim_{x \to \infty} f(x) \text{ exist in } \mathbb{R} \right\}.$$

More generally, for $n \in \mathbb{N}$, we define

$$\mathfrak{S}(\mathbb{R}^n) := \overline{\operatorname{span}} \left\{ \prod_{j=1}^m (g_j \circ p_{a_j}) \;\middle|\; m \in \mathbb{N}, g_j \in \mathfrak{S}(\mathbb{R}), a_j \in \mathbb{R}^n \right\},$$

where we recall that $p_a(x) = a \cdot x$.

We note that $\mathfrak{S}(\mathbb{R}^n)$ is a closed subalgebra of $C_{\mathrm{b}}(\mathbb{R}^n)$. We may give a more explicit characterisation of $\mathfrak{S}(\mathbb{R}^n)$.

Lemma 5.2. *For all $n \in \mathbb{N}$ we have*

$$\mathfrak{S}(\mathbb{R}^n) = \overline{\operatorname{span}} \left\{ \prod_{j=1}^m (\tanh \circ p_{a_j}) \;\middle|\; m \in \mathbb{Z}_{\geq 0}, a_j \in \mathbb{R}^n \right\}.$$

In the above formula, tanh *can be replaced by any strictly monotonous element of $\mathfrak{S}(\mathbb{R})$.*

Remark 5.3. The above remarks can be formulated in C*-algebraic language quite concisely. Namely, for any fixed strictly monotonous $\sigma \in \mathfrak{S}(\mathbb{R})$, $\mathfrak{S}(\mathbb{R}^n)$ is the smallest C*-subalgebra of the real C*-algebra $C_{\mathrm{b}}(\mathbb{R}^n)$ that contains the functions 1 and $\sigma \circ p_a$ ($a \in \mathbb{R}^n$).

Proof of Lemma 5.2. The inclusion $\supseteq$ follows by taking $g_j = \tanh$. The right-hand side is therefore a subalgebra of $\mathfrak{S}(\mathbb{R}^n)$. Because composition with p_a is a continuous mapping, it thus suffices to show that, for all $g \in \mathfrak{S}(\mathbb{R})$,

$$g \in \overline{\operatorname{span}} \left\{ \prod_{j=1}^m \sigma \;\middle|\; m \in \mathbb{Z}_{\geq 0} \right\},$$

for a fixed strictly monotonous element $\sigma \in \mathfrak{S}(\mathbb{R})$ such as $\sigma = \tanh$. The above set contains all limits of all polynomials in σ, and hence, by Stone–Weierstrass, it contains $f \circ \sigma$ for every continuous function $f \in C(\overline{\operatorname{ran}\sigma})$. It therefore also contains $g = (g \circ \sigma^{-1}) \circ \sigma$, as required. □

The algebra $\mathfrak{S}(\mathbb{R})$ is closely related to the space of one-layer neural networks.

Lemma 5.4. *For any $\varphi \in \mathfrak{S}(\mathbb{R})$ we have $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} \subseteq \mathfrak{S}(\mathbb{R})$. If moreover $\varphi(-\infty) \neq \varphi(\infty)$, then we have $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} = \mathfrak{S}(\mathbb{R})$.*

Proof. For any $\varphi \in \mathfrak{S}(\mathbb{R})$, the space $\mathcal{N}_\varphi^1(\mathbb{R})$ is spanned by functions

$$x \mapsto \varphi(ax + b) \qquad (a, b \in \mathbb{R}),$$

which are also functions in $\mathfrak{S}(\mathbb{R})$. Since $\mathfrak{S}(\mathbb{R})$ is a closed linear space, we obtain $\overline{\mathcal{N}_\varphi^1(\mathbb{R})} \subseteq \mathfrak{S}(\mathbb{R})$.

If we furthermore assume that $\varphi(-\infty) \neq \varphi(\infty)$, then for every $f \in \mathfrak{S}(\mathbb{R})$ there exist $a, b \in \mathbb{R}$ such that $f - a\varphi - b \in C_0(\mathbb{R})$. Hence, by using Theorem 2.1,

$$f \in a\varphi + b + C_0(\mathbb{R}) \subseteq \overline{\mathcal{N}_\varphi^1(\mathbb{R})}.$$

The combination of both inclusions finishes the proof. □

One of the two desired inclusions is now derived as follows.

