FINITE SAMPLE COMPLEXITY GUARANTEES FOR FEEDFORWARD NEURAL NETWORKS VIA O-MINIMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove that, in a precise sense, a broad class of feedforward neural networks (FFNN) learn (have finite sample complexity) in the PAC model: every fixed finite feedforward architecture whose layers are definable in an o-minimal structure has finite sample complexity in the agnostic PAC setting, even with unbounded parameters. Quantitative versions are obtained for general FFNNs, possibly using *multivariate* non-linearities, which are Pfaffian arranged in a general computational graph (DAG). This covers standard fixed-size MLPs, CNNs, GNNs, and transformers with fixed sequence length, together with the operations and layers typically used in such architectures, including linear projections, residual connections, attention mechanisms, pooling layers, normalization layers, and admissible positional encodings. Hence, distribution-free learnability for modern non-recurrent architectures is not an exceptional property of particular activations or architecture-specific VC arguments, but a consequence of tame feedforward computation. Our results reposition finite-sample PAC learnability as a baseline rather than a differentiator: they shift the focus of architectural comparison toward inductive biases, symmetries and geometric priors, scalability, and optimization behaviour.

1 INTRODUCTION

Modern neural architectures appear strikingly diverse, ranging from classical multilayer perceptrons (MLPs) to graph neural networks (GNNs) and contemporary transformers. Each major architectural innovation introduces new layers, symmetries, and "basic computational units," and has, in turn, sparked a corresponding line of architecture-specific statistical theory proving that the resulting model class can learn. Examples include guarantees for piecewise-linear MLPs (Bartlett et al., 2019), message-passing GNNs (Garg et al., 2020; Morris et al., 2023), and transformer architectures (Edelman et al., 2022; Trauger & Tewari, 2024; Limmer et al., 2024).

From the viewpoint of computation, however, these architectures share a common form. At inference time, they are finite feedforward programs: directed acyclic computational graphs whose nodes execute elementary parametrized operations; cf. circuit-complexity theory (Jukna, 2012; 2023; Lloris Ruiz et al., 2014). This observation suggests a simple question.

> *Do all standard fixed finite feedforward neural networks have finite sample complexity?*

We answer this question *affirmatively* for the broad class of feedforward architectures whose primitive operations are definable in an o-minimal structure. In simple terms: *modern artificial neural networks are mathematically well-behaved enough to be theoretically learnable.* More precisely, we prove that every fixed finite feedforward architecture assembled from definable layers/gates and arranged in a directed acyclic computational graph has finite agnostic PAC sample complexity for both classification (Theorem 3.1) and regression (Theorem 3.2). This holds even when their parameter space is unbounded.

Throughout the paper, a "fixed network" or "fixed architecture" means that the computational graph, input and output dimensions, widths, depths, sequence lengths, and the number and placement of trainable parameters are fixed. The numerical values of the parameters themselves are not fixed; they are allowed to vary over their prescribed definable parameter spaces (free to vary during training).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We verify that the definability assumption is mild by checking it for the standard fixed finite building blocks of MLPs, CNNs, GNNs, and transformers, including common activations, convolution, attention, normalization, pooling, and admissible positional encodings. Here, learnability is statistical and distribution-free: uniform convergence yields finite agnostic PAC sample complexity through Borel measurable approximate empirical risk minimization; see Lemma B.1.

Takeaways. Our results complement the ongoing work in the foundations of statistical learning theory seeking to characterize the exact conditions under which PAC learning is possible by uncovering sharp learning-theoretic dimensions; see, e.g., Attias et al. (2023); Bressan et al. (2025). Here, we demonstrate that most architectures satisfy the well-known necessary conditions of finite fat-shattering and finite VC dimension, whose sufficiency dates back at least to Alon et al. (1997). Thus, our *theoretical takeaway* is a simple, virtually always guaranteed **finite-sample-complexity** result under a unified framework, relying on clarified connections between fat-shattering dimension Bartlett et al. (1997), VC dimension, and o-minimal structures in *mathematical logic* Pillay & Steinhorn (1986), particularly model theory van den Dries (1998); Coste (1999); Speissegger (1999).

The *practical takeaway* is that finite sample complexity is too coarse to distinguish standard feedforward architectures. Once a model is fixed, finite, feedforward, and built from definable primitives, distribution-free learnability follows automatically. Thus, model selection should be guided by finer criteria: problem-specific inductive bias, symmetry and geometric priors, as studied in geometric deep learning Bronstein et al. (2016; 2021); de Ocáriz Borde & Bronstein (2025), as well as scalability Kaplan et al. (2020) and optimization behavior.

2 Background and Preliminaries

One of the most important bridges between mathematical logic and learning theory is shattering. In PAC learning, finite VC dimension is the fundamental combinatorial certificate for distribution-free learnability (Blumer et al., 1989; Anthony & Bartlett, 1999); in model theory, the same phenomenon appears as the independence property for uniformly definable families (Laskowski, 1992; Chase & Freitag, 2019). Early neural-network theory already exploited this connection: tame real geometry was used to prove finiteness results for sigmoidal networks (Macintyre & Sontag, 1993), polynomial VC bounds for Pfaffian networks (Karpinski & Macintyre, 1997), and general bounds for real-parametrized concept classes generated by elementary operations (Goldberg & Jerrum, 1995); see also (Bartlett & Maass, 2003). Our work revisits this logic-learning bridge in the language of modern architectures: a fixed feedforward network is a finite composition of parametrized gates, so joint definability of the gates propagates to the whole model, while o-minimal tameness rules out unbounded shattering. Thus, model theory supplies a clean statistical certificate: definable feedforward computation has finite sample complexity.

2.1 Notation and Terminology

We begin by fixing the notation and terminology. We denote the indicator function of a set A by I_A. We also use I to convert Boolean variables into $\{0,1\}$-valued variables, so that $I(\text{True}) = 1$ and $I(\text{False}) = 0$. An interval means a set of one of the forms $(a,b), (-\infty,a)$, or (b,∞), where $a,b \in \mathbb{R}$. For every $n \in \mathbb{N}_+$, we equip $\mathbb{R}^n$ with its usual Euclidean topology.

