Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An ideal-theoretic characterization of uni- versal approximation for hypercomplex- valued neural networks

Muhammad Saad

m.saad@alexu.edu.eg

Alexandria University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11790710/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An ideal-theoretic characterization of universal approximation for hypercomplex-valued neural networks

Muhammad Saad

Abstract. Hypercomplex-valued neural networks take weights in a finite-dimensional real algebra, such as the quaternions, and apply a real activation function coordinatewise. Whether such networks are universal approximators depends on the basis in which the activation acts: for the dual numbers, one basis yields universal single-hidden-layer networks, whereas in another universality fails at every depth. We explain this phenomenon through the ideal theory of the algebra, for every finite-dimensional real algebra and every continuous non-polynomial activation function. Given an ordered basis $(e_1, \ldots, e_d)$ with dual basis $(e_1^*, \ldots, e_d^*)$, call an index j a *Frobenius coordinate* when the bilinear form $(w, x) \mapsto e_j^*(wx)$ is non-degenerate. With one hidden layer, networks with algebra-valued output weights are universal if and only if the left multiples of the basis vectors e_j corresponding to Frobenius coordinates span the algebra; with real output weights, universality holds if and only if every coordinate is a Frobenius coordinate. For an associative algebra with identity, a basis satisfying the algebra-valued criterion exists if and only if the algebra is a Frobenius algebra. For deeper networks, a recursion on the support digraph of the multiplication table provides an exact criterion. When the algebra has a left identity, universality holds at some depth if and only if no proper nonzero left ideal is spanned by basis vectors, in which case depth $d - 1$ suffices in dimension $d \geq 2$, and this bound is sharp. Associative non-Frobenius algebras with identity therefore exhibit a depth separation between one and two hidden layers, which we quantify explicitly. These results explain the failures observed for degenerate Clifford algebras, resolve the extension left open by Valle et al. [2024], and yield a counterexample to a recently published universal approximation theorem.

Mathematics Subject Classification (2010). Primary 68T07, 15A66; Secondary 16G10, 13C05.

ORCID: https://orcid.org/0000-0002-0697-6325.

Keywords. Clifford algebra, Depth separation, Frobenius algebra, Hypercomplex-valued neural network, Left ideal, Split activation function, Universal approximation.

1. Introduction

Hypercomplex-valued neural networks process multidimensional signals by allowing inputs, weights, and biases to take values in a finite-dimensional real algebra, such as the complex numbers, the quaternions [Arena et al., 1997], the octonions [Popa, 2016], or a Clifford algebra [Buchholz and Sommer, 2008]. Left multiplication by an algebra-valued weight couples signal coordinates in a structured manner, reducing the number of parameters; parameterized hypercomplex layers go further by learning the multiplication rule directly from data [Zhang et al., 2021]. Geometric (Clifford) algebras, including degenerate ones such as the projective algebra $\mathrm{Cl}(3,0,1)$, also play a central role in recent equivariant architectures [Brehmer et al., 2023, de Haan et al., 2024].

A simple example shows that the most basic theoretical question about these networks—whether they are universal approximators—is more subtle than in the real-valued setting. Consider the dual numbers $\mathbb{D} = \mathbb{R}[\varepsilon]/(\varepsilon^2)$ equipped with a split activation function, which applies a real function to each coordinate. In the basis $(1, 1+\varepsilon)$, single-hidden-layer networks approximate every continuous function on compact sets. In the basis $(1, \varepsilon)$, by contrast, the real part of the output of every network, regardless of width or depth, depends only on the real parts of the inputs (Example 6.1). The algebra, the activation function, and the architecture are unchanged; only the coordinate basis in which the activation acts has been altered. The purpose of this paper is to determine precisely when and why such obstructions occur across all finite-dimensional real algebras.

Universality is well understood for real- and complex-valued networks. Single-hidden-layer real networks are universal if and only if the activation is non-polynomial [Leshno et al., 1993, Pinkus, 1999], and for complex-valued networks the activations yielding universality have been completely characterized [Voigtlaender, 2023]. For hypercomplex-valued networks with split activations, universality was established for quaternions by Arena et al. [1997] and for non-degenerate Clifford algebras by Buchholz and Sommer, who also stated that degenerate Clifford algebras do not yield universal approximation [Buchholz and Sommer, 2008, Sections 5.1 and 8]. Vital et al. [2022] introduced non-degenerate hypercomplex algebras and stated a universal approximation theorem for them, which Valle et al. [2024] proved and extended, unifying the earlier results: if every bilinear form $(w, x) \mapsto e_j^*(wx)$ attached to a basis $(e_1, \ldots, e_d)$ is non-degenerate, then single-hidden-layer networks

with real output weights are universal, as are those with algebra-valued output weights when the algebra has an identity. They noted that this condition depends on the choice of basis, observed numerically that it cannot simply be dropped, and remarked that algebra-valued output weights could be admitted for broader classes of algebras under additional assumptions left unspecified [Valle et al., 2024, Remark 3]. In a tensor formulation, Niemczynowicz and Kycia [2026] replaced this condition with a weaker one. Related universality results have also been established for networks with structured weight matrices [Zhao et al., 2017] and for equivariant networks built on regular representations [Ravanbakhsh, 2020].

These results leave four questions open: Is the non-degeneracy condition necessary? What is the exact criterion when the output weights are algebra-valued? Is the dependence on the basis merely an artifact, or does it reflect an intrinsic property of the algebra? Finally, what changes when the network has more than one hidden layer? We answer all four questions within an ideal-theoretic framework.

The central observation is that the obstructions to universality are left ideals of the algebra spanned by basis vectors. Because a split activation acts on each coordinate separately, it maps cosets of such an ideal into cosets, and left multiplication by weights preserves the ideal; consequently, no network can distinguish inputs that differ by an element of the ideal (Proposition 5.5). In the basis $(1, \varepsilon)$ of the dual numbers, $\mathbb{R}\varepsilon$ is precisely such an ideal. At depth one, a second, finer obstruction arises, governed by whether a coordinate functional is a Frobenius form—a classical notion from the representation theory of algebras. Our main contributions are as follows.

- For any finite-dimensional real algebra, associative or not, and any continuous non-polynomial activation, single-hidden-layer networks with algebra-valued output weights are universal if and only if the left multiples of the basis vectors whose slice forms are non-degenerate span the algebra (Theorem 4.1); for an associative algebra with identity, these are the basis vectors whose coordinate functionals are Frobenius forms, and the condition states that they generate the algebra as a left ideal. With real output weights, the condition of Valle et al. [2024] is both necessary and sufficient (Theorem 4.2). This identifies the exact assumptions anticipated in their Remark 3, and Example 4.3 shows that algebra-valued output weights are strictly more expressive.
- For an associative algebra with identity, a basis yielding universal single-hidden-layer networks exists if and only if the algebra is a Frobenius algebra, in which case almost every basis succeeds (Theorem 4.4).
- For networks of depth k over an arbitrary algebra, we obtain an exact criterion via a recursion on blind subspaces (Theorem 5.1). When the algebra has a left identity, universality holds at some depth if and only if the support digraph G of the multiplication table is strongly connected—equivalently, no proper nonzero left ideal is spanned by basis vectors; in dimension $d \geq 2$, depth $\min\{1 + \operatorname{diam} G, d - 1\}$ then

suffices (Theorem 5.2), and the bound $d-1$ is sharp (Example 5.4). Without a left identity, this equivalence fails (Remark 5.3).
- Consequently, for an associative non-Frobenius algebra with identity, networks with two hidden layers are universal in almost every basis, whereas no basis yields universal single-hidden-layer networks (Corollary 5.6). For $\mathbb{R}[x,y]/(x,y)^2$, we compute the exact error floor of all single-hidden-layer networks on an explicit target function (Proposition 6.4).
- In their standard bases, split-activation networks over the dual numbers, the dual quaternions, and degenerate Clifford algebras fail to be universal at any depth (Example 6.1, Proposition 6.2). This confirms the assertion of Buchholz and Sommer for the blade basis while showing that it depends on the choice of basis, since every Clifford algebra is a Frobenius algebra.
- The weaker non-degeneracy hypothesis of Niemczynowicz and Kycia [2026] is satisfied by every algebra with identity, and the dual numbers provide a counterexample to their Theorem 3 (Remark 6.5).

All of these criteria reduce to finite linear-algebraic and graph-theoretic computations on the multiplication table, as detailed in Remark 5.7. The remainder of the paper is organized as follows. Section 2 establishes notation and recalls the required facts on Frobenius algebras. Section 3 proves three auxiliary approximation lemmas. Sections 4 and 5 analyze networks with one hidden layer and with multiple hidden layers, respectively. Section 6 develops the main examples and counterexamples, Section 7 presents numerical experiments, and Section 8 concludes with open problems.

2. Algebras, bases, and networks

Throughout this paper, an *algebra* is a finite-dimensional real vector space $\mathcal{A}$ equipped with a bilinear multiplication $(x,y)\mapsto xy$. Neither associativity nor the existence of an identity is assumed unless explicitly stated; this setting therefore encompasses the octonions, the sedenions, the networks with arbitrary bilinear products of Fan et al. [2020], and the learned multiplication rules of parameterized hypercomplex layers [Zhang et al., 2021]. We say that $\mathcal{A}$ has a *left identity* if there exists $1\in\mathcal{A}$ such that $1x=x$ for all $x\in\mathcal{A}$, and an *identity* if $1x=x1=x$ for all $x\in\mathcal{A}$.