Proposition 5.5. *For every $n \in \mathbb{N}$ and every $\varphi \in \mathfrak{S}(\mathbb{R})$, we have*

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} \subseteq \mathfrak{S}(\mathbb{R}^n).$$

Proof. The space $\mathcal{N}_\varphi^1(\mathbb{R}^n)$ is spanned by functions of the form $f = g \circ p_a$ for $g \in \mathcal{N}_\varphi^1(\mathbb{R})$ and $a \in \mathbb{R}^n$. By Lemma 5.4, we have $g \in \mathfrak{S}(\mathbb{R})$, which implies that $f \in \mathfrak{S}(\mathbb{R}^n)$. Therefore, $\mathcal{N}_\varphi^1(\mathbb{R}^n) \subseteq \mathfrak{S}(\mathbb{R}^n)$, and since $\mathfrak{S}(\mathbb{R}^n)$ is a closed subalgebra of $C_{\mathrm{b}}(\mathbb{R}^n)$, Lemma 4.3 implies the proposition. □

We proceed with the converse inclusion, in order to obtain equality of the spaces $\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)}$ and $\mathfrak{S}(\mathbb{R}^n)$.

5.1. Converse inclusion

Denote by $\mathfrak{S}_{\mathrm{c}}(\mathbb{R})$ the set of functions $f \in \mathfrak{S}(\mathbb{R})$ such that $f(\mathbb{R}) = [0, 1]$ and $f^{-1}((0, 1))$ is bounded.

Lemma 5.6. *The span of $\mathfrak{S}_{\mathrm{c}}(\mathbb{R})$ is dense in $\mathfrak{S}(\mathbb{R})$.*

Proof. Our definitions imply that $C_0(\mathbb{R}) + \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}) \subseteq \mathfrak{S}(\mathbb{R})$. Let $f \in \mathfrak{S}(\mathbb{R})$ be arbitrary. If $f(-\infty) = f(\infty)$ then $f - f(\infty)v \in C_0(\mathbb{R})$, where $v \in \mathfrak{S}_{\mathrm{c}}(\mathbb{R})$ is defined as $v(x) := \min(1, |x|)$. It follows that $f \in C_0(\mathbb{R}) + \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R})$.

Define $v_1, v_2 \in \mathfrak{S}_{\mathrm{c}}(\mathbb{R})$ by

$$v_1(x) = \max\{0, \min\{1, x\}\} \qquad \text{and} \qquad v_2(x) = v_1(-x),$$

so that $v_1(\infty) = 1 = v_2(-\infty)$ and $v_1(-\infty) = 0 = v_2(\infty)$. Then, setting $g := f - f(\infty)v_1 - f(-\infty)v_2$, we have $g \in C_0(\mathbb{R})$ and

$$f = g + f(\infty)v_1 + f(-\infty)v_2 \in C_0(\mathbb{R}) + \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}).$$

Therefore,

$$C_0(\mathbb{R}) + \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}) = \mathfrak{S}(\mathbb{R}). \tag{23}$$

For an arbitrary $f \in C_{\mathrm{c}}(\mathbb{R})$, decompose $f = f_+ - f_-$ for the positive and negative parts $f_+, f_- \geq 0$ of f. If $f_\pm \neq 0$, we have $\frac{f_\pm}{\|f_\pm\|_\infty} \in \mathfrak{S}_{\mathrm{c}}(\mathbb{R})$, which implies $f_\pm \in \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R})$. Hence,

$$C_{\mathrm{c}}(\mathbb{R}) \subseteq \operatorname{span}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}). \tag{24}$$

By taking closures of (23) and (24), we obtain $\overline{\operatorname{span}}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}) = C_0(\mathbb{R}) + \overline{\operatorname{span}}\mathfrak{S}_{\mathrm{c}}(\mathbb{R}) = \mathfrak{S}(\mathbb{R})$, as claimed. □

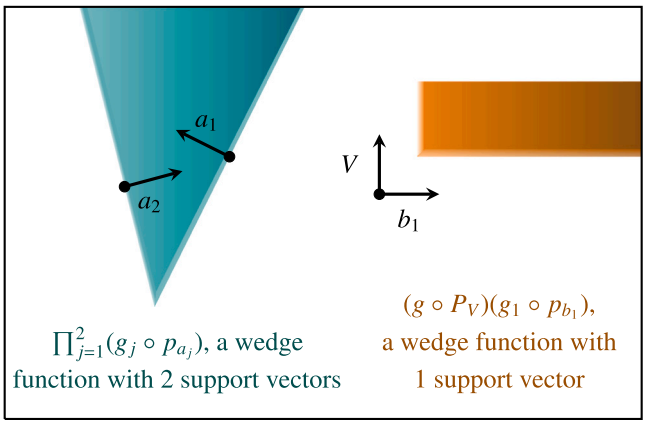


Fig. 3. Contour plot of two wedge functions on $\mathbb{R}^2$.