Definition 2.1 (Directed Acyclic Graph (DAG))**.** A *directed graph* is a pair $G = (V,E)$, where V is a non-empty set of *nodes* and $E \subseteq V^2 \setminus \{(v,v)\}_{v\in V}$ is a set of directed edges. A *directed path* from v to w is a finite sequence of directed edges $(v,v_1),(v_1,v_2),\ldots,(v_k,w) \in E$. The graph G is called *acyclic* if there is no non-empty directed path from any node to itself.

2.2 Background in Mathematical Logic: O-Minimality

We now review the background required to formulate our results. Standard learning-theoretic notions used later, such as VC dimension and pseudo-dimension, are recalled in Appendix A.2. We begin with the basic language of o-minimal geometry; see, for instance, (van den Dries, 1998; Coste, 1999). The central object is a structure on the real field whose definable subsets form a tame class of sets: stable under the usual logical and geometric operations.

Definition 2.2 (O-Minimal Structure Expanding the Real Field)**.** A *structure expanding the real field* is a sequence $\mathcal{S} = (S_n)_{n\in\mathbb{N}_+}$, where each S_n is a collection of subsets of $\mathbb{R}^n$, satisfying the following axioms:

(i) Every semi-algebraic subset of $\mathbb{R}^n$ belongs to S_n.

(ii) For every $n \in \mathbb{N}_+$, the collection S_n is a Boolean subalgebra of $\mathcal{P}(\mathbb{R}^n)$.

(iii) If $A \in S_m$ and $B \in S_n$, then $A \times B \in S_{m+n}$.

(iv) If $p : \mathbb{R}^{n+1} \to \mathbb{R}^n$ denotes the projection onto the first n coordinates and $A \in S_{n+1}$, then $p(A) \in S_n$.

The elements of S_n are called the *$\mathcal{S}$-definable subsets* of $\mathbb{R}^n$. The structure $\mathcal{S}$ is called *o-minimal* if *every set in S_1 is a finite union of points and intervals*.

Example 1 (Semi-Algebraic + (Global) Exponential). The real exponential field $\mathbb{R}_{\exp} \overset{\text{def.}}{=} (\mathbb{R}, <, +, \cdot, 0, 1, \exp)$ is o-minimal by Wilkie (1996).

More broadly, we may add in all restricted analytic functions.

Example 2 (Restricted Analytic Functions + (Global) Exponentials). The structure $\mathbb{R}_{\text{an,exp}}$, obtained by adding all restricted analytic functions and the global exponential map, is also o-minimal and contains the standard analytic operations used in softmax-type layers by van den Dries & Miller (1996).

Furthermore, Pfaffian functions on open $\mathcal{S}$-definable domains are definable in the o-minimal Pfaffian closure $\mathcal{P}(\mathcal{S})$. Briefly, Pfaffian functions form a robust class of tame functions: their graphs may bend, grow, and interact non-linearly, but cannot oscillate with unrestricted complexity (cf. Appendix A.1 for a precise definition of Pfaffianity). The implication of learnability for neural networks with Pfaffian activation functions was considered in Karpinski & Macintyre (1997) and more recently for GNNs in D'Inverno et al. (2025). The former results do not cover neural networks with multivariate non-linearities; eg. transformers, and the latter does not provide quantitative complexity quantifications based on the depth of the model. Neither of those can handle the elementary ReLU activation, and we note that Bartlett et al. (1998) can handle ReLU cannot handle ReLU combined with attention.

Example 3 (Pfaffian Closures). The Pfaffian closure of an o-minimal structure is obtained by adjoining the tame solution sets of definable Pfaffian differential systems. More precisely, the Pfaffian closure $\mathcal{P}(\mathcal{S})$ of any o-minimal expansion $\mathcal{S}$ of the real field is again o-minimal, and contains the Pfaffian functions generated from $\mathcal{S}$-definable C^1 data Speissegger (1999).

A function is well-behaved, in the o-minimal sense, if its graph is definable in the ambient structure.

Definition 2.3 (Definable Map)**.** Let $A \subset \mathbb{R}^n$. A map $f : A \to \mathbb{R}^p$ is called *definable* if its graph $\Gamma_f \overset{\text{def.}}{=} \{(x, f(x)) \in \mathbb{R}^{n+p} : x \in A\}$ is a definable subset of $\mathbb{R}^{n+p}$.

In our examples section, Section 4, we show that most neural networks are definable in the o-minimal structures above. We do so by verifying that the standard components used in deep learning are definable therein, and then using the fact that compositions of definable maps are again definable. Definable functions enjoy several closure properties: their linear combinations, coordinate-wise concatenations, and compositions are all again definable in $\mathcal{S}$; cf. (Coste, 1999, Chapter 1.3).

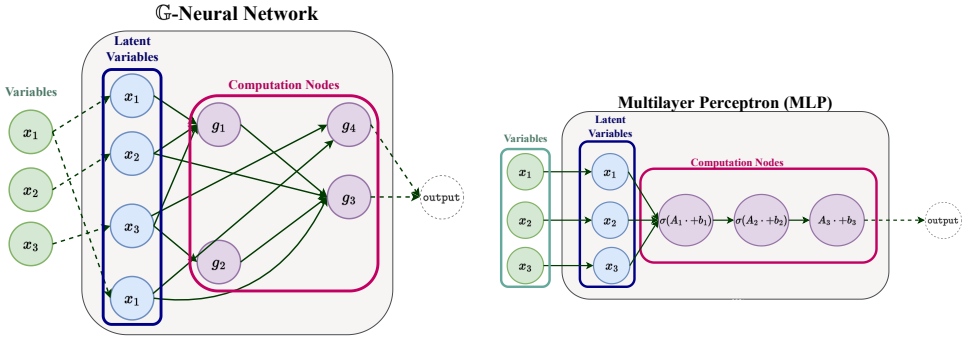
2.3 Definable (Feedforward) Neural Networks

As illustrated in Figure 1, feedforward neural-network computations are naturally organized by a DAG, in close analogy with the circuit-complexity literature, including Boolean circuits Jukna (2012); Chiang (2025); Li et al. (2026), tropical circuits Jukna (2023), probabilistic circuits Choi et al. (2020); Zhang et al. (2021), and algebraic circuits Wang et al. (2024); see also Kratsios et al. (2025). Here, we work with a broader class of feedforward computational graphs, subsuming most modern deep learning architectures as well as these semi-classical models of computational complexity theory.