Fix an ordered basis $\mathcal{E}=(e_1,\dots,e_d)$ of $\mathcal{A}$ and let $e_1^*,\dots,e_d^*$ denote the dual basis, so that $x=\sum_{j=1}^{d} e_j^*(x)\,e_j$ for all $x\in\mathcal{A}$. For each $j\in\{1,\dots,d\}$, we define the *j-th slice* of the multiplication to be the bilinear form

$$\beta_j\colon \mathcal{A}\times\mathcal{A}\to\mathbb{R},\qquad \beta_j(w,x)=e_j^*(wx), \tag{2.1}$$

and denote its right radical by

$$N_j=\{x\in\mathcal{A}:\ e_j^*(wx)=0 \text{ for all } w\in\mathcal{A}\}. \tag{2.2}$$

In the basis $\mathcal{E}$, the form β_j is represented by the matrix $B_j = \big(e_j^*(e_a e_b)\big)_{a,b=1}^{d}$, which coincides with the matrices introduced by Valle et al. [2024, Definition 2]. Since $\mathcal{A}$ is finite-dimensional, $N_j = 0$ if and only if B_j is non-singular, or equivalently, if and only if β_j is non-degenerate. We set

$$S(\mathcal{E}) = \{j \in \{1, \dots, d\} : \ N_j = 0\},$$

and call its elements the *Frobenius coordinates* of $\mathcal{E}$. For any $e \in \mathcal{A}$, we write $\mathcal{A}e = \{we : \ w \in \mathcal{A}\}$, which is a linear subspace of $\mathcal{A}$.

Definition 2.1. The *support digraph* $G(\mathcal{A}, \mathcal{E})$ has vertex set $\{1, \dots, d\}$ and a directed edge $i \to j$ whenever $e_j^*(\mathcal{A}e_i) \neq 0$, that is, whenever $e_j^*(e_a e_i) \neq 0$ for some $a \in \{1, \dots, d\}$. A subspace $L \subseteq \mathcal{A}$ is a *coordinate subspace* if $L = \operatorname{span}\{e_i : \ i \in T\}$ for some $T \subseteq \{1, \dots, d\}$, and a *coordinate left ideal* if, in addition, $\mathcal{A}L \subseteq L$.

A coordinate subspace $\operatorname{span}\{e_i : \ i \in T\}$ is a coordinate left ideal if and only if no edge of $G(\mathcal{A}, \mathcal{E})$ leaves T. Indeed, since $we_i = \sum_{j=1}^{d} e_j^*(we_i)e_j$, we have $\mathcal{A}e_i \subseteq \operatorname{span}\{e_j : \ j \in T\}$ for all $i \in T$ if and only if $i \in T$ and $i \to j$ imply $j \in T$.

When $\mathcal{A}$ is associative with identity, N_j is the largest left ideal of $\mathcal{A}$ contained in $\ker e_j^*$: it is a left ideal because $w(ux) = (wu)x$, it lies in $\ker e_j^*$ because $x = 1x$, and every left ideal $I \subseteq \ker e_j^*$ satisfies $e_j^*(\mathcal{A}I) \subseteq e_j^*(I) = 0$. We now recall the relevant facts on Frobenius algebras from Skowroński and Yamagata [2011]. A finite-dimensional associative algebra with identity is a *Frobenius algebra* if its left and right regular representations are equivalent. By the Brauer–Nesbitt–Nakayama theorem [Skowroński and Yamagata, 2011, Theorem IV.2.1], this holds if and only if $\mathcal{A}$ admits a non-degenerate associative bilinear form (that is, a non-degenerate bilinear form β satisfying $\beta(uv, w) = \beta(u, vw)$ for all $u, v, w \in \mathcal{A}$), or equivalently, a linear form $\varphi\colon \mathcal{A} \to \mathbb{R}$ whose kernel contains no nonzero left ideal (equivalently, no nonzero right ideal). We call such a functional φ a *Frobenius form*. The bilinear form $(w, x) \mapsto \varphi(wx)$ is automatically associative, and it is non-degenerate precisely when φ is a Frobenius form [Skowroński and Yamagata, 2011, proof of Theorem IV.2.1]. Consequently, for an associative algebra $\mathcal{A}$ with identity,

$$j \in S(\mathcal{E}) \iff e_j^* \text{ is a Frobenius form}, \tag{2.3}$$

and $S(\mathcal{E}) \neq \emptyset$ for some basis $\mathcal{E}$ only if $\mathcal{A}$ is a Frobenius algebra.

Definition 2.2. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$. The *split activation* on $\mathcal{A}$ with respect to $\mathcal{E}$ is defined by $\sigma_{\mathcal{E}}\big(\sum_{j=1}^{d} x_j e_j\big) = \sum_{j=1}^{d} \sigma(x_j)e_j$, and is applied componentwise on $\mathcal{A}^m$. An *affine $\mathcal{A}$-map* $\mathcal{A}^n \to \mathcal{A}^m$ is a map of the form $x \mapsto Wx + b$, where $W = (w_{rl}) \in \mathcal{A}^{m \times n}$, $b \in \mathcal{A}^m$, and $(Wx)_r = \sum_{l=1}^{n} w_{rl}x_l$, with weights acting by left multiplication. For $k \geq 1$, a *depth-k $\mathcal{A}$-network* is a map $F\colon \mathcal{A}^n \to \mathcal{A}^m$ of the form

$$x^{(0)} = x, \qquad x^{(\ell)} = \sigma_{\mathcal{E}}\big(W_\ell x^{(\ell-1)} + b_\ell\big) \ (1 \leq \ell \leq k), \qquad F(x) = W_{k+1}x^{(k)},$$

where the affine $\mathcal{A}$-maps may have arbitrary widths. We denote the set of such maps by $\mathcal{N}_k(n,m)$, and by $\mathcal{N}_1^{\mathbb{R}}(n,m)$ the set of depth-one networks whose output matrix W_2 has real entries. A class $\mathcal{F}$ of continuous maps $\mathcal{A}^n \to \mathcal{A}^m$ is *universal* if, for every compact set $K \subset \mathcal{A}^n$, the restrictions $\{F|_K : F \in \mathcal{F}\}$ are dense in $C(K, \mathcal{A}^m)$ with respect to the uniform norm.

For $k = 1$ and $m = 1$, these coincide with the single-hidden-layer V-MLP networks of Valle et al. [2024, Definition 6], with vector-valued output weights for $\mathcal{N}_1$ and scalar output weights for $\mathcal{N}_1^{\mathbb{R}}$. The quaternion networks of Arena et al. [1997], the Clifford networks of Buchholz and Sommer [2008, Eq. (44)], and the octonion networks of Popa [2016] employ the same layer structure: left multiplication by algebra-valued weights followed by a split activation.

Because two networks of the same depth can be combined in parallel, $\mathcal{N}_k(n,m)$ is a real vector space, and a map belongs to $\mathcal{N}_k(n,m)$ if and only if each of its m output components belongs to $\mathcal{N}_k(n,1)$; the same holds for $\mathcal{N}_1^{\mathbb{R}}$. Universality may therefore be verified for $m = 1$, and all results below are stated in that setting. Throughout the paper, calligraphic symbols ($\mathcal{N}_k$, $\mathcal{N}_1^{\mathbb{R}}$) denote classes of networks, whereas italic symbols (N_j, $N_j^{(k)}$) denote subspaces of $\mathcal{A}$.

For a compact set $K \subset \mathcal{A}^n$, we write $C(K) = C(K, \mathbb{R})$, equipped with the uniform norm. Given a subspace $N \subseteq \mathcal{A}$, let

$$C(K; N) = \{f \in C(K) : \ f(x) = f(y) \text{ whenever } x, y \in K \text{ and } x - y \in N^n\}.$$

Thus $C(K; 0) = C(K)$, and $N \subseteq N'$ implies $C(K; N') \subseteq C(K; N)$. For subspaces $V \subseteq \mathcal{A}$ and $\mathcal{F} \subseteq C(K)$, we write $V \otimes \mathcal{F}$ for the linear span of the maps $x \mapsto v f(x)$ with $v \in V$ and $f \in \mathcal{F}$.

3. Three approximation lemmas

The first lemma is a variant of the density theorem of Leshno et al. [1993] in which only differences of two translates of σ sharing a common scale are available. This is precisely the form that an algebra-valued hidden unit can isolate in a single coordinate.

Lemma 3.1. *Let $\sigma \in C(\mathbb{R})$ be non-polynomial and define*

$$D(\sigma) = \operatorname{span}\{\, t \mapsto \sigma(\lambda t + \theta) - \sigma(\lambda t + \theta') \ : \ \lambda, \theta, \theta' \in \mathbb{R} \,\}.$$

Then $D(\sigma)$ is dense in $C(\mathbb{R})$ in the topology of uniform convergence on compact sets.

Proof. Let $\overline{D}$ denote the closure of $D(\sigma)$ in $C(\mathbb{R})$ with respect to uniform convergence on compact sets; then $\overline{D}$ is a closed linear subspace. Setting $\lambda = 0$ yields the constants $\sigma(\theta) - \sigma(\theta')$, which do not all vanish because σ is not constant; hence $1 \in \overline{D}$.

Assume first that $\sigma \in C^\infty(\mathbb{R})$. For fixed $\theta, \theta' \in \mathbb{R}$, the map $\lambda \mapsto \sigma(\lambda t + \theta) - \sigma(\lambda t + \theta')$ is smooth, and its difference quotients in λ converge uniformly for t in any compact set. Consequently, $\overline{D}$ contains

$$\frac{\partial^k}{\partial \lambda^k}\Big|_{\lambda=0} \big[\sigma(\lambda t + \theta) - \sigma(\lambda t + \theta')\big] = t^k \big(\sigma^{(k)}(\theta) - \sigma^{(k)}(\theta')\big), \qquad k \geq 1.$$

If $\sigma^{(k)}$ were constant for some $k \geq 1$, then σ would be a polynomial; hence for every $k \geq 1$ there exist $\theta, \theta' \in \mathbb{R}$ such that $\sigma^{(k)}(\theta) \neq \sigma^{(k)}(\theta')$, giving $t^k \in \overline{D}$. Thus $\overline{D}$ contains all polynomials and is dense in $C(\mathbb{R})$ by the Weierstrass theorem, as in Pinkus [1999, Proposition 3.4].