Definition 5.7. Let J be a finite set. A **wedge function** with **support vectors** $\{a_j\}_{j\in J} \subseteq \mathbb{R}^n$ is a function of the form

$$(g\circ P_V)\prod_{j\in J}(g_j\circ p_{a_j}),$$

for some linear subspace $V \subseteq \mathbb{R}^n$, $g \in C_c(V)$ and $g_j \in \mathfrak{S}_c(\mathbb{R})$ for each $j \in J$.

The following proposition is the key to Theorem 5.9.

Proposition 5.8. *Fix $\varphi \in \mathfrak{S}(\mathbb{R})$ with $\varphi(-\infty) \neq \varphi(\infty)$ and let $m \in \mathbb{N}$. Any wedge function with m support vectors is in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$.*

Proof. The proof proceeds by induction on m. The induction basis follows immediately, since wedge functions with no support vectors are in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$ by Corollary 4.6.

Let J be an index set of support vectors $\{a_j\}_{j\in J}$ such that all wedge functions with strictly less support vectors are in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$ – this is the induction hypothesis. To obtain the induction step, we shall fix a generic wedge function $(g\circ P_V)\prod_{j\in J}(g_j\circ p_{a_j})$ and prove that it is in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$. That is, we fix a linear subspace $V \subseteq \mathbb{R}^n$, functions $g \in C_c(V)$ and $g_j \in \mathfrak{S}_c(\mathbb{R})$ and support vectors $a_j \in \mathbb{R}^n$ for all $j \in J$. Without loss of generality, we may assume that $0 \leq g \leq 1$. We define a function $h \in C_b(\mathbb{R})$ by

$$h(x) := \begin{cases} 0 & \text{if} \quad x \leq 0, \\ x & \text{if} \quad 0 \leq x \leq 1, \\ 1 & \text{if} \quad 1 \leq x, \end{cases}$$

as well as the shifts $h_m(x) := h(x-m)$. For every $I \subseteq J$, we recursively define

$$\begin{aligned} f_\emptyset &:= 0, \\ f_I &:= (g\circ P_V)\prod_{j\in I}(g_j\circ p_{a_j}) - h_{\#I}\circ\Big(g\circ P_V + \sum_{j\in I}(g_j\circ p_{a_j})\Big) \\ &\quad - \sum_{\substack{H\subseteq I\\ H\neq I}} f_H \prod_{j\in I\setminus H}(g_j\circ p_{a_j}). \end{aligned} \tag{25}$$

We note that the two definitions are consistent by taking $I = \emptyset$ and using $g\circ P_V - h_0\circ(g\circ P_V) = 0$. We also define the space $W_I := V + \operatorname{span}\{a_j \mid j \in I\}$. By induction on $\#I$, and using that $p_{a_j} = p_{a_j}\circ P_{W_I}$ for every $j \in I$, we find that $f_I = f_I \circ P_{W_I}$. Moreover, we claim that $f_I\restriction_{W_I} \in C_c(W_I)$. The latter claim is shown by proving the following statement by induction (i.e., we are using nested induction). Here, $x \in \mathbb{R}^n$ is fixed.

$$\alpha(I): \quad \begin{array}{l} \text{“If } g(P_V(x)) = 0 \\ \text{or } g_{j_0}(a_{j_0}\cdot x) \in \{0,1\} \text{ for some } j_0 \in I, \\ \text{then } f_I(x) = 0.\text{"} \end{array}$$

We prove $\alpha(I)$ by induction on $\#I$. By definition, $\alpha(\emptyset)$ is true. Now suppose $\alpha(H)$ is true for all $H \subsetneq I$. Then $\alpha(I)$ follows from the following three statements:

- If $g(P_V(x)) = 0$, then $f_H(x) = 0$, so (25) gives $f_I(x) = 0$.
- If $g_{j_0}(a_{j_0}\cdot x) = 0$ for a certain $j_0 \in I$, then the first and second term of (25) drop out, leaving us with
$$f_I(x) = -\sum_{\substack{H\subseteq I\\ H\neq I}} f_H(x)\prod_{j\in I\setminus H} g_j(a_j\cdot x).$$
If $j_0 \in H$, then we have $f_H(x) = 0$, and if $j_0 \notin H$, then $\prod_{j\in I\setminus H} g_j(a_j\cdot x) = 0$. Hence, in both cases, $f_I(x) = 0$.
- If $g_{j_0}(a_{j_0}\cdot x) = 1$ for a certain $j_0 \in I$, then
$$\begin{aligned} f_I(x) =& g(P_V(x))\prod_{j\in I\setminus\{j_0\}} g_j(a_j\cdot x) \\ &- h_{\#I-1}\Big(g(P_V(x)) + \sum_{j\in I\setminus\{j_0\}} g_j(a_j\cdot x)\Big) \\ &- \sum_{\substack{H\subseteq I\\ H\neq I}} f_H(x)\prod_{j\in(I\setminus\{j_0\})\setminus H} g_j(a_j\cdot x). \end{aligned}$$
For all $H \subsetneq I$ with $j_0 \in H$ we have $f_H(x) = 0$, so the third term becomes
$$\begin{aligned} &-\sum_{\substack{H\subseteq I\\ H\neq I}} f_H(x)\prod_{j\in(I\setminus\{j_0\})\setminus H} g_j(a_j\cdot x) \\ &= -\sum_{\substack{H\subseteq I\setminus\{j_0\}\\ H\neq I\setminus\{j_0\}}} f_H(x)\prod_{j\in I\setminus H} g_j(a_j\cdot x) - f_{I\setminus\{j_0\}}(x). \end{aligned}$$
We conclude that $f_I(x) = f_{I\setminus\{j_0\}}(x) - f_{I\setminus\{j_0\}}(x) = 0$.

Therefore $\alpha(H)$ is true for every $H \subseteq J$. We will now deduce that $f_H\restriction_{W_H}$ has compact support. The assertion $\alpha(H)$ shows that for $x \in W_H$ with $f_H(x) \neq 0$ we have $P_V(x) \in \operatorname{supp} g$ and $a_j\cdot x \in \overline{g_j^{-1}((0,1))}$ for $j \in H$, which by compactness of $\operatorname{supp} g$ and $\overline{g_j^{-1}((0,1))}$ implies that there exists $R > 0$ independent of x such that $\|P_V(x)\| \leq R$ and $|a_j\cdot x| \leq R$. It is not hard to see that

$$\|x\|_* := \|P_V(x)\| + \sum_{j\in H}|a_j\cdot x| \qquad (x \in W_H),$$

defines a seminorm, and also positive definiteness is proven as follows. If $\|x\|_* = 0$, then $x \perp V$ and $x \perp a_j$ for all $j \in H$. Since $x \in W_H = V + \operatorname{span}\{a_j \mid j \in I\}$, this implies $x \perp x$ and hence $x = 0$. Now, the considerations from above imply that the set $\{x \in W_H \mid f_H(x) \neq 0\}$ is bounded with respect to the norm $\|\cdot\|_*$ (hence bounded with respect to any norm on W_H), which implies that $f_H\restriction_{W_H}$ has compact support. We have noted earlier that $f_H = f_H\circ P_{W_H}$, so we conclude that

$$f_H = f_H\restriction_{W_H}\circ P_{W_H} \quad \text{and} \quad f_H\restriction_{W_H} \in C_c(W_H), \tag{26}$$

in particular, f_H is a wedge function without support vectors.

By rearranging (25) in the case $I = J$, we obtain

$$\begin{aligned} (g\circ P_V)\prod_{j\in J}(g_j\circ p_{a_j}) =& \sum_{\emptyset\neq H\subseteq J} f_H\prod_{j\in J\setminus H}(g_j\circ p_{a_j}) \\ &+ h_{\#J}\circ\Big(g\circ P_V + \sum_{j\in J}(g_j\circ p_{a_j})\Big). \end{aligned} \tag{27}$$

We can summarise (26) and (27) by saying that any wedge function can be written as a sum of wedge functions with strictly less support

vectors (that are therefore in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$ by the induction hypothesis) plus a remainder term which we denote by

$$r := h_{\#J} \circ \Big(g \circ P_V + \sum_{j \in J} (g_j \circ p_{a_j}) \Big).$$

We now show that $r \in \overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$. Corollary 4.6 implies that $g \circ P_V \in C_{\mathcal{R}}(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}$. Moreover, Lemma 5.4 shows that $g_j \in \mathfrak{S}_c(\mathbb{R}) \subseteq \overline{\mathcal{N}_\varphi^1(\mathbb{R})}$ and hence $g_j \circ p_{a_j} \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}$. Therefore,

$$g \circ P_V + \sum_{j \in J} (g_j \circ p_{a_j}) \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}.$$

By Lemma 5.4, $h_{\#J} \in \overline{\mathcal{N}_\varphi^1(\mathbb{R})}$. By uniform continuity of elements in $\mathcal{N}_\varphi^1(\mathbb{R})$, it follows that $r \in \overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$.