We fix a family of finite-dimensional parametric maps, which will serve as the source of non-linearities in our deep learning models. We refer to this family as our *gate dictionary*

$$\mathbb{G} \subseteq \bigcup_{n,p,m=1}^{\infty} \left\{ g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m \right\}. \tag{2.1}$$

For any $\mathbb{G} \ni g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$, we interpret the first n coordinates as *inputs* and the last p coordinates as *parameters*, and we write $g_\theta \overset{\text{def.}}{=} g(\cdot, \theta)$ for every fixed parameter $\theta \in \mathbb{R}^p$.



(a) General feedforward $\mathbb{G}$-neural network *(b)* Special case: MLP

Figure 1: **$\mathbb{G}$-NN; cf. Definition 2.4 (left)**: The structure of a *general* non-recursive (feedforward) neural network ($\mathbb{G}$-NN). Parameterizable/trainable computations (gates) chosen from the *gate dictionary* $\mathbb{G}$ (2.1) are executed in the order specified by a DAG. **The "standardized" *MLP* special case (right):** The standard multilayer perceptron (MLP) is recovered as a special case. The MLP model arises when the computational DAG is a directed line graph and all gates are either of the form $\sigma(A \cdot + b)$ or with $A \cdot + b$ in the final layer.

Let $G = (V, E)$ be a DAG. For any $v \in V$, the *parents* of v are the nodes $\mathrm{Pa}_G(v) \overset{\text{def.}}{=} \{u \in V : (u, v) \in E\}$. Nodes without parents are called *input* nodes and are denoted by $\mathrm{in}(G)$; all non-input nodes are called *computation* nodes and are denoted by $\mathrm{comp}(G)$. Nodes without children are called *output* nodes and are denoted by $\mathrm{out}(G)$. Every DAG induces a natural partial order (*i.e., the direction of computation during the forward pass*) on its nodes: we write $u \lesssim v$ if either $u = v$ or there is a directed path from u to v; see, e.g., (Bang-Jensen & Gutin, 2008, Proposition 2.1.3).

Definition 2.4 ($\mathbb{G}$-Neural Network)**.** Fix dimensions $d_{in}, P, d_{out} \in \mathbb{N}_+$ and a gate dictionary $\mathbb{G}$, cf. (2.1). A map $f : \mathbb{R}^{d_{in}} \times \mathbb{R}^P \to \mathbb{R}^{d_{out}}$ is called a $\mathbb{G}$-*neural network* if there exist a finite DAG $G = (V, E)$, a binary lifting channel $\Pi \in \{0,1\}^{|\mathrm{in}(G)| \times d_{in}}$, gates $\mathbb{G} \ni g_v : \mathbb{R}^{M_v} \times \mathbb{R}^{p_v} \to \mathbb{R}^{m_v}$, $v \in \mathrm{comp}(G)$, and a linear readout matrix $A_{\mathrm{out}} \in \mathbb{R}^{d_{out} \times M_{\mathrm{out}}}$, such that f can be iteratively represented as follows. First, set $z^0 \overset{\text{def.}}{=} \Pi x$ for each $x \in \mathbb{R}^{d_{in}}$, and identify the coordinates of z^0 with the input nodes of G. Then, for every $v \in \mathrm{comp}(G)$, define the *latent variables*

$$z_v \overset{\text{def.}}{=} (g_v)_{\theta_v}\left((z_u)_{u \in \mathrm{Pa}_G(v)}\right), \tag{2.2}$$

where the parent states are concatenated in any fixed order compatible with the partial order on G, and $M_v \overset{\text{def.}}{=} \sum_{u \in \mathrm{Pa}_G(v)} m_u$ is the total dimension of the parent state. Finally, the *output*

$$f_\theta(x) \overset{\text{def.}}{=} A_{\mathrm{out}}(z_v)_{v \in \mathrm{out}(G)}, \tag{2.3}$$

where $M_{\mathrm{out}} \overset{\text{def.}}{=} \sum_{v \in \mathrm{out}(G)} m_v$. We write $\theta \overset{\text{def.}}{=} \left(A_{\mathrm{out}}, (\theta_v)_{v \in \mathrm{comp}(G)}\right) \in \mathbb{R}^P$, for its parameter vector, where $P = d_{out} M_{\mathrm{out}} + \sum_{v \in \mathrm{comp}(G)} p_v$. The *domain* of f, denoted $\mathrm{dom}(f)$, consists of all $x \in \mathbb{R}^{d_{in}}$ for which the compositions in (2.2) and (2.3) are well-defined.

The Key Insight: Joint Definability in Inputs and Parameters. The core of this work is to show that, if a model family is *jointly definable* in its inputs and parameters, and if the loss function is definable in the same structure, then the resulting model necessarily has finite sample complexity.

Proposition 2.5 (Definable *Gates* Imply Definable $\mathbb{G}$*-Neural Networks*)**.** *Fix dimensions $d_{in}, P, d_{out} \in \mathbb{N}_+$ and an o-minimal structure $\mathcal{S}$ expanding the real field. If every gate in $\mathbb{G}$ is definable in $\mathcal{S}$, then every $\mathbb{G}$-NN is definable in $\mathcal{S}$, jointly in its inputs and parameters.*

Proposition 2.5 motivates our only key assumption, Assumption 2.6, whose mildness is truly communicated via our broad list of practical examples in Section 4.