Now let $\sigma \in C(\mathbb{R})$ be arbitrary and non-polynomial; we follow the argument of Pinkus [1999, Proposition 3.7]. For $\phi \in C_c^\infty(\mathbb{R})$, set $\sigma_\phi = \sigma * \phi \in C^\infty(\mathbb{R})$. Since

$$\sigma_\phi(\lambda t + \theta) - \sigma_\phi(\lambda t + \theta') = \int_{\mathbb{R}} \big[\sigma(\lambda t + \theta - y) - \sigma(\lambda t + \theta' - y)\big] \phi(y)\, dy$$

is a uniform limit on compact sets of Riemann sums of functions in $D(\sigma)$, we have $D(\sigma_\phi) \subseteq \overline{D}$. Suppose toward a contradiction that $\overline{D} \neq C(\mathbb{R})$. Then $t^k \notin \overline{D}$ for some $k \geq 1$, so $t^k \notin \overline{D(\sigma_\phi)}$ for every $\phi \in C_c^\infty(\mathbb{R})$, and the smooth case implies that $\sigma_\phi^{(k)}$ is constant—that is, σ_ϕ is a polynomial of degree at most k—for every $\phi \in C_c^\infty(\mathbb{R})$. Choosing a sequence of mollifiers $\phi_n \in C_c^\infty(\mathbb{R})$ with $\sigma_{\phi_n} \to \sigma$ uniformly on compact sets, and using the fact that the polynomials of degree at most k form a closed finite-dimensional subspace, we conclude that σ is a polynomial, a contradiction. □

Lemma 3.2. *Let E be a finite-dimensional real vector space, let $\mathcal{D} \subseteq E^*$ be a linear subspace with annihilator $\mathcal{D}^\perp = \{x \in E : \ \ell(x) = 0 \text{ for all } \ell \in \mathcal{D}\}$, and let $K \subset E$ be compact. Then the closure in $C(K)$ of* $\operatorname{span}\{\psi \circ \ell|_K : \ \ell \in \mathcal{D},\ \psi \in C(\mathbb{R})\}$ *equals*

$$C_{\mathcal{D}^\perp}(K) = \{f \in C(K) : \ f(x) = f(y) \text{ whenever } x, y \in K \text{ and } x - y \in \mathcal{D}^\perp\}.$$

Proof. Every function $\psi \circ \ell|_K$ with $\ell \in \mathcal{D}$ belongs to $C_{\mathcal{D}^\perp}(K)$, which is closed in $C(K)$. Conversely, let $\ell_1, \dots, \ell_r$ be a basis of $\mathcal{D}$ and define $Q = (\ell_1, \dots, \ell_r) \colon E \to \mathbb{R}^r$, so that $\ker Q = \mathcal{D}^\perp$. The restriction $Q|_K$ is a continuous closed map onto the compact set $Q(K)$, hence a quotient map, and therefore $C_{\mathcal{D}^\perp}(K) = \{g \circ Q|_K : \ g \in C(Q(K))\}$. By the polarization identity

$$\ell_{i_1} \cdots \ell_{i_k} = \frac{1}{2^k k!} \sum_{\epsilon \in \{\pm 1\}^k} \epsilon_1 \cdots \epsilon_k \Big(\sum_{s=1}^{k} \epsilon_s \ell_{i_s} \Big)^k,$$

every monomial in $\ell_1, \dots, \ell_r$ is a linear combination of powers ℓ^k with $\ell \in \mathcal{D}$. Thus the span contains $p \circ Q|_K$ for every polynomial p on $\mathbb{R}^r$, and these functions are dense in $\{g \circ Q|_K : \ g \in C(Q(K))\}$ by the Stone–Weierstrass theorem on $Q(K)$. □

Lemma 3.2 corresponds to the case of a full subspace of directions in the ridge-function theory of Lin and Pinkus [1993]; we include the short proof above to keep the presentation self-contained.

Lemma 3.3. *Let $V_1, \dots, V_d \subseteq \mathcal{A}$ and $M_1, \dots, M_d \subseteq \mathcal{A}$ be linear subspaces, and set $J = \{j \in \{1, \dots, d\} : \ M_j \neq 0\}$. Let $K \subset \mathcal{A}^n$ be a compact set with nonempty interior. If $\sum_{j \notin J} V_j \neq \mathcal{A}$, then the closure of $\sum_{j=1}^{d} V_j \otimes C(K; M_j)$ is a proper subspace of $C(K, \mathcal{A})$.*

Proof. Choose a nonzero functional $u \in \mathcal{A}^*$ vanishing on $\sum_{j \notin J} V_j$. For each $j \in J$, pick a nonzero vector $v_j \in M_j$ and set $\tilde{v}_j = (v_j, 0, \dots, 0) \in M_j^n$. Choose a point x_0 in the interior of K and a scalar $t > 0$ small enough that all points $x_T = x_0 + t \sum_{j \in T} \tilde{v}_j$ $(T \subseteq J)$ lie in K, and define the signed measure $\mu = \sum_{T \subseteq J} (-1)^{|J|-|T|} \delta_{x_T}$. The linear functional $\Phi(F) = \int u(F)\, d\mu$ is continuous on $C(K, \mathcal{A})$. If $j \notin J$ and $v \in V_j$, then $\Phi(vf) = u(v) \int f\, d\mu = 0$. If $j \in J$ and $f \in C(K; M_j)$, then pairing each subset $T \subseteq J \setminus \{j\}$ with $T \cup \{j\}$ yields $\int f\, d\mu = 0$, because $x_{T \cup \{j\}} - x_T = t\tilde{v}_j \in M_j^n$. Hence Φ vanishes on the closure of $\sum_{j=1}^{d} V_j \otimes C(K; M_j)$.

To show that $\Phi \neq 0$, choose $a \in \mathcal{A}$ with $u(a) = 1$ and $\xi \in (\mathcal{A}^n)^*$ satisfying $\xi(\tilde{v}_j) \neq 0$ for all $j \in J$, and define $F(x) = a\, \xi(x - x_0)^{|J|}$. For a homogeneous polynomial of degree $r = |J|$, the r-th iterated difference equals the corresponding mixed directional derivative, giving $\Phi(F) = t^r\, r! \prod_{j \in J} \xi(\tilde{v}_j) \neq 0$ (with $\Phi(F) = u(a) = 1$ when $J = \emptyset$). Therefore $\Phi \neq 0$, and the closure is proper. □

4. Networks with one hidden layer

In a single-hidden-layer network, each hidden unit contributes to the output through the individual coordinates of its pre-activation; universality therefore depends on which input variations are visible to a single coordinate of $w \cdot x$. The directions invisible to the j-th coordinate form precisely the right radical N_j of the j-th slice.

For $w = (w_1, \dots, w_n) \in \mathcal{A}^n$ and $x = (x_1, \dots, x_n) \in \mathcal{A}^n$, write $w \cdot x = \sum_{l=1}^{n} w_l x_l$. For each $j \in \{1, \dots, d\}$, define the linear functional $\ell_{w,j}(x) = e_j^*(w \cdot x)$ and the subspace $\mathcal{D}_j = \{\ell_{w,j} : \ w \in \mathcal{A}^n\} \subseteq (\mathcal{A}^n)^*$. Then $\ell_{w,j}(x) = 0$ for all $w \in \mathcal{A}^n$ if and only if $x_l \in N_j$ for every $l \in \{1, \dots, n\}$, so that

$$\mathcal{D}_j^{\perp} = N_j^n. \tag{4.1}$$

A hidden unit with output weight $\alpha \in \mathcal{A}$ and bias $b = \sum_{j=1}^{d} b_j e_j$ decomposes along the basis $\mathcal{E}$ as

$$\alpha\, \sigma_{\mathcal{E}}(w \cdot x + b) = \sum_{j=1}^{d} \sigma\big(\ell_{w,j}(x) + b_j\big)\, \alpha e_j. \tag{4.2}$$

Theorem 4.1. *Let $\mathcal{A}$ be an algebra with basis $\mathcal{E}$ and let $\sigma \in C(\mathbb{R})$ be non-polynomial. For every compact set $K \subset \mathcal{A}^n$,*

$$\overline{\mathcal{N}_1(n,1)|_K} = \overline{\sum_{j=1}^{d} (\mathcal{A}e_j) \otimes C(K; N_j)}.$$

Consequently, $\mathcal{N}_1$ is universal if and only if

$$\sum_{j \in S(\mathcal{E})} \mathcal{A}e_j = \mathcal{A}. \tag{4.3}$$

Proof. By (4.1) and (4.2), every network in $\mathcal{N}_1(n,1)|_K$ lies in $\sum_{j=1}^{d} (\mathcal{A}e_j) \otimes C(K; N_j)$. Conversely, fix $j \in \{1, \dots, d\}$, $\alpha \in \mathcal{A}$, $w \in \mathcal{A}^n$, $b \in \mathcal{A}$, and $\lambda, h \in \mathbb{R}$. The difference of the two hidden units $\alpha\sigma_{\mathcal{E}}(\lambda w \cdot x + b + he_j)$ and $\alpha\sigma_{\mathcal{E}}(\lambda w \cdot x + b)$ belongs to $\mathcal{N}_1(n,1)$, and by (4.2) all coordinates except the j-th cancel:

$$\begin{aligned} &\alpha\sigma_{\mathcal{E}}(\lambda w \cdot x + b + he_j) - \alpha\sigma_{\mathcal{E}}(\lambda w \cdot x + b) \\ &\qquad = \big[\sigma(\lambda \ell_{w,j}(x) + b_j + h) - \sigma(\lambda \ell_{w,j}(x) + b_j)\big]\, \alpha e_j. \end{aligned}$$

By Lemma 3.1, the closure $\overline{\mathcal{N}_1(n,1)|_K}$ contains $\psi(\ell_{w,j}(x))\, \alpha e_j$ for every $\psi \in C(\mathbb{R})$; by Lemma 3.2 and (4.1) (noting that $C_{\mathcal{D}_j^{\perp}}(K) = C(K; N_j)$), it therefore contains $(\mathcal{A}e_j) \otimes C(K; N_j)$. This establishes the displayed identity.