Hence, the function in (27) is in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$, which completes the induction step initiated at the beginning of this proof. We conclude that all wedge functions are in $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$.

Theorem 5.9. *Let $\varphi \in C(\mathbb{R})$ be such that $\lim_{x \to -\infty} \varphi(x)$ and $\lim_{x \to \infty} \varphi(x)$ are finite and unequal. We have*

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n)} = \overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)} = \mathfrak{S}(\mathbb{R}^n).$$

Regarding systems with m output nodes, we therefore have

$$\overline{\mathcal{N}_\varphi^\infty(\mathbb{R}^n, \mathbb{R}^m)} = \mathfrak{S}(\mathbb{R}^n)^{\oplus m}.$$

Proof. By Proposition 5.8 we know that all wedge functions belong to $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$. In particular, taking $V = \{0\}$, wedge functions of the form $\prod_{j \in J} (g_j \circ p_{a_j})$ belong to $\overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$. By using Lemma 5.6, this implies that $\mathfrak{S}(\mathbb{R}^n) \subseteq \overline{\mathcal{N}_\varphi^2(\mathbb{R}^n)}$. Combining the latter inclusion with the inclusion from Proposition 5.5, we obtain equality. □

5.2. Difference between one-layer and two-layer networks

Let $\varphi \in \mathfrak{S}(\mathbb{R})$ with $\varphi(-\infty) \neq \varphi(\infty)$. Let $h \in \mathfrak{S}(\mathbb{R})$ be a continuous function satisfying $h(x) = 0$ for $x \leq 0$ and $h(x) = 1$ for $x \geq 1$. Then

$$f(x, y) := h(x)h(y) \tag{28}$$

(a mollified AND function) is approximable by two-layer neural networks by Theorem 5.9. However, the following reasoning shows that it is at least a (supremum norm) distance of 1/4 from all one-layer neural networks.

Theorem 5.10. *Let $\varphi \in \mathfrak{S}(\mathbb{R})$ with $\varphi(-\infty) \neq \varphi(\infty)$, and define $f \in \overline{\mathcal{N}_\varphi^2(\mathbb{R}^2)}$ by (28). Then*

$$\|f - g\|_\infty \geq \frac{1}{4}$$

for all $g \in \mathcal{N}_\varphi^1(\mathbb{R}^2)$.

Proof. For any function $g \in C(\mathbb{R}^2)$ we define

$$l_g(v) := \lim_{t \to \infty} \big(g(tv) + g(-tv) \big),$$

for the $v \in S^1 \subseteq \mathbb{R}^2$ for which this limit exists.

If $g(x) = c\varphi(a \cdot x + b)$, then l_g will be defined on the full unit circle S^1, and constant almost everywhere, namely on all $v \in S^1$ satisfying $a \cdot v \neq 0$ (or everywhere, if $a = 0$). Taking a sum, linearity of limits implies that, for all $g \in \mathcal{N}_\varphi^1(\mathbb{R}^2)$, the partial function l_g is defined everywhere and constant almost everywhere on S^1. (In fact, this holds for all $g \in \overline{\mathcal{N}_\varphi^1(\mathbb{R}^2)}$.) However,

$$l_f((\cos\theta, \sin\theta)) = \begin{cases} 1 & \theta \in (0, \frac{1}{2}\pi) \cup (\pi, \frac{3}{2}\pi) \\ 0 & \theta \in (\frac{1}{2}\pi, \pi) \cup (\frac{3}{2}\pi, 2\pi), \end{cases}$$

which is manifestly not constant almost everywhere. Hence, there exists a $v \in S^1$ with

$$\Big| \lim_{t \to \infty} g(tv) - \lim_{t \to \infty} f(tv) \Big| \geq \frac{1}{4},$$

which implies that $\|g - f\|_\infty \geq \frac{1}{4}$. □

The above reasoning for showing $f \notin \overline{\mathcal{N}_\varphi^1(\mathbb{R}^n)}$ will work for all $f \in C(\mathbb{R}^n)$ with l_f not constant almost everywhere on S^{n-1}, hence supplying a large list of examples of functions not uniformly approximable by one-layer neural networks. By scaling, the uniform distance can be made arbitrary large, so there are functions in $\mathcal{N}_\varphi^2(\mathbb{R}^n)$ with arbitrarily large distance from $\mathcal{N}_\varphi^1(\mathbb{R}^n)$, proving Theorem 2.4.