Assumption 2.6 (Definable Gates in an O-Minimal Structure)**.** Fix an o-minimal structure $\mathcal{S}$ expanding the real field and suppose that every gate $g \in \mathbb{G}$, cf. (2.1), is definable in $\mathcal{S}$.

3 Main Statistical Guarantees

We begin with the classification formulation of our general statistical guarantee.

Theorem 3.1 (Definability Implies Learning Is Possible—Binary Classification Case)**.** *Suppose Assumption 2.6 holds. Let $n, N \in \mathbb{N}_+$, $f : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}$ be a $\mathbb{G}$-NN, let $\mathbb{P}$ be a Borel probability measure on $\mathbb{R}^n \times \{0,1\}$, and let $((X_i, Y_i))_{i=1}^N$ be i.i.d. random variables with law $\mathbb{P}$.*
There exist constants[1] *$C, K, c > 0$ such that: for every error $\varepsilon > 0$ and every failure probability $0 < \delta \leq 1$*

$$\sup_{\theta \in \mathbb{R}^P} \left| \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[I(I_{(0,\infty)} \circ f_\theta(X) = Y)\big] - \frac{1}{N}\sum_{i=1}^{N} I(I_{(0,\infty)} \circ f_\theta(X_i) = Y_i) \right| \leq \varepsilon$$

holds with probability at least $1 - \delta$, provided that

$$N \geq \frac{C}{\varepsilon^2}\left[(c + 2P\log_2(2P)) + \ln\left(\frac{2}{\delta}\right)\right].$$

We emphasize that Theorem 3.1 imposes *no* weight constraints. In particular, it applies to the class of all ReLU-MLPs with fixed depth and width but *unbounded* parameters: for such classes, standard chaining-based covering arguments conducted over the parameter space cannot yield a finite upper bound on the sample size N (equivalently, a finite complexity bound), since the relevant parameter-space covering numbers are infinite; see, e.g., (Petersen & Zech, 2024, Theorem 14.15). Consequently, existing approaches that permit unbounded parameters typically proceed model-by-model and can be technically demanding, cf. Bartlett et al. (2019) for ReLU-MLPs of fixed depth and width. By contrast, Theorem 3.1 gives this conclusion automatically for every fixed finite feedforward architecture built from the standard definable primitives verified in Section 4.

The next result is a regression analogue of the preceding one. It holds for bounded definable losses, including *clipped* or normalized versions of MSE and MAE, and KL-type losses on definable domains where the divergence is finite and bounded. The takeaway is the same: virtually every reasonable deep learning model class will learn to generalize on *regression* problems.

Theorem 3.2 (Definability Implies Learning Is Possible—Regression Case)**.** *Suppose Assumption 2.6 holds. Let $m, n, N \in \mathbb{N}_+$, $f : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}^m$ be a $\mathbb{G}$-NN, let $\mathbb{P}$ be a Borel probability measure on $\mathbb{R}^n \times \mathbb{R}^m$, let $((X_i, Y_i))_{i=1}^N$ be i.i.d. random variables with law $\mathbb{P}$, and let $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0,1]$ be a definable loss function (in the same o-minimal structure $\mathcal{S}$).*
There exist constants[1] *$C, K, c > 0$ (not depending on N) such that: for every error $\varepsilon > 0$ and every failure probability $0 < \delta \leq 1$*

$$\sup_{\theta \in \mathbb{R}^P} \left| \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[\ell(f_\theta(X), Y)\big] - \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(X_i), Y_i) \right| \leq \varepsilon$$

holds with probability at least $1 - \delta$, provided that

$$N \geq \frac{C}{\varepsilon^2}\left[(c + 2P\log_2(2P))\ln^2\left(\frac{eK}{\varepsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

Intuition of Proofs: What Definability Does. The main results are derived in three steps. First, Proposition 2.5 shows that any $\mathbb{G}$-neural network is jointly definable in its parameters and inputs; whence the composite map $((x,y),\theta) \mapsto I(I_{(0,\infty)} \circ f_\theta(x) = y)$ (resp. $((x,y),\theta) \mapsto \ell(f_\theta(x), y)$ in the regression case) is jointly definable in (x, y, θ). Next, we use this to show that this composite

[1] Where C is universal and K depends only on the fixed architecture, the defining formulas of its gates, and, in the regression case, the loss ℓ, but not on N, ε, δ, $\mathbb{P}$, or the magnitude of the parameters.

map has a bounded number of osculations at any scale, implying that $I_{(0,\infty)} \circ f_\theta$ (resp. f_θ) cannot (resp. fat-)shatter arbitrarily many points at any scale; whence the family must have a finite VC (resp. fat-shattering) dimension as the parameter θ varies. Finally, we use classical results of Valiant (1984); Blumer et al. (1989) (resp. Alon et al. (1997)) to deduce finite sample-complexity bounds in the classification (resp. regression) case.

3.1 Quantitative Guarantees Via Multivariate Pfaffian Gates

We complete this setting by answering the question: *Is it possible to obtain general **quantitative** sample complexity estimates under assumptions satisfied by most contemporary feedforward neural networks?* Indeed, we see that Pfaffianity, cf. A.1, is enough.

Setting 1 (Pfaffian Gates)**.** *Let $f : \mathbb{R}^{d_{in}} \times \mathbb{R}^P \to \mathbb{R}^{d_{out}}$ be a $\mathbb{G}$-neural network as in Definition 2.4. Suppose that, for every $v \in \mathrm{comp}(G)$ and $1 \leq j \leq m_v$, the coordinate $(u, \theta_v) \mapsto g_{v,j}(u, \theta_v)$ is globally Pfaffian on $\mathbb{R}^{M_v+p_v}$ of format at most (q_0, D_0, d_0), where $q_0, D_0 \in \mathbb{N}_0$ and $d_0 \in \mathbb{N}_+$. Let $1 \leq L$ denote the maximal number of computational vertices on a directed path in G, and set $1 \leq U \overset{\text{def.}}{=} \sum_{v \in \mathrm{comp}(G)} m_v$, $\Theta \overset{\text{def.}}{=} \mathbb{R}^P$, and $\mathcal{Z} \overset{\text{def.}}{=} \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}$. Consider the clipped MSE loss*

$$\ell(u, y) \overset{\text{def.}}{=} \min\{1, \|u - y\|_2^2\}$$

for each $u, y \in \mathbb{R}^{d_{out}}$. Write $h_\theta(x, y) \overset{\text{def.}}{=} \ell(f(x, \theta), y)$ and $\mathcal{F}_\ell \overset{\text{def.}}{=} \{h_\theta : \theta \in \Theta\}$.