If (4.3) holds, then since $C(K; N_j) = C(K)$ for all $j \in S(\mathcal{E})$, the closure contains $\sum_{j \in S(\mathcal{E})} (\mathcal{A}e_j) \otimes C(K) = \mathcal{A} \otimes C(K) = C(K, \mathcal{A})$. Conversely, if (4.3) fails, then Lemma 3.3 with $V_j = \mathcal{A}e_j$ and $M_j = N_j$ (so that J is the complement of $S(\mathcal{E})$) shows that for any compact set $K \subset \mathcal{A}^n$ with nonempty interior, such as a closed ball, the closure is a proper subspace of $C(K, \mathcal{A})$. □

Theorem 4.2. *Let $\mathcal{A}$, $\mathcal{E}$, and σ be as in Theorem 4.1. Then $\mathcal{N}_1^{\mathbb{R}}$ is universal if and only if $S(\mathcal{E}) = \{1, \dots, d\}$, that is, if and only if every slice β_j is non-degenerate.*

Proof. With real output weights, the j-th coordinate of any network in $\mathcal{N}_1^{\mathbb{R}}(n,1)|_K$ has the form $\sum_k \alpha_k \sigma(\ell_{w_k,j}(x) + b_{k,j}) \in C(K; N_j)$, so the entire class is contained in $\sum_{j=1}^{d} (\mathbb{R}e_j) \otimes C(K; N_j)$. Applying the difference construction from the proof of Theorem 4.1 with $\alpha \in \mathbb{R}$ shows that the closure contains $(\mathbb{R}e_j) \otimes C(K; N_j)$ for every j. If $N_j = 0$ for all j, this closure equals $C(K, \mathcal{A})$. If $N_j \neq 0$ for some j, then Lemma 3.3 applies with $V_i = \mathbb{R}e_i$ and $M_i = N_i$, since $\sum_{i \notin J} \mathbb{R}e_i$ does not contain e_j. □

Theorem 4.2 shows that the non-degeneracy condition of Valle et al. [2024, Theorem 1] is both necessary and sufficient for scalar output weights, and that it suffices for every continuous non-polynomial activation. Theorem 4.1 resolves the question left open in Valle et al. [2024, Remark 3] by identifying the exact assumption on the algebra that permits vector-valued output weights: namely, (4.3). When $\mathcal{A}$ has a left identity, condition (4.3) follows from $S(\mathcal{E}) = \{1, \dots, d\}$, which recovers their second statement, yet (4.3) is strictly weaker in general.

Example 4.3. Let $\mathcal{A} = \mathbb{R}[x]/(x^3)$ with basis $\mathcal{E} = (x,\, 1,\, 1 + x^2)$. Writing a generic element $u = a + bx + cx^2$ as $u = b\,x + (a - c)\,1 + c\,(1 + x^2)$, we see that the coordinate functionals are given by $e_1^*(u) = b$, $e_2^*(u) = a - c$, and $e_3^*(u) = c$. Every nonzero ideal of $\mathcal{A}$ contains x^2, and $x^2 \in \ker e_1^*$ whereas $x^2 \notin \ker e_2^* \cup \ker e_3^*$. Hence $S(\mathcal{E}) = \{2, 3\}$, and (4.3) holds because $\mathcal{A}e_2 = \mathcal{A}$. Thus $\mathcal{N}_1$ is universal despite the degeneracy of the first slice, whereas $\mathcal{N}_1^{\mathbb{R}}$ is not.

Theorem 4.4. *Let $\mathcal{A}$ be an associative algebra with identity and let $\sigma \in C(\mathbb{R})$ be non-polynomial.*

(a) If $\mathcal{A}$ is not a Frobenius algebra, then $\mathcal{N}_1$ is not universal for any basis of $\mathcal{A}$.

(b) If $\mathcal{A}$ is a Frobenius algebra, then $\mathcal{N}_1$ is universal for every basis whose dual basis consists of Frobenius forms, and such bases form a nonempty Zariski-open, hence dense, subset of the space of all bases of $\mathcal{A}$.

Proof. (a) By (2.3), we have $S(\mathcal{E}) = \emptyset$ for every basis $\mathcal{E}$, so (4.3) fails because $\mathcal{A} \neq 0$.

(b) Fix an arbitrary reference basis $(u_1, \dots, u_d)$ of $\mathcal{A}$. A linear form $\varphi \in \mathcal{A}^*$ is a Frobenius form if and only if $\det\big(\varphi(u_a u_b)\big) \neq 0$, a polynomial condition on φ that holds for at least one φ by hypothesis. The bases $(\varphi_1, \dots, \varphi_d)$ of $\mathcal{A}^*$ consisting of Frobenius forms therefore form the complement in $(\mathcal{A}^*)^d$ of finitely many proper algebraic hypersurfaces, which is a nonempty Zariski-open set; passing to dual bases is a homeomorphism of the space of bases. For any such dual basis $\mathcal{E} = (e_1, \dots, e_d)$, we have $S(\mathcal{E}) = \{1, \dots, d\}$ and $\sum_{j=1}^{d} \mathcal{A}e_j \supseteq \operatorname{span}\{1e_j :\ 1 \le j \le d\} = \mathcal{A}$, so Theorem 4.1 applies. □

5. Deep networks

In a network with multiple hidden layers, the post-activation output in coordinate i is mapped by the subsequent weight matrix into every coordinate j such that $i \to j$; consequently, an input direction invisible to one coordinate may become visible to another, and obstructions propagate along the support digraph.

Define subspaces $N_j^{(k)} \subseteq \mathcal{A}$ recursively by

$$N_j^{(1)} = N_j, \qquad N_j^{(k+1)} = \bigcap_{i:\ i \to j} N_i^{(k)}, \tag{5.1}$$

where the intersection is taken over the in-neighbors of j in $G(\mathcal{A}, \mathcal{E})$ (and is understood to be $\mathcal{A}$ if j has no in-neighbors). We call $N_j^{(k)}$ the *blind subspaces*: $N_j^{(k)}$ collects the input directions to which the j-th coordinate of a depth-k pre-activation is insensitive. Because information reaches coordinate j only along edges terminating at j, the recursion runs over in-neighbors.

Theorem 5.1. *Let $\mathcal{A}$ be an algebra with basis $\mathcal{E}$, let $\sigma \in C(\mathbb{R})$ be non-polynomial, and let $k \geq 1$. For every compact set $K \subset \mathcal{A}^n$,*

$$\overline{\mathcal{N}_k(n,1)|_K} = \overline{\textstyle\sum_{j=1}^{d} (\mathcal{A}e_j) \otimes C(K; N_j^{(k)})}.$$

Consequently, $\mathcal{N}_k$ is universal if and only if $\sum\{\mathcal{A}e_j : \ N_j^{(k)} = 0\} = \mathcal{A}$.

Proof. Let $\mathcal{P}_k$ denote the set of depth-k pre-activations $z\colon \mathcal{A}^n \to \mathcal{A}$, that is, maps of the form $z(x) = w \cdot x^{(k-1)} + b$, where $x^{(k-1)}$ is the output of $k-1$ hidden layers (with $x^{(0)} = x$). Combining networks in parallel shows that $\mathcal{P}_k$ is a real vector space closed under the addition of constants and under scalar multiplication $z \mapsto \lambda z$, and that every map $\sum_r \alpha_r \sigma_{\mathcal{E}}(z_r)$ with $z_r \in \mathcal{P}_k$ and $\alpha_r \in \mathcal{A}$ belongs to $\mathcal{N}_k(n,1)$. Write $z_j = e_j^* \circ z$, and let Ψ_k^j be the closure in $C(K)$ of $\operatorname{span}\{\psi \circ z_j : \ z \in \mathcal{P}_k, \ \psi \in C(\mathbb{R})\}$. We prove by induction on k that

$$z_j \in C(K; N_j^{(k)}) \text{ for all } z \in \mathcal{P}_k, \qquad \text{and} \qquad \Psi_k^j = C(K; N_j^{(k)}). \tag{5.2}$$

For $k = 1$, we have $z_j = \ell_{w,j} + b_j$, and (5.2) follows from (4.1) and Lemma 3.2. Assume now that (5.2) holds for k. Any pre-activation of depth $k+1$ has j-th coordinate

$$z_j = \sum_r e_j^*\big(w_r \sigma_{\mathcal{E}}(z^r)\big) + b_j = \sum_r \sum_{i=1}^{d} e_j^*(w_r e_i)\, \sigma(z_i^r) + b_j, \qquad z^r \in \mathcal{P}_k,$$

where $e_j^*(w_r e_i) \neq 0$ only when $i \to j$. Since $\sigma(z_i^r) \in C(K; N_i^{(k)}) \subseteq C(K; N_j^{(k+1)})$ for every $i \to j$, the first half of (5.2) holds for $k+1$, and consequently $\Psi_{k+1}^j \subseteq C(K; N_j^{(k+1)})$.

For the reverse inclusion, let $i \to j$ and choose $w \in \mathcal{A}$ such that $c_i = e_j^*(we_i) \neq 0$. For any $z \in \mathcal{P}_k$, $\lambda, h \in \mathbb{R}$, and $b \in \mathcal{A}$, the map

$$w\big[\sigma_{\mathcal{E}}(\lambda z + b + he_i) - \sigma_{\mathcal{E}}(\lambda z + b)\big] = \big[\sigma(\lambda z_i + b_i + h) - \sigma(\lambda z_i + b_i)\big]\, we_i$$

lies in $\mathcal{P}_{k+1}$ (as a depth-$(k+1)$ pre-activation with two hidden units and zero bias), and its j-th coordinate is $c_i\big[\sigma(\lambda z_i + b_i + h) - \sigma(\lambda z_i + b_i)\big]$. By Lemma 3.1 and the induction hypothesis, the closure Y_{k+1}^j of $\{z_j : \ z \in \mathcal{P}_{k+1}\}$ in $C(K)$, which is a linear subspace, contains $\Psi_k^i = C(K; N_i^{(k)})$ for every $i \to j$, and hence contains $\mathcal{L} = \sum_{i \to j} C(K; N_i^{(k)})$.