6. Nonzero one-layer networks do not vanish at infinity

As an encore, we prove a claim made in the introduction. We also prove a claim made in the caption of Fig. 2, which is stated without proof in Pinkus (1999, Section 7). An elegant proof of the latter (i.e. the proof of point 2 below) was supplied by a very generous anonymous referee.

Theorem 6.1. *Let $m \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}$ be numbers, $a_1, \ldots, a_m \in \mathbb{R}^n$ be vectors, and $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ be functions. Define $f(x) := \sum_{j=1}^m f_j(a_j \cdot x)$.*

1. *If $f \in C_0(\mathbb{R}^n)$ then $f = 0$. In particular,*

   $$\mathcal{N}_\varphi^1(\mathbb{R}^n) \cap C_0(\mathbb{R}^n) = \{0\},$$

   for every function $\varphi : \mathbb{R} \to \mathbb{R}$.
2. *If $f \in L^p(\mathbb{R}^n)$ for some $p \in (0, \infty)$, then $f = 0$ almost everywhere. In particular,*

   $$\mathcal{N}_\varphi^1(\mathbb{R}^n) \cap L^p(\mathbb{R}^n) = \{0\},$$

   for every function $\varphi : \mathbb{R} \to \mathbb{R}$ and every $p \in (0, \infty)$.

Proof. Since $n \geq 2$, we can choose for each $j \in \{1, \ldots, m\}$ a nonzero vector $v_j \in \mathbb{R}^n$ satisfying

$$a_j \cdot v_j = 0 \quad \text{and} \quad \|v_j\| > \|v_1\| + \ldots + \|v_{j-1}\|.$$

For brevity, write $[m] := \{1, \ldots, m\}$. For $I \subseteq [m]$, define

$$v(I) := \sum_{i \in I} v_i, \quad \text{in particular,} \quad v(\emptyset) = 0.$$

By choice of the v_j we have that $v(I) \neq 0$ for all $\emptyset \neq I \subseteq [m]$.

For all $j \in [m]$, $x \in \mathbb{R}^n$, and $t \in \mathbb{R}$, we note that, since $a_j \cdot v_j = 0$,

$$\sum_{I \subseteq [m], j \notin I} (-1)^{\#I} f_j(a_j \cdot (x + tv(I))) = \sum_{I \subseteq [m], j \in I} (-1)^{\#I - 1} f_j(a_j \cdot (x + tv(I))),$$

and hence,

$$\begin{aligned} \sum_{I \subseteq [m]} (-1)^{\#I} f_j(a_j \cdot (x + tv(I))) &= \sum_{I \subseteq [m], j \in I} (-1)^{\#I} f_j(a_j \cdot (x + tv(I))) \\ &\quad + \sum_{I \subseteq [m], j \notin I} (-1)^{\#I} f_j(a_j \cdot (x + tv(I))) \\ &= 0. \end{aligned}$$

By summing the above over j and interchanging the sums, we see $\sum_{I \subseteq [m]} (-1)^{\#I} f(x + tv(I)) = 0$. Since the summand for $I = \emptyset$ is simply $f(x)$, we thus get

$$f(x) = -\sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I} f(x + tv(I)) \qquad (x \in \mathbb{R}^n, t \in \mathbb{R}). \tag{29}$$

The above formula allows us to prove both 1 and 2.

1. If $f \in C_0(\mathbb{R}^n)$ then for every $x \in \mathbb{R}^n$ and every $\epsilon > 0$ there exists a $t \in \mathbb{R}$ large enough such that the right-hand side of 29 is smaller than ϵ, hence $|f(x)| < \epsilon$ for every x and every ϵ, implying $f = 0$.

2. If $f \in L^p(\mathbb{R}^n)$ then $g = |f|^p \in L^1(\mathbb{R}^n)$. For completeness, let us prove a folklore assertion related to the Poisson summation formula, namely the claim that for every $g \in L^1(\mathbb{R}^n)$ and any lattice $\Lambda \subseteq \mathbb{R}^n$, the series

$$h(x) := \sum_{k\in\Lambda} g(x+k)$$

converges for almost all $x \in \mathbb{R}^n$. Indeed, if $U \subseteq \mathbb{R}^n$ is a measurable subset such that $\mathbb{R}^n = \sqcup_{k\in\Lambda}(U+k)$, then

$$\infty > \int_{\mathbb{R}^n} |g(x)| = \int_U \sum_{k\in\Lambda} |g(x+k)|\, dx\,.$$