Then, Theorem 3.2 refines to the following quantitative statement.

Theorem 3.3 (Quantitative Learning with Clipped Squared Loss)**.** *Under Setting 1, there exists a constant $C = C(q_0, D_0, d_0) > 0$ such that, for every Borel probability measure $\mathbb{P}$ on $\mathcal{Z}$, every $N \in \mathbb{N}_+$, every i.i.d. sample $((X_i, Y_i))_{i=1}^N$ with law $\mathbb{P}^N$, and every $0 < \varepsilon, \delta < 1$, the following holds with probability at least $1 - \delta$:*

$$\sup_{\theta \in \mathbb{R}^P} \left| \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(f_\theta(X), Y)] - \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(X_i), Y_i) \right| \leq \varepsilon,$$

provided that

$$N \geq \frac{C}{\varepsilon^2} \left[P^2 U^2 \ln^2\left(\frac{eP^2U^2}{\varepsilon}\right) + \ln\left(\frac{2}{\delta}\right) \right]. \tag{3.1}$$

Since $U \leq LW$, a sufficient alternative condition is obtained by replacing P^2U^2 by $P^2L^2W^2$.

4 From Theory to Practice: A Large Class of Neural Networks Are Definable

We demonstrate the power of our result by establishing the **generality and scope** of Assumption 2.6. It suffices to show that most components used to build a modern deep learning model satisfies this assumption; Proposition 2.5 directly extends the conclusion to the models themselves.

4.1 Neural-Network Building Blocks

Affine and Polynomial Maps The affine functions core to classical MLPs (linear layers) and CNNs (convolutional kernels) are *quadratic polynomial* functions jointly of their inputs and parameters, i.e., $(A, b, x) \mapsto Ax + b$, are among the simplest examples of definable maps; more generally, the same is true of polynomial maps, which cover polynomial feature maps (e.g., $x \mapsto (x_i x_j)_{i,j}$ or all monomials up to a fixed degree) and polynomial activations (e.g., $t \mapsto t^2$ or $t \mapsto t^3$). Since the graph of a polynomial map is defined by polynomial equations, polynomial maps are semi-algebraic and hence definable in the real field; which in particular include the linear (really affine) layers of MLPs.

Proposition 4.1 (Definability of Polynomial Maps)**.** *Let $n, m \in \mathbb{N}_+$, and let $P : \mathbb{R}^n \to \mathbb{R}^m$ be a polynomial map. Then P is semi-algebraic. In particular, P is definable in the real field $(\mathbb{R}, +, \cdot, <)$, and hence in every o-minimal expansion of the real field.*

Residual/skip connections, such as those used in Highway Networks Srivastava et al. (2015b;a), ResNets He et al. (2016) and transformer blocks (Vaswani et al., 2017), fall under this framework too.

Corollary 4.2 (Residual and Gated Residual Polynomial Blocks). *Let $F, G : \mathbb{R}^n \to \mathbb{R}^n$ be polynomial maps. Then the residual block $x \mapsto x+F(x)$ and the gated residual block $x \mapsto x+G(x)\odot F(x)$, where $\odot$ denotes coordinatewise multiplication (or Hadamard product), are polynomial maps. In particular, they are semi-algebraic and definable in the real field.*

Activation Functions in Deep Learning. Activation functions provide the basic nonlinearities used to construct feedforward neural architectures. Rectified Linear Units (ReLUs) became a standard choice in modern deep learning (Nair & Hinton, 2010; Glorot et al., 2011), while other alternatives such as Leaky ReLU (Maas et al., 2013), ELU (Clevert et al., 2016), Swish (also called SiLU) (Ramachandran et al., 2018) and GELU (Hendrycks & Gimpel, 2016), to mention a few, are also popular. For instance, GELU is used in Gemma 3 (Team, 2025), and other gated variants such as SwiGLU have also been widely adopted as part of the transformer blocks (Shazeer, 2020) comprising modern large language models (LLMs) such as DeepSeek-V3 (DeepSeek-AI, 2024), DeepSeek-R1 (Guo et al., 2025), and Llama-3 (Team, 2024). Also, learnable spline activations form the basis of Kolmogorov–Arnold Networks (KANs) (Liu et al., 2025). Piecewise-polynomial activations are semi-algebraic. The sigmoid, hyperbolic tangent, softplus, ELU, Swish, SwiGLU, and the tanh approximation to GELU are definable in $\mathbb{R}_{\exp}$. The exact GELU is definable in the Pfaffian closure $\mathcal{P}(\mathbb{R}_{\exp})$, as established in Proposition C.3.

Proposition 4.3 (Definability of Modern Activation Functions[2]). *Many activation functions: sigmoid* $\sigma(x) \overset{\text{def.}}{=} \frac{1}{1+e^{-x}}$, $\tanh(x)$, $\operatorname{softplus}(x)$, $\operatorname{ReLU}(x)$, $\operatorname{LeakyReLU}_\alpha(x) \overset{\text{def.}}{=} \max\{x, \alpha\, x\}$, $\operatorname{ELU}_\alpha(x) \overset{\text{def.}}{=} xI_{x\geq 0} + \alpha(e^x - 1)I_{x<0}$, $\operatorname{GELU}_{\tanh}(x)$ *(the tanh approximation to GELU),* $\operatorname{Swish}_\beta(x) \overset{\text{def.}}{=} \frac{x}{1+e^{-\beta\, x}}$, $\operatorname{SwiGLU}_\beta(x_1, x_2) \overset{\text{def.}}{=} x_1 \operatorname{Swish}_\beta(x_2)$ *are definable in* $\mathbb{R}_{\text{an}}$, $\mathbb{R}_{\exp}$, *or* $\mathbb{R}_{\text{an,exp}}$.