By the uniform continuity of ψ on compact sets, Ψ_{k+1}^j contains $\psi \circ f$ for every $f \in Y_{k+1}^j$ and $\psi \in C(\mathbb{R})$. Since Y_{k+1}^j is a linear subspace, taking $\psi(t) = t^m$ and applying the polarization identity from the proof of Lemma 3.2 to $f_1, \ldots, f_m \in Y_{k+1}^j$ shows that Ψ_{k+1}^j contains every product $f_1 \cdots f_m$ ($m \geq 1$). Thus Ψ_{k+1}^j contains the algebra generated by $\mathcal{L}$ and the constants. If $x, y \in K$ satisfy $x - y \notin (N_j^{(k+1)})^n$, then $x - y \notin (N_i^{(k)})^n$ for some $i \to j$, and a linear form vanishing on $(N_i^{(k)})^n$ but not at $x - y$ separates x and y inside $C(K; N_i^{(k)})$. By the Stone–Weierstrass theorem on the image

of K in $(\mathcal{A}/N_j^{(k+1)})^n$, this algebra is dense in $C(K; N_j^{(k+1)})$, completing the induction.

Finally, applying the difference construction of Theorem 4.1 to depth-k pre-activations shows that $\overline{\mathcal{N}_k(n,1)|_K} \supseteq (\mathcal{A}e_j) \otimes \Psi_k^j$ for every j, whereas every network in $\mathcal{N}_k(n,1)|_K$ lies in $\sum_{j=1}^d (\mathcal{A}e_j) \otimes C(K; N_j^{(k)})$ by (5.2). The universality criterion then follows as in Theorem 4.1, using Lemma 3.3 with $M_j = N_j^{(k)}$. □

Theorem 5.2. *Let $\mathcal{A}$ be an algebra with basis $\mathcal{E}$ and let $\sigma \in C(\mathbb{R})$ be non-polynomial.*

(a) If some coordinate subspace $\operatorname{span}\{e_i : i \in T\}$ *with* $\emptyset \neq T \neq \{1, \dots, d\}$ *is a left ideal—equivalently, if $G(\mathcal{A}, \mathcal{E})$ is not strongly connected—then $\mathcal{N}_k$ is not universal for any $k \geq 1$.*

(b) If $\mathcal{A}$ has a left identity and $G(\mathcal{A}, \mathcal{E})$ is strongly connected, then $\mathcal{N}_k$ is universal for every $k \geq 1 + \operatorname{diam} G(\mathcal{A}, \mathcal{E})$ and, if $d \geq 2$, for every $k \geq d - 1$.

Proof. (a) If $G = G(\mathcal{A}, \mathcal{E})$ is not strongly connected, there exist vertices i_0, j_0 with no directed path from i_0 to j_0. The set T of vertices reachable from i_0 is nonempty, omits j_0, and has no outgoing edges, so $L = \operatorname{span}\{e_i : i \in T\}$ is a proper nonzero coordinate left ideal. Conversely, a coordinate left ideal $\operatorname{span}\{e_i : i \in T\}$ with $\emptyset \neq T \neq \{1, \dots, d\}$ has no edge leaving T, so G is not strongly connected. For $j \notin T$, we have $\mathcal{A}L \subseteq L \subseteq \ker e_j^*$, hence $L \subseteq N_j^{(1)}$; moreover, an edge $i \to j$ with $j \notin T$ forces $i \notin T$, so by induction $L \subseteq N_j^{(k)}$ for all $k \geq 1$ and all $j \notin T$. Therefore $\sum\{\mathcal{A}e_j : N_j^{(k)} = 0\} \subseteq \sum_{j \in T} \mathcal{A}e_j \subseteq L \neq \mathcal{A}$, and Theorem 5.1 implies that $\mathcal{N}_k$ is not universal.

(b) Because $\mathcal{A}$ has a left identity 1, we have $e_j^*(1e_j) = 1$ for every j, so every vertex of G carries a self-loop and (5.1) gives

$$N_j^{(k)} = \bigcap \{N_i : \text{ there is a path from } i \text{ to } j \text{ of length at most } k-1\}.$$

If G is strongly connected and $k - 1 \geq \operatorname{diam} G$, every vertex i qualifies, so $N_j^{(k)} = \bigcap_{i=1}^d N_i = \{x \in \mathcal{A} : \mathcal{A}x = 0\} = 0$ for all j, using $1x = x$. Consequently, $\sum_{j=1}^d \mathcal{A}e_j \supseteq \operatorname{span}\{1e_j : 1 \leq j \leq d\} = \mathcal{A}$, and Theorem 5.1 yields universality.

To establish the bound $d-1$, assume $d \geq 2$. Setting $w = 1$ in (2.2) yields $N_i \subseteq \ker e_i^*$, so $\bigcap_{i \neq i_0} N_i \subseteq \bigcap_{i \neq i_0} \ker e_i^* = \mathbb{R}e_{i_0}$ for every $i_0 \in \{1, \dots, d\}$. If this intersection were nonzero, then $e_{i_0} \in N_i$—that is, $e_i^*(\mathcal{A}e_{i_0}) = 0$—for all $i \neq i_0$, meaning no edge would leave i_0, contradicting strong connectivity. Hence any $d-1$ of the subspaces N_i have trivial intersection. For a fixed vertex j, let R_m denote the set of vertices with a path to j of length at most m. Then $R_0 = \{j\}$, and as long as $R_m \neq \{1, \dots, d\}$, strong connectivity ensures that at least one edge enters R_m from outside, giving $|R_{m+1}| \geq |R_m| + 1$. Thus $|R_{d-2}| \geq d-1$. For $k \geq d-1$ we have $R_{k-1} \supseteq R_{d-2}$, so $N_j^{(k)} = \bigcap_{i \in R_{k-1}} N_i = 0$ for every j, and Theorem 5.1 again yields universality. □

Remark 5.3. The left-identity hypothesis in Theorem 5.2(b) cannot be omitted. Let $\mathcal{A}$ be the two-dimensional algebra with basis (e_1, e_2) and multiplication $e_1^2 = e_2$, $e_2^2 = e_1$, and $e_1e_2 = e_2e_1 = 0$. Then $\mathcal{A}e_1 = \mathbb{R}e_2$ and $\mathcal{A}e_2 = \mathbb{R}e_1$, so $G(\mathcal{A}, \mathcal{E})$ is the directed 2-cycle, which is strongly connected, and neither $\mathbb{R}e_1$ nor $\mathbb{R}e_2$ is a left ideal. Nevertheless, $N_1 = \mathbb{R}e_1$ and $N_2 = \mathbb{R}e_2$, and recursion (5.1) alternates between $(N_1^{(k)}, N_2^{(k)}) = (\mathbb{R}e_1, \mathbb{R}e_2)$ and $(\mathbb{R}e_2, \mathbb{R}e_1)$. By Theorem 5.1, $\mathcal{N}_k$ is not universal for any $k \geq 1$.

Example 5.4. The bound $d-1$ in Theorem 5.2(b) is attained in every dimension $d \geq 2$. With indices taken modulo d, define a multiplication on $\mathbb{R}^d$ in the standard basis $\mathcal{E} = (e_1, \ldots, e_d)$ by $e_1e_i = e_i$, $e_2e_i = e_{i+1}$, and $e_ke_i = 0$ for $k \geq 3$. Then e_1 is a left identity and $we_i = w_1e_i + w_2e_{i+1}$, so $G(\mathcal{A}, \mathcal{E})$ is the directed cycle $1 \to 2 \to \cdots \to d \to 1$ with a self-loop at every vertex, which is strongly connected. Since $e_i^*(wx) = w_1x_i + w_2x_{i-1}$, we obtain $N_i = \{x \in \mathbb{R}^d : \ x_{i-1} = x_i = 0\}$. For $m \leq d-1$, the vertices with a path of length at most m to j are $j-m, \ldots, j$, so the characterization of $N_j^{(k)}$ in the proof of Theorem 5.2(b) gives $N_j^{(k)} = \{x \in \mathbb{R}^d : \ x_{j-k} = \cdots = x_j = 0\}$ for $k \leq d-1$. When $k \leq d-2$, this subspace contains e_{j+1} for every j, so $\mathcal{N}_k$ is not universal by Theorem 5.1, whereas $\mathcal{N}_{d-1}$ is universal. For $d = 2$, this algebra coincides with the split-complex numbers in the basis $(1, u)$; for $d \geq 3$, it is non-associative, since $(e_2e_2)e_1 = 0 \neq e_3 = e_2(e_2e_1)$.

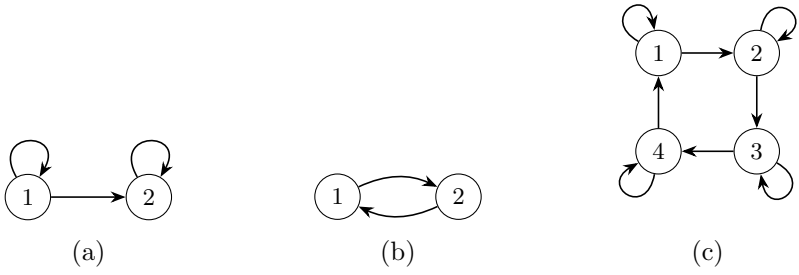

FIGURE 1. Support digraphs illustrating the three regimes of Section 5. (a) Dual numbers in the basis $(1, \varepsilon)$: no edge leaves vertex 2, so $\mathbb{R}\varepsilon$ is a coordinate ideal and no depth yields universality (Theorem 5.2(a)); the real coordinate receives edges only from itself, and its blind subspace $\mathbb{R}\varepsilon$ never shrinks. (b) The algebra of Remark 5.3: a directed 2-cycle without self-loops. Although the digraph is strongly connected, the blind subspaces $\mathbb{R}e_1$ and $\mathbb{R}e_2$ are interchanged at every layer and never vanish. (c) Example 5.4 for $d = 4$: a directed cycle with a self-loop at every vertex. At depth k, the blind subspace $N_j^{(k)}$ is defined by the vanishing of the $k+1$ coordinates $j-k, \ldots, j$, and therefore first vanishes at depth $3 = d-1$.