So there exists a null-set $N_0 \subseteq U$ such that $h(x) < \infty$ for $x \in U \setminus N_0$. As $h(x+l) = h(x)$ for every $l \in \Lambda$, we have for the null-set $N_\Lambda := N_0 + \Lambda$ that $h(x) < \infty$ for every $x \in \mathbb{R}^n \setminus N_\Lambda$. For every non-empty $I \subseteq [m]$, $\Lambda := \mathbb{Z}v(I)$ is a nontrivial lattice and so the above statement in particular supplies a null-set $N_I \subseteq \mathbb{R}^n$ such that

$$\lim_{l\in\mathbb{N}, l\to\infty} g(x+lv(I)) = 0 \qquad (x \in \mathbb{R}^n \setminus N_I).$$

Hence,

$$\lim_{l\in\mathbb{N}, l\to\infty} f(x+lv(I)) = 0 \qquad (x \in \mathbb{R}^n \setminus N_I).$$

Now, $N := \bigcup_{\emptyset\neq I\subseteq[m]} N_I$ is a null-set, and 29 shows for $x \in \mathbb{R}^n \setminus N$ that

$$f(x) = -\sum_{\emptyset\neq I\subseteq[m]} (-1)^{\#I} f(x+lv(I)) \to 0 \quad (l \in \mathbb{N}, l\to\infty).$$

Thus, $f = 0$ almost everywhere. □

7. Open questions

As practitioners are asking for mathematically founded ways to choose the right architecture for their specific problems, there is still much to learn regarding the influence of the number of neurons, the width, the depth, et cetera, on the approximation capabilities of neural networks. This paper indicates that the uniform topology on $\mathbb{R}^n$ is apt to gain new insights. Yet, there remains a lot of unexplored terrain.

Density theorems and analytic bounds expressing the quality of the optimal approximation in terms of the width and depth of the neural network, as well as the regularity of the activation function, have thus far been developed with respect to the topology of compact convergence and with respect to L^p-convergence (Gripenberg, 2003; Shen, Yang, & Zhang, 2021, 2022; Yarotsky, 2021). The spaces $C(\mathbb{R}^n)$ and $L^p(\mathbb{R}^n)$ naturally form the arenas of functions for which those bounds can be sought. By the present results, the spaces $C_0(\mathbb{R}^n)$, $C_{\mathcal{R}}(\mathbb{R}^n)$, and $\mathfrak{S}(\mathbb{R}^n)$ are the analogous arenas in which one should search for similar bounds and density theorems for global uniform convergence.

One concrete and fascinating question would be whether there exists an activation function such that every element of $C_0(\mathbb{R}^n)$ can be globally uniformly approximated up to arbitrary accuracy with a fixed width and depth dependent only on n, i.e., whether there exist globally superexpressive activations.

It would also be very informative to get theoretical bounds on the number of nodes that are needed to obtain a certain uniform precision (possibly constraining the amount of nodes per layer (Gripenberg, 2003; Kidger & Lyons, 2020) and also the amount of nodes in totality (Ismailov, 2021, Section 5)).

The same questions can be asked for specific classes of neural networks, such as convolutional neural networks or finite impulse recurrent neural networks, as then the 'upper bound' of Proposition 5.5 still holds. For instance, if a certain class of convolutional neural networks can be expressed as feedforward ANNs (the neural networks considered in this paper) with sigmoidal activation function, it follows that a function outside $\mathfrak{S}(\mathbb{R}^n)$ will not be uniformly approximated. Whether the converse is true is not immediately clear.

A reason to favour the uniform norm over the L^p-norms has already been highlighted in the caption of Fig. 2. Namely, while the latter is infinite for all neural networks, the former can be finite and actually describe convergence to local (C_0) functions. One may also compare our results with those of weighted L^p spaces, although they surrender translation invariance of the norm. A related class of topologies is given by the weighted supremum norms, as considered in Cuchiero, Schmocker, and Teichmann (2023). They form a natural class of topologies on $C(\mathbb{R}^n)$ that lie between the uniform topology and the topology of compact convergence, and it would therefore be interesting to relate the results of Cuchiero et al. (2023) to those obtained here.

On a different note, precisely because our results concern the uniform topology, they may yield quite tangible statements about the structure of a neural network *after* training. For instance, the results of Sections 4 and 5 imply that deep neural networks can be represented well as one-layer or two-layer neural networks without significant information loss at any scale, i.e., preserving the generalisation. Similarly, Theorem 5.9 relates deep feedforward neural networks to Sum Of Product Neural Networks, as in Lin and Li (2000), Long et al. (2007), which might be more light-weight than the corresponding deep neural network. It is unclear whether these observations lead to a practically feasible compression method, but the fact that there are ways to alternatively represent neural networks with arbitrary small loss deserves further investigation.