Multivariate Gating: Maxout, Winner-Take-All, and Mixtures of Experts. Many architectures also use comparison-based multivariate non-linearities, such as maxout, winner-take-all, and hard-routed experts. These maps are piecewise affine over finitely many regions.

Proposition 4.4 (Definability of Multivariate Piecewise-Affine Gating Layers). *Fix $d, m, K \in \mathbb{N}_+$. The following maps are semi-algebraic, hence definable in $(\mathbb{R}, +, \cdot, <)$ and in every o-minimal expansion of the real field.*

(i) ***Maxout:*** *The map $(x, (a_k, b_k)_{k=1}^K) \mapsto \max_{1\leq k\leq K}(a_k^\top x + b_k)$, from $\mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^K$ to $\mathbb{R}$.*

(ii) ***Winner-take-all****: The map obtained by partitioning $\mathbb{R}^d$ into finitely many polyhedral cells and applying one affine map $x \mapsto A_k x + b_k$ on each cell.*

These cover hard-routed mixture-of-experts (MoE) layers, where a finite polyhedral rule selects an affine expert on each cell. More general MoE layers, including for instance those used in GLaM (Du et al., 2022), Mixtral (Jiang et al., 2024), and DeepSeek-V3/R1 (DeepSeek-AI, 2024; Guo et al., 2025), are definable whenever their experts and routing weights are definable: top-k selection is semi-algebraic, while softmax routing is definable in $\mathbb{R}_{\exp}$.

Attention Layers. Attention mechanisms are a central component of modern deep learning, originating in neural machine translation (Bahdanau et al., 2015) and becoming the defining operation of transformer architectures through multi-head self-attention (Vaswani et al., 2017).

Proposition 4.5 (Definability of Multi-Head (self-)Attention). *Fix $N, H, d_{in}, d_k, d_v, d_{out} \in \mathbb{N}_+$ and $\lambda > 0$. For each $h \in \{1, \ldots, H\}$, let $W_Q^h, W_K^h \in \mathbb{R}^{d_k \times d_{in}}$ and $W_V^h \in \mathbb{R}^{d_v \times d_{in}}$, and let $W_O \in \mathbb{R}^{d_{out} \times H d_v}$. For $X \in \mathbb{R}^{N \times d_{in}}$, define*

$$\operatorname{Attn}_h(X)_n \overset{\text{def.}}{=} \sum_{m=1}^{N} \alpha_{n,m}^{(h)}(X)\, W_V^h X_m, \qquad \alpha_{n,m}^{(h)}(X) \overset{\text{def.}}{=} \frac{\exp\big(\lambda \langle W_Q^h X_n, W_K^h X_m\rangle / \sqrt{d_k}\big)}{\sum_{\ell=1}^{N} \exp\big(\lambda \langle W_Q^h X_n, W_K^h X_\ell\rangle / \sqrt{d_k}\big)}.$$

Then the multi-head attention map $(X, (W_Q^h, W_K^h, W_V^h)_{h=1}^H, W_O) \mapsto \operatorname{MHA}(X) \overset{\text{def.}}{=} W_O\big(\operatorname{Attn}_1(X)_n, \ldots, \operatorname{Attn}_H(X)_n\big)$, is definable in $\mathbb{R}_{\exp}$, jointly in X and the parameters.

[2] A comprehensive version of this proposition is available in the appendix; cf. Proposition C.3.

Extending the result above to cross-attention is trivial, see Corollary C.5. The same is true for sliding-window attention (Corollary C.6) which is typical in modern long-context architectures. For example, the Gemma 4 model card describes a hybrid attention mechanism interleaving local sliding-window attention with full global attention (Google, 2026). A similar softmax-weighted aggregation principle (linear additive self-attention (Sáez De Ocáriz Borde, 2024)) also underlies graph attention networks (GATs), where attention is restricted to fixed node neighbourhoods (Velickovic et al., 2018). This then directly implies the GNN version of that result.

Normalization Layers. Normalization layers are now standard components of modern deep architectures, beginning with batch normalization (Ioffe & Szegedy, 2015), and later variants such as layer normalization (Ba et al., 2016), instance normalization (Ulyanov et al., 2016), group normalization (Wu & He, 2018), and RMS normalization (Zhang & Sennrich, 2019). From the o-minimal viewpoint, these layers introduce no obstruction: after fixing the finite input dimension, parameters, and stabilization constant $\varepsilon > 0$, the usual normalization maps are semi-algebraic.

Proposition 4.6 (Definability of Standard Normalization Layers)**.** *Fix $d \in \mathbb{N}_+$, $\varepsilon > 0$, parameters $\gamma, \beta \in \mathbb{R}^d$, and a partition $\mathcal{G}$ of $\{1, \dots, d\}$. For each $i \in \{1, \dots, d\}$, let $G(i) \in \mathcal{G}$ denote the unique block containing i, and define $\mu_G(x) \overset{\text{def.}}{=} \frac{1}{|G|} \sum_{j \in G} x_j$ and $\sigma_G^2(x) \overset{\text{def.}}{=} \frac{1}{|G|} \sum_{j \in G} \big(x_j - \mu_G(x)\big)^2$. Then, the normalization map $\mathrm{Norm}_{\gamma,\beta,\varepsilon} : \mathbb{R}^d \to \mathbb{R}^d$ defined coordinate-wise by*

$$\mathrm{Norm}_{\gamma,\beta,\varepsilon}(x)_i \overset{\text{def.}}{=} \gamma_i \frac{x_i - \mu_{G(i)}(x)}{\sqrt{\sigma^2_{G(i)}(x) + \varepsilon}} + \beta_i$$

is definable in every o-minimal expansion of the real field.