Figure 1 illustrates these three regimes. The obstruction in Theorem 5.2(a) admits a transparent geometric interpretation: if L is a coordinate left ideal, then every affine $\mathcal{A}$-map sends $x + L^n$ into $(Wx + b) + L^m$ and every split activation sends $x + L$ into $\sigma_{\mathcal{E}}(x) + L$. Consequently, every network F of arbitrary depth satisfies $F(x+y) - F(x) \in L$ whenever $y \in L^n$, meaning that the output modulo L depends only on the input modulo L. The next proposition shows that coordinate subspaces are the only subspaces compatible with split activations, so that among left ideals, the coordinate left ideals are precisely those whose cosets are respected by split-activation networks.

Proposition 5.5. *Let $\sigma \in C(\mathbb{R})$ be non-affine and let $L \subseteq \mathcal{A}$ be a subspace. Then $\sigma_{\mathcal{E}}(x + y) - \sigma_{\mathcal{E}}(x) \in L$ for all $x \in \mathcal{A}$ and $y \in L$ if and only if L is a coordinate subspace.*

Proof. If $L = \operatorname{span}\{e_i : i \in T\}$ and $y \in L$, then $\sigma_{\mathcal{E}}(x + y) - \sigma_{\mathcal{E}}(x) = \sum_{i \in T}[\sigma(x_i + y_i) - \sigma(x_i)]e_i \in L$. Conversely, suppose $y \in L$ has $y_j \neq 0$. If the difference $\tau \mapsto \sigma(s + \tau) - \sigma(s)$ were independent of s for every $\tau \in \mathbb{R}$, then $\sigma(s + \tau) = \sigma(s) + \sigma(\tau) - \sigma(0)$, and the only continuous solutions to this Cauchy functional equation are affine. Hence there exist $s, s', \tau \in \mathbb{R}$ such that $\sigma(s + \tau) - \sigma(s) \neq \sigma(s' + \tau) - \sigma(s')$; setting $t = \tau/y_j \in \mathbb{R}$ gives $\tau = ty_j$. Choosing $x, x' \in \mathcal{A}$ that agree in all coordinates except the j-th, where $x_j = s$ and $x'_j = s'$, and noting that $ty \in L$, we see that the difference $[\sigma_{\mathcal{E}}(x+ty)-\sigma_{\mathcal{E}}(x)]-[\sigma_{\mathcal{E}}(x'+ty)-\sigma_{\mathcal{E}}(x')]$ is a nonzero multiple of e_j belonging to L. Thus $e_j \in L$ whenever some $y \in L$ satisfies $y_j \neq 0$, so L is a coordinate subspace. □

Corollary 5.6. *Let $\mathcal{A}$ be an algebra with identity and let $\sigma \in C(\mathbb{R})$ be non-polynomial. For a nonempty Zariski-open set of bases, $G(\mathcal{A}, \mathcal{E})$ is the complete digraph and $\mathcal{N}_2$ is universal. If, moreover, $\mathcal{A}$ is associative with identity and not a Frobenius algebra, then for all such bases $\mathcal{N}_2$ is universal whereas $\mathcal{N}_1$ is not, and $\mathcal{N}_1$ is not universal for any basis of $\mathcal{A}$.*

Proof. The set U of elements $x \in \mathcal{A}$ for which right multiplication $w \mapsto wx$ is bijective is Zariski-open and contains the identity; therefore, the bases with $e_i \in U$ for all i form a nonempty Zariski-open set. When every $e_i \in U$, we have $\mathcal{A}e_i = \mathcal{A} \not\subseteq \ker e_j^*$ for all i, j, so $G(\mathcal{A}, \mathcal{E})$ is complete with $\operatorname{diam} G(\mathcal{A}, \mathcal{E}) \leq 1$, and Theorem 5.2(b) implies universality at depth two. The second assertion follows directly from Theorem 4.4(a). □

Remark 5.7 (Using the criteria in practice). For a given multiplication table and basis, Theorems 4.1– 5.2 reduce universality testing to finite linear-algebraic and graph-theoretic computations on the structure constants $e_j^*(e_a e_b)$: the ranks of the d slice matrices B_j determine $S(\mathcal{E})$, the span of the vectors $e_a e_j$ for $j \in S(\mathcal{E})$ tests (4.3), and the support digraph together with the recursion (5.1) determines universality at each depth. Three practical consequences stand out. For a Frobenius algebra, a basis that fails at depth one can be repaired by a generic change of basis (Theorem 4.4(b)). For an

associative non-Frobenius algebra with identity, no change of basis yields universality at depth one, but two hidden layers suffice in almost every basis (Corollary 5.6). Finally, for a learned multiplication rule [Zhang et al., 2021], a structure tensor in general position has non-singular slices and bijective right multiplications by every basis vector, so single-hidden-layer networks are universal with either real or algebra-valued output weights; the exceptions lie on the algebraic set where either $\det B_j$ or the determinant of right multiplication by e_j vanishes for some j.

6. Examples

We now apply the criteria of Sections 4 and 5 to the hypercomplex algebras most widely used in practice, as well as to a minimal non-Frobenius algebra that makes the depth separation explicit.

Example 6.1 (Dual numbers and dual quaternions). Let $\mathbb{D} = \mathbb{R}[\varepsilon]/(\varepsilon^2)$. In the standard basis $(1, \varepsilon)$, the subspace $\operatorname{span}\{\varepsilon\}$ is a coordinate ideal, so by Theorem 5.2(a) split-activation networks over $\mathbb{D}$ in this basis fail to be universal at any depth: the real part of every network output depends only on the real parts of the inputs. In the basis $(1, 1+\varepsilon)$, by contrast, the coordinate functionals are given by $e_1^*(a+b\varepsilon) = a - b$ and $e_2^*(a+b\varepsilon) = b$; neither kernel contains $\operatorname{span}\{\varepsilon\}$, so both slices are non-degenerate, and Theorems 4.1 and 4.2 guarantee universality at depth one. Valle et al. [2024] observed an analogous change in behavior using the basis $(1+\varepsilon, 1-\varepsilon)$, which defines their algebra $\mathbb{E}$. Similarly, for the dual quaternions $\mathbb{H}\otimes\mathbb{D}$ in the standard basis, $\varepsilon\mathbb{H}$ is a coordinate two-sided ideal, so the non-dual part of every output depends only on the non-dual part of the input at every depth.

Proposition 6.2 (Clifford algebras). *Let $\mathrm{Cl}(p,q,r)$ be the Clifford algebra generated by $g_1, \ldots, g_n$ ($n = p+q+r$), where p generators satisfy $g_i^2 = 1$, q satisfy $g_i^2 = -1$, the remaining r satisfy $g_i^2 = 0$, and $g_i g_j = -g_j g_i$ for $i \neq j$. Let $\mathcal{E}$ be the blade basis consisting of $g_A = g_{a_1} \cdots g_{a_s}$ for $A = \{a_1 < \cdots < a_s\} \subseteq \{1, \ldots, n\}$.*

- *(a) The coefficient φ of the pseudoscalar $g_{\{1,\ldots,n\}}$ is a Frobenius form. Hence every $\mathrm{Cl}(p,q,r)$ is a Frobenius algebra and, by Theorem 4.4(b), admits bases for which $\mathcal{N}_1$ is universal.*
- *(b) If $r \geq 1$, the blades containing at least one generator of square zero span a proper nonzero coordinate two-sided ideal, so in the blade basis $\mathcal{N}_k$ is not universal for any $k \geq 1$.*

Proof. (a) For any two blades g_A and g_B, we have $g_A g_B = \pm\big(\prod_{i\in A\cap B} g_i^2\big) g_{A\triangle B}$. The pseudoscalar coefficient of this product is nonzero only when $A\triangle B = \{1, \ldots, n\}$, which forces $A \cap B = \emptyset$ and $B = A^c$, in which case the coefficient equals ± 1. Thus the matrix $\big(\varphi(g_A g_B)\big)$ has exactly one nonzero entry, equal to ± 1, in each row and column. It is therefore

non-singular, so the associative bilinear form $(w, x) \mapsto \varphi(wx)$ is non-degenerate and φ is a Frobenius form [Skowroński and Yamagata, 2011, Theorem IV.2.1].

(b) Let Z be the set of indices of generators satisfying $g_i^2 = 0$, and put $I = \operatorname{span}\{g_A : A \cap Z \neq \emptyset\}$. If $i \in A \cap Z$, then $g_A g_B = 0$ whenever $i \in B$ (since a factor $g_i^2 = 0$ appears), whereas $g_A g_B$ is a scalar multiple of $g_{A \triangle B}$ with $i \in A \triangle B$ when $i \notin B$; the same holds for $g_B g_A$. Hence I is a two-sided ideal that is proper ($1 \notin I$) and nonzero, and Theorem 5.2(a) applies. □

Proposition 6.2(b) rigorously confirms, for the blade basis and at arbitrary depth, the assertion of Buchholz and Sommer [2008, Section 8] that degenerate Clifford algebras do not yield universal approximation, while part (a) demonstrates that this failure is an artifact of the blade basis rather than an intrinsic obstruction of the algebra. In particular, the obstruction applies to the projective geometric algebra $\mathrm{Cl}(3, 0, 1)$ whenever split activations are employed; architectures such as the geometric algebra transformer [Brehmer et al., 2023] rely on different equivariant layer operations and therefore fall outside this setting.