A final question is whether one can use the results of Theorem 2.4 to reasonably decide whether to believe that an unknown model is a neural network, based on its responses to suitably chosen inputs.

The author expects further investigation of uniform universal approximation outside $[0,1]^n$ to be worthwhile, not in the least because it will require thinking outside the box.

CRediT authorship contribution statement

Teun D.H. van Nuland: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known 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.

Acknowledgements

The author thanks Marjolein Troost for the crucial observations that sparked this paper, and for providing important guidance. The author is moreover indebted to Klaas Landsman, Walter van Suijlekom, Wim Wiegerinck, and the anonymous referees for helpful comments. This work was supported in part by NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) Physics Projectruimte 680-91-101, in part by NSF (National Science Foundation) grant DMS-1554456, in part by ARC (Australian Research Council) grant FL17010005, and in part by NWO grant OCENW.M.22.070.

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930–945.

Bauer, W., & Fulsche, R. (2023). Resolvent algebra in Fock-Bargmann representation. In *Semigroups, Algebras and Operator Theory*: *vol. 436*, (pp. 195–228). Springer Proceedings in Mathematics & Statistics.

Buchholz, D., & Grundling, H. (2007). Algebraic supersymmetry: A case study. *Communications in Mathematical Physics*, *272*(3), 699–750.

Buchholz, D., & Grundling, H. (2008). The resolvent algebra: A new approach to canonical quantum systems. *Journal of Functional Analysis*, *254*(11), 2725–2779.

Buchholz, D., & van Nuland, T. D. H. (2023). The basic resolvents of position and momentum operators form a total set in the resolvent algebra. *Letters in Mathematical Physics*, *113*(119).

Cuchiero, C., Schmocker, P., & Teichmann, J. (2023). Global universal approximation of functional input maps on weighted spaces. ArXiv preprint, 2306.03303 [stat.ML].

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

Gripenberg, G. (2003). Approximation by neural networks with a bounded number of nodes at each level. *J. Approx. Theory*, *122*(2), 260–266.

Hartman, E. J., Keeler, J. D., & Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*, *2*(2), 210–215.

Hashimoto, Y., Wang, Z., & Matsui, T. (2022). C*-algebra net: A new approach generalizing neural network parameters to C*-algebra. In *Proceedings of the 39th international conference on machine learning*: *vol. 162*, (pp. 8523–8534). PMLR.

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

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

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

J., Murphy G. (1990). *C*-algebras and operator theory*. Boston, MA: Academic Press Inc..

Kidger, P., & Lyons, T. (2020). Universal approximation with deep narrow networks. In *Proceedings of thirty third conference on learning theory*: *vol. 125*, (pp. 2306–2327). PMLR.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867.

Lin, C. S., & Li, C. K. (2000). A sum-of-product neural network (SOPNN). *Neurocomputing*, *30*(1–4), 273–291.

Long, J., Wu, W., & Nan, D. (2007). Uniform approximation capabilities of sum-of-product and sigma-pi-sigma neural networks. In *International symposium on neural networks* (pp. 1110–1116). Berlin, Heidelberg: Springer.

Maragos, P., Charisopoulos, V., & Theodosis, E. (2021). Tropical geometry and machine learning. *Proceedings of the IEEE*, *109*(5), 728–755.

Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*: *vol. 27*, (pp. 2924–2932).

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143–195.

Schuld, M., Sinayskiy, I., & Petruccione, F. (2014). The quest for a quantum neural network. *Quantum Information Processing*, *13*, 2567–2586.

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

Shen, Z., Yang, H., & Zhang, S. (2022). Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *The Journal of Machine Learning Research*, *23*(1), 12653–12712.

van Nuland, T. D. H. (2019). Quantization and the resolvent algebra. *Journal of Functional Analysis*, *277*(8), 2815–2838.

van Nuland, T. D. H. (2022). Strict deformation quantization of Abelian lattice gauge fields. *Letters in Mathematical Physics*, *112*(34), 1–29.

van Nuland, T. D. H., & Stienstra, R. (2020). Classical and quantised resolvent algebras for the cylinder. arXiv preprint. ArXiv preprint, 2003.13492 [math-ph].

Yarotsky, D. (2021). Elementary superexpressive activations. In *International conference on machine learning* (pp. 11932–11940). PMLR.

Zhang, L., Naitzat, G., & Lim, L. H. (2018). Tropical geometry of deep neural networks. In *Proceedings of the 35th international conference on machine learning*: *vol. 80*, (pp. 5824–5832). PMLR.