Embeddings and Fourier Positional-Encodings. Embedding and positional-encoding layers are standard components of transformer-type architectures. Embedding layers with a fixed finite vocabulary are definable because they are simply lookup maps from a finite set into Euclidean space.

Proposition 4.7 (Definability of Embedding Layers)**.** *Let $N, d \in \mathbb{N}_+$. The parametrized embedding*

$$\mathrm{Emb} : \{1, \dots, N\} \times \mathbb{R}^{N \times d} \to \mathbb{R}^d, \qquad \mathrm{Emb}(i, E) \overset{\text{def.}}{=} E_{i,:},$$

where $E_{i,:}$ denotes the ith row of E, is semi-algebraic. In particular, it is definable in the real field $(\mathbb{R}, +, \cdot, <)$, and hence in every o-minimal expansion of the real field.

Fourier positional encodings also appear broadly in coordinate-based learning and representation learning (Tancik et al., 2020). From the o-minimal viewpoint, the only subtlety is that the global sine and cosine functions are not definable in $\mathbb{R}_{\mathrm{an,exp}}$. Thus, Fourier positional encodings must be interpreted either on a fixed finite set of positions, or on a fixed bounded position domain. Under this standard bounded-context interpretation, they are definable in $\mathbb{R}_{\mathrm{an}}$.

Definition 4.8 (*Bounded* Fourier Positional Encoding)**.** Let $D \subseteq \mathbb{R}^r$ be a bounded definable set, let $M \in \mathbb{N}_+$, and fix frequencies $\omega_1, \dots, \omega_M \in \mathbb{R}^r$ and phases $\varphi_1, \dots, \varphi_M \in \mathbb{R}$. The associated bounded Fourier positional encoding is the map $\mathrm{PE}_{\Omega,\varphi} : D \to \mathbb{R}^{2M}$ defined by

$$\mathrm{PE}_{\Omega,\varphi}(t) \overset{\text{def.}}{=} \Big(\sin\big(2\pi\langle \omega_j, t\rangle + \varphi_j\big), \cos\big(2\pi\langle \omega_j, t\rangle + \varphi_j\big) \Big)_{j=1}^{M}.$$

For standard sinusoidal positional encodings (and in RoPE Su et al. (2024)), the frequencies are fixed deterministically. If frequencies or phases are trainable, joint $\mathbb{R}_{\mathrm{an}}$-definability follows when their parameter domain is bounded and $\mathbb{R}_{\mathrm{an}}$-definable. More generally, uniformly bounded phase ranges suffice; see Remark C.7.

Proposition 4.9 (Definability of Bounded Fourier Positional Encodings)**.** *Let $D \subseteq \mathbb{R}^r$ be bounded and $\mathbb{R}_{\mathrm{an}}$-definable. For fixed frequencies $\omega_1, \dots, \omega_M \in \mathbb{R}^r$ and fixed phases $\varphi_1, \dots, \varphi_M \in \mathbb{R}$, the bounded Fourier positional encoding $\mathrm{PE}_{\Omega,\varphi} : D \to \mathbb{R}^{2M}$ is definable in $\mathbb{R}_{\mathrm{an}}$, and hence in $\mathbb{R}_{\mathrm{an,exp}}$. If $D = \{1, \dots, N\}$ is a fixed finite set of positions, then $\mathrm{PE}_{\Omega,\varphi}$ is semi-algebraic.*

4.2 How to Apply our Result: Feedforward Architectures

We now package the preceding layer-wise definability results into architecture-level corollaries. The aim is not to strengthen Proposition 2.5 or Theorems 3.1 and 3.2, but to spell out their consequences for standard fixed finite architectures.

Corollary 4.10 (MLPs Are Definable and Have Finite Sample Complexity)**.** *Fix $L \in \mathbb{N}_+$ and widths $d_0, d_1, \ldots, d_L, d_{L+1} \in \mathbb{N}_+$. For each $l \in \{0, \cdots, L-1\}$, let σ_l be as in Proposition 4.3, and let $P \overset{\text{def.}}{=} \sum_{l=0}^{L} d_{l+1}(d_l+1)$. Then, the MLP $f_\theta : \mathbb{R}^{d_0} \times \mathbb{R}^P \to \mathbb{R}^{d_{L+1}}$ given, for any (x, θ), by*

$$f_\theta(x) \overset{\text{def.}}{=} A_L x^{(L)} + b_L, \quad x^{(l+1)} \overset{\text{def.}}{=} \sigma_l \bullet \left(A_l x^{(l)} + b_l\right) \text{ for } l \in \{0, \cdots, L-1\} \text{ and } \quad x^{(0)} \overset{\text{def.}}{=} x$$

where $\bullet$ denotes component-wise composition and $\theta \overset{\text{def.}}{=} (\mathrm{vec}(A_l), b_l)_{l=0}^{L}$, is definable in $\mathbb{R}_{\mathrm{an,exp}}$. In particular, the conclusions of Theorems 3.1 and 3.2 apply.

Corollary 4.11 (Transformers are Definable and Have Finite Sample Complexity)**.** *Fix $N, L, T, d_{\mathrm{out}} \in \mathbb{N}_+$, fix $d_0 \in \mathbb{N}_+$, and, for each $l \in [L]_+$, let $H_l, d_l, d_{k,l}, d_{v,l}, d_{\mathrm{ff},l} \in \mathbb{N}_+$. Let $\lambda, \varepsilon > 0$. Let $E \in \mathbb{R}^{N \times d_0}$ be an embedding parameter, and let Emb be as in Proposition 4.7, let $\mathrm{PE}_{\Omega,\varphi} : \{1, \ldots, T\} \to \mathbb{R}^{d_0}$ be as in Proposition 4.9, for fixed Ω and φ, and, for each $l \in [L]_+$, let Norm_l denote $\mathrm{Norm}_{\gamma_l, \beta_l, \varepsilon}$ as in Proposition 4.6 (two per block), for fixed $\gamma_l, \beta_l \in \mathbb{R}^{d_l}$. For each $l \in [L]_+$, let σ_l be as in Proposition 4.3. Set*

$$P \overset{\text{def.}}{=} N d_0 + \sum_{l=1}^{L} \Big(H_l d_{l-1}(2d_{k,l} + d_{v,l}) + H_l d_{v,l} d_l + d_{l-1} d_l + d_{\mathrm{ff},l}(2d_l + 1) + d_l \Big) + d_{\mathrm{out}}(d_L + 1).$$