Example 6.3 (Group algebras). For a finite group Γ equipped with the group basis of $\mathbb{R}[\Gamma]$, we have $e_g^*(xy) = \sum_{hk=g} x_h y_k$, so every slice matrix is a permutation matrix. All slices are therefore non-degenerate, and both $\mathcal{N}_1$ and $\mathcal{N}_1^{\mathbb{R}}$ are universal. When Γ is cyclic, the weight matrices are circulant, and the scalar-output case is contained in Zhao et al. [2017, Theorem 4.2 and Corollary 4.1]. If biases are restricted to scalar multiples of $\sum_{g \in \Gamma} g$, the resulting networks are equivariant under right multiplication by Γ; universality for equivariant target functions is treated by Ravanbakhsh [2020].

The next proposition makes the depth separation of Corollary 5.6 quantitative.

Proposition 6.4 (An explicit depth separation). *Let $\mathcal{A} = \mathbb{R}[x, y]/(x, y)^2$ with basis $\mathcal{E} = (1, 1 + x, 1 + y)$, take one input ($n = 1$), and write inputs as $u = a + bx + cy$. Let $K = \{u \in \mathcal{A} : a, b, c \in [-1, 1]\}$ be equipped with the uniform probability measure, and define $F(u) = bc(b - c) \cdot 1$.*

(a) $\mathcal{A}$ is not a Frobenius algebra, so $\mathcal{N}_1$ is not universal for any basis; however, $G(\mathcal{A}, \mathcal{E})$ is complete, so $\mathcal{N}_2$ is universal.

(b) For every $H \in \mathcal{N}_1(1, 1)$,

$$\int_K \big(e_1^*(F(u) - H(u))\big)^2 \, du \;\geq\; \frac{8}{135} = \frac{4}{9} \operatorname{Var}\big(e_1^*(F)\big),$$

and the constant $8/135$ is the exact infimum over $\mathcal{N}_1(1, 1)$.

Proof. (a) The subspace $\mathfrak{m} = \operatorname{span}\{x, y\}$ satisfies $\mathfrak{m}^2 = 0$ and $(a + bx + cy)v = av$ for all $v \in \mathfrak{m}$, so every linear subspace of $\mathfrak{m}$ is an ideal of $\mathcal{A}$. Because the kernel of any linear functional on $\mathcal{A}$ intersects the two-dimensional subspace $\mathfrak{m}$ nontrivially, it contains a nonzero ideal; hence no Frobenius form exists, and Theorem 4.4(a) applies. On the other hand, each element of $\mathcal{E}$ is a unit

of the commutative algebra $\mathcal{A}$, so right multiplication by each basis vector is bijective. As in the proof of Corollary 5.6, $G(\mathcal{A}, \mathcal{E})$ is complete, and Theorem 5.2(b) gives universality at depth two.

(b) In the standard coordinates (a, b, c), the dual basis functionals are given by $e_1^*(u) = a - b - c$, $e_2^*(u) = b$, and $e_3^*(u) = c$, and (2.2) yields $N_1 = \mathbb{R}(y - x)$, $N_2 = \mathbb{R}y$, and $N_3 = \mathbb{R}x$. Since $\mathcal{A}e_j = \mathcal{A}$ for every j, Theorem 4.1 shows that the first coordinates of shallow networks are uniformly dense in

$$\mathcal{H} = \{g_1(a, b + c) + g_2(a, b) + g_3(a, c) : \ g_1, g_2, g_3 \in C([-2, 2]^2)\}.$$

Set $f(b, c) = bc(b - c)$ and $r(b, c) = f(b, c) + \frac{1}{3}(b - c) = (b - c)(bc + \frac{1}{3})$. Direct integration gives $\int_{-1}^{1} r(b, c)\, dc = 0$ for every $b \in [-1, 1]$ and $\int_{-1}^{1} r(b, c)\, db = 0$ for every $c \in [-1, 1]$; moreover, r integrates to zero along every segment $\{b + c = s\} \cap [-1, 1]^2$ because $r(c, b) = -r(b, c)$ and the segment is symmetric under exchanging b and c. By Fubini's theorem, r is orthogonal in $L^2(K)$ to every function in $\mathcal{H}$. Since $f - r = -\frac{1}{3}(b - c) \in \mathcal{H}$, the orthogonal projection of f onto the $L^2(K)$-closure $\overline{\mathcal{H}}$ is $-\frac{1}{3}(b - c)$, and the squared distance equals $\int_K r^2\, du = \frac{8}{135}$, whereas $\mathrm{Var}(f) = \frac{2}{15}$. □

Proposition 6.4(b) bounds every single-hidden-layer network regardless of its width or continuous non-polynomial activation, whereas by part (a) networks with two hidden layers approximate F arbitrarily closely. Section 7 corroborates this separation numerically.

Remark 6.5 (A counterexample). Niemczynowicz and Kycia [2026, Definition 1] define an algebra to be non-degenerate if either $xy = 0$ for all y, or $yx = 0$ for all y, implies $x = 0$. Every algebra with identity satisfies this condition. Their Theorem 3 asserts that, for any such algebra V, every continuous map $g\colon X \to V$ on a compact set $X \subset V^N$ can be uniformly approximated by sums $\sum_k \alpha_k f(\langle v_k | x\rangle + \Theta_k)$ with real coefficients α_k, componentwise sigmoid activation f, and $\langle v | x\rangle = \sum_l v_l x_l$. Taking $V = \mathbb{D}$ in the basis $(1, \varepsilon)$, $N = 1$, $X = \{x_0 + x_1\varepsilon : \ |x_0|, |x_1| \le 1\}$, and $g(x) = x_1 \cdot 1$, the real part of $vx + \Theta$ is $v_0 x_0 + \Theta_0$. Consequently, the real part of every approximant depends on x_0 alone, and the uniform approximation error in the real coordinate is at least 1. The step in the proof of their Lemma 2 that deduces $v_i = 0$ from $\pi_i(\langle v_i | y\rangle) = 0$ for all y implicitly requires the i-th slice to be non-degenerate, which fails for $i = 0$. Theorem 4.2 characterizes precisely the algebras and bases for which their conclusion holds.

7. Numerical experiments

Our theoretical results apply to networks of arbitrary width. The computations in this section evaluate the algebraic criteria across a broad collection of algebras and bases, and test whether trained networks of moderate width attain the predicted error floors.

The criteria of Theorems 4.1, 4.2, 5.1, and 5.2 reduce to finite computations on the structure constants $e_j^*(e_a e_b)$: computing the ranks of the

slice matrices B_j, the spans of the vectors $e_a e_j$, the support digraph, and the recursion (5.1). We implemented these procedures in Python and evaluated them on 32 algebra–basis pairs, a representative selection of which is reported in Table 1.

TABLE 1. Universality criteria for selected algebras and bases. "All slices" denotes the condition of Valle et al. [2024]; "Depth one" denotes criterion (4.3); "Some depth" denotes strong connectivity of $G(\mathcal{A}, \mathcal{E})$. For the non-associative octonions and sedenions, "Frobenius" indicates that some linear form φ renders $(w, x) \mapsto \varphi(wx)$ non-degenerate.

Algebra, basis	Frobenius	All slices	Depth one	Some depth	Least depth
$\mathbb{C}$, $\mathbb{H}$, octonions, sedenions	yes	yes	yes	yes	1
Split-complex numbers, standard	yes	yes	yes	yes	1
Split-complex numbers, idempotent	yes	no	no	no	–
$\mathbb{D}$, basis $(1, \varepsilon)$	yes	no	no	no	–
$\mathbb{D}$, basis $(1, 1+\varepsilon)$	yes	yes	yes	yes	1
Tessarines, standard / idempotent	yes	yes / no	yes / no	yes / no	1 / –
Dual-complex numbers, dual quaternions	yes	no	no	no	–
$\mathrm{Cl}(3,0,0)$, $\mathrm{Cl}(1,3,0)$, blade basis	yes	yes	yes	yes	1
$\mathrm{Cl}(2,0,1)$, $\mathrm{Cl}(3,0,1)$, blade basis	yes	no	no	no	–
$\mathbb{R}[\Gamma]$, $\Gamma \in \{C_2, C_3, C_4, S_3\}$, group basis	yes	yes	yes	yes	1
$T_2(\mathbb{R})$, matrix units / random basis	no	no	no	no / yes	– / 2
$\mathbb{R}[x,y]/(x,y)^2$, monomials / random basis	no	no	no	no / yes	– / 2
$\mathbb{R}[x,y]/(x^2,y^2)$, monomials / random basis	yes	no / yes	no / yes	no / yes	– / 1

Across 2,662 random bases with entries in $\{-1, 0, 1\}$ drawn for the 26 unital algebras of dimensions 2 through 8 in our catalog, the two characterizations of deep universality in Theorem 5.2—strong connectivity of $G(\mathcal{A}, \mathcal{E})$ and termination of the recursion (5.1)—agreed in every instance. Criterion (4.3) held in 426 bases that possessed at least one singular slice, confirming that Theorem 4.1 covers a substantial class of bases outside the hypothesis of Valle et al. [2024]. Depth two or greater was necessary in 259 bases, including bases of the Frobenius algebras $M_2(\mathbb{R})$, $\mathbb{R}[C_4]$, and $\mathbb{R}[S_3]$, with one basis of $\mathbb{R}[S_3]$ requiring depth three.

We also trained neural networks to verify the predicted error floors empirically. All models use left multiplication by algebra-valued weights, split tanh activations, and algebra-valued output weights unless noted otherwise, and were trained using full-batch Adam with a cosine learning-rate schedule for 4,000 steps (3,000 in E4) on 4,096 training points (8,192 in E2). Reported errors are mean squared errors on 4,096 independent test points (8,192 in E2), averaged over the seeds indicated below and measured in the standard coordinates of the algebra, except in E2, where the error is evaluated in the first coordinate $e_1^*(u) = a - b - c$ with respect to the basis $\mathcal{E}$, as in Proposition 6.4(b).

E1. Dual numbers, one hidden layer of width 64, three seeds, with target $x_0 + x_1\varepsilon \mapsto \sin(2x_1) + \cos(x_0 + x_1)\varepsilon$ on $[-1, 1]^2$. In the basis $(1, \varepsilon)$, the real part of every network depends only on x_0, so its real-coordinate error is at least $\mathrm{Var}(\sin 2x_1) = \frac{1}{2} - \frac{\sin 4}{8}$.

E2. The depth-separation setting of Proposition 6.4, with 8,192 training and test points, the target scaled to unit variance, and two seeds.
E3. The setting of Remark 6.5: dual numbers in the basis $(1, \varepsilon)$, real output weights, width 128, three seeds, and target x_1 in the real coordinate, giving an error floor of $\mathrm{Var}(x_1) = \frac{1}{3}$.
E4. Same setup as E1 in the basis $(1, \varepsilon)$, using three hidden layers of width 32 and one seed.

TABLE 2. Test errors compared against the theoretical floors proved in Sections 4–6. "Floor" is the infimum of the population error over the network class. "Optimal" is the error, evaluated on the same test points, of the function attaining this infimum (for E2, the correspondingly scaled projection $-\frac{1}{3}(b - c)$ from Proposition 6.4); it differs from the population floor only by finite-sample test variance.

Experiment	Network	Floor	Optimal	Test error
E1	$\mathbb{D}$, basis $(1, \varepsilon)$, real coordinate	0.5946	0.5896	0.5902
E1	$\mathbb{D}$, basis $(1, 1 + \varepsilon)$, real coordinate	0	0	0.00066
E2	One hidden layer, width 64	$4/9 \approx 0.4444$	0.4421	0.4434
E2	One hidden layer, width 256	$4/9 \approx 0.4444$	0.4421	0.4435
E2	Two hidden layers, widths 32–32	0	0	0.0013
E2	Two hidden layers, widths 64–64	0	0	0.0008
E3	Real output weights, real coordinate	1/3	0.3281	0.3286
E4	Three hidden layers, real coordinate	0.5946	0.5896	0.5909

Table 2 shows that every network subject to a theoretical floor approaches the empirical error of the optimal function to within 0.0015 without falling below it, whereas all architectures predicted to be universal approximate their targets closely. In particular, increasing depth does not overcome the coordinate-ideal obstruction for dual numbers in the basis $(1, \varepsilon)$ (E4), while a single change of basis removes it completely (E1). Conversely, adding a second hidden layer breaks the shallow error floor of Proposition 6.4 by more than two orders of magnitude (E2).

8. Conclusion and open problems

Universal approximation by split-activation networks over a finite-dimensional real algebra is governed by two ideal-theoretic invariants of the chosen basis: the Frobenius coordinates, which control single-hidden-layer networks, and the left ideals spanned by basis vectors, which—whenever the algebra has a left identity—constitute the only obstructions at arbitrary depth. Both invariants are directly computable from the multiplication table. This framework accounts for the basis dependence observed by Valle et al. [2024]: for a Frobenius algebra, non-universality disappears in almost every basis, whereas for an associative non-Frobenius algebra with identity it persists at depth one in every basis and disappears at depth two in

almost every basis. The dual numbers introduced in Section 1 illustrate both mechanisms: the standard basis $(1, \varepsilon)$ contains the coordinate ideal $\mathbb{R}\varepsilon$, whereas in $(1, 1+\varepsilon)$ both coordinates are Frobenius coordinates.

Several directions remain open for future investigation.

1. *Least universal depth.* In dimension $d \geq 2$, Theorem 5.2 bounds the least universal depth by $\min\{1+\operatorname{diam} G(\mathcal{A}, \mathcal{E}),\ d-1\}$, and Example 5.4 shows that $d-1$ is attained among non-associative algebras with a left identity. For an associative algebra with identity, generic bases require depth at most 2 (Corollary 5.6), and in our computations no basis required depth exceeding 3. What is the largest least universal depth achievable over all bases of a given associative algebra?
2. *Non-split activations.* For activations defined natively through the algebra multiplication, such as convergent power series $f(x) = \sum_n c_n x^n$, every ideal I of a commutative algebra is compatible in the sense that $x - y \in I$ implies $f(x) - f(y) \in I$, so the quotient obstruction applies to every proper nonzero ideal. Characterizing universal approximation for general activations would extend the theorem of Voigtlaender [2023] from the complex numbers to arbitrary finite-dimensional algebras.
3. *Weight conventions.* Replacing left multiplication by right multiplication substitutes right ideals for left ideals, while two-sided layers $x \mapsto \sum_k a_k x b_k$ lead naturally to sub-bimodules. Determining the corresponding criteria and characterizing when these conventions yield distinct expressive power remain open.
4. *Approximation rates.* How do quantitative approximation rates depend on the algebraic structure of $\mathcal{A}$—for instance, through its Wedderburn–Malcev decomposition into a semisimple part and the Jacobson radical?
5. *Learned and equivariant architectures.* For learned multiplication rules [Zhang et al., 2021], the non-universal structure tensors form an explicit algebraic subset of the parameter space. For group algebras with biases constrained to enforce equivariance, the target class consists of equivariant maps [Ravanbakhsh, 2020], and the role of coordinate ideals in that equivariant setting remains open.

A further application arises in hypercomplex transfer learning: Vieira Neto and Valle [2026] embed real-valued layers into hypercomplex layers under the assumption that all slices are non-degenerate (their Section 2.1), yet their construction inverts only the matrix of the real-part slice (their Eq. (12)). In the terminology of this paper, that embedding requires only that the real coordinate be a Frobenius coordinate, suggesting that the present ideal-theoretic framework may sharpen the theory of hypercomplex transfer learning as well.

CRediT authorship contribution statement

Muhammad Saad: Conceptualization, Methodology, Formal analysis, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No external datasets were used in this study. All Python code reproducing the algebraic computations, symbolic verifications, and numerical experiments is publicly available at `https://github.com/MuhammadSaad87/hvnn-universal-approximation`.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Paolo Arena, Luigi Fortuna, Giovanni Muscato, and Maria Gabriella Xibilia. Multilayer perceptrons to approximate quaternion valued functions. *Neural Networks*, 10(2):335–342, 1997. doi: 10.1016/S0893-6080(96)00048-2.

Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sven Buchholz and Gerald Sommer. On Clifford neurons and Clifford multilayer perceptrons. *Neural Networks*, 21(7):925–935, 2008. doi: 10.1016/j.neunet.2008.03.004.

Pim de Haan, Taco Cohen, and Johann Brehmer. Euclidean, projective, conformal: choosing a geometric algebra for equivariant transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3088–3096, 2024.

Zhe-Cheng Fan, Tak-Shing T. Chan, Yi-Hsuan Yang, and Jyh-Shing Roger Jang. Backpropagation with N-D vector-valued neurons using arbitrary bilinear products. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2638–2652, 2020. doi: 10.1109/TNNLS.2019.2933882.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Vladimir Ya. Lin and Allan Pinkus. Fundamentality of ridge functions. *Journal of Approximation Theory*, 75:295–311, 1993. doi: 10.1006/jath.1993.1104.

Agnieszka Niemczynowicz and Radosław Antoni Kycia. Fully tensorial approach to hypercomplex-valued neural networks. *Information Sciences*, 755:123796, 2026. doi: 10.1016/j.ins.2026.123796.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999. doi: 10.1017/S0962492900002919.

Călin-Adrian Popa. Octonion-valued neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2016, Part I*, volume 9886 of *Lecture Notes in Computer Science*, pages 435–443. Springer, 2016. doi: 10.1007/978-3-319-44778-0_51.

Siamak Ravanbakhsh. Universal equivariant multilayer perceptrons. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7996–8006, 2020.

Andrzej Skowroński and Kunio Yamagata. *Frobenius Algebras I: Basic Representation Theory*. EMS Textbooks in Mathematics. European Mathematical Society, Zürich, 2011.

Marcos Eduardo Valle, Wington L. Vital, and Guilherme Vieira. Universal approximation theorem for vector- and hypercomplex-valued neural networks. *Neural Networks*, 180:106632, 2024. doi: 10.1016/j.neunet.2024.106632.

Guilherme Vieira Neto and Marcos Eduardo Valle. Ghost features and spooky transfer learning for hypercomplex-valued neural networks. arXiv:2608.07735, 2026.

Wington L. Vital, Guilherme Vieira, and Marcos Eduardo Valle. Extending the universal approximation theorem for a broad class of hypercomplex-valued neural networks. In *Intelligent Systems (BRACIS 2022)*, volume 13654 of *Lecture Notes in Computer Science*, pages 646–660, Cham, 2022. Springer. doi: 10.1007/978-3-031-21689-3_45.

Felix Voigtlaender. The universal approximation theorem for complex-valued neural networks. *Applied and Computational Harmonic Analysis*, 64:33–61, 2023. doi: 10.1016/j.acha.2022.12.002.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. Beyond fully-connected layers with quaternions: parameterization of hypercomplex multiplications with $1/n$ parameters. In *International Conference on Learning Representations*, 2021.

Liang Zhao, Siyu Liao, Yanzhi Wang, Zhe Li, Jian Tang, and Bo Yuan. Theoretical properties for neural networks with weight matrices of low displacement rank. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4082–4090, 2017.

Muhammad Saad
Department of Mathematics and Computer Science, Faculty of Science,
Alexandria University, Alexandria 21568, Egypt
e-mail: m.saad@alexu.edu.eg