Then, the transformer $f_\theta : \{1, \ldots, N\}^T \times \mathbb{R}^P \to \mathbb{R}^{T \times d_{\mathrm{out}}}$ given, for any (X, θ), by

$$\begin{aligned}
f_\theta(X) &\overset{\text{def.}}{=} Z^{(L)} W_{\mathrm{out}} + \mathbf{1}_T b_{\mathrm{out}}^\top \\
Z^{(l)} &\overset{\text{def.}}{=} \mathrm{Norm}^{\bullet}_{l,2}\left(Y^{(l)} + \left(\sigma_l \bullet \left(Y^{(l)} W_l^1 + \mathbf{1}_T (b_l^1)^\top\right) W_l^2 + \mathbf{1}_T (b_l^2)^\top\right)\right) && \text{for } l \in [L]_+ \\
Y^{(l)} &\overset{\text{def.}}{=} \mathrm{Norm}^{\bullet}_{l,1}\left(Z^{(l-1)} W_l^R + \mathrm{MHA}_l(Z^{(l-1)})\right) && \text{for } l \in [L]_+ \\
Z_t^{(0)} &\overset{\text{def.}}{=} \mathrm{Emb}(X_t, E) + \mathrm{PE}_{\Omega,\varphi}(t) && \text{for } t \in [T]_+,
\end{aligned}$$

where $\mathrm{Norm}^{\bullet}_l$ denotes row-wise normalization; where $\mathrm{MHA}_l(Z) \overset{\text{def.}}{=} \left(\bigoplus_{h=1}^{H_l} \mathrm{Attn}_{l,h}(Z)\right) W_l^O$,

$$\mathrm{Attn}_{l,h}(Z)_n \overset{\text{def.}}{=} \sum_{m=1}^{T} \alpha_{n,m}^{(l,h)}(Z)\, Z_m W_{l,h}^V, \quad \alpha_{n,m}^{(l,h)}(Z) \overset{\text{def.}}{=} \frac{\exp\left(\lambda \langle Z_n W_{l,h}^Q, Z_m W_{l,h}^K \rangle / \sqrt{d_{k,l}}\right)}{\sum_{r=1}^{T} \exp\left(\lambda \langle Z_n W_{l,h}^Q, Z_r W_{l,h}^K \rangle / \sqrt{d_{k,l}}\right)},$$

where $\oplus$ denotes concatenation over attention heads, $\bullet$ denotes component-wise composition, and $\theta \overset{\text{def.}}{=} \left(E, (W_{l,h}^Q, W_{l,h}^K, W_{l,h}^V)_{l,h}, (W_l^O, W_l^R, W_l^1, b_l^1, W_l^2, b_l^2)_{l=1}^{L}, W_{\mathrm{out}}, b_{\mathrm{out}}\right) \in \mathbb{R}^P$.
Then, f_θ is definable in $\mathbb{R}_{\mathrm{an,exp}}$; and the conclusions of Theorems 3.1 and 3.2 apply.

Analogous conclusions can also be derived for pre-LN transformers (Xiong et al., 2020) and MoEs, CNNs, most GNNs, and most other modern feedforward architectures in a similar fashion, using the previous propositions and main theorems.

5 Conclusion

We showed that distribution-free generalization is a universal consequence of tame feedforward computation. More precisely, every fixed finite feedforward architecture (numerical values of the parameters themselves are not fixed: they are trainable) whose layers are definable in an o-minimal structure has finite sample complexity in the agnostic PAC setting for classification (Theorem 3.1) and for bounded definable regression losses (Theorem 3.2), even when its parameters are unbounded. This covers the standard non-recurrent architectures used in modern deep learning, including MLPs, CNNs, finite-size GNNs, and transformers, to name a few. The main conceptual consequence is that PAC learnability is *not* a *fragile property* requiring architecture-specific VC arguments, nor is it tied to particular activations or parameter bounds. Instead, it follows automatically from the definable structure shared by essentially all finite feedforward models used in practice. This shifts the role of generalization theory: for modern non-recurrent architectures, finite sample complexity should be viewed as a baseline property, not as the feature that distinguishes one architecture from another. For future work see Appendix D.

The Practical Takeaway. Accordingly, the meaningful differences between architectures must be sought elsewhere: in their inductive biases, symmetries and geometric prior, optimization geometry, scalability, and computational efficiency. Therefore o-minimality provides a unifying language for separating what is generic about deep learning from what is genuinely architecture-specific.

REFERENCES

Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. Open problem: Are all VC-classes CPAC learnable? In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4636–4641. PMLR, 2021.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: PAC learning and online learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer Science & Business Media, 2008.

P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE Trans. Inform. Theory*, 43(5):1721–1724, 1997.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. *Advances in neural information processing systems*, 11, 1998.

Peter L. Bartlett and Wolfgang Maass. Vapnik-Chervonenkis dimension of neural nets. In Michael A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, pp. 1188–1192. MIT Press, 2 edition, 2003.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20: 63:1–63:17, 2019.

Gal Binyamini and Dmitry Novikov. Tameness in geometry and arithmetic: beyond o-minimality. In *International Congress of Mathematicians*, pp. 1440–1461. European Mathematical Society-EMS-Publishing House GmbH, 2023.

Gal Binyamini, Dmitri Novikov, and Benny Zack. Sharply o-minimal structures and sharp cellular decomposition. *arXiv preprint arXiv:2209.10972*, 2022.

Gal Binyamini, Dmitry Novikov, and Benny Zak. Wilkie's conjecture for Pfaffian structures. *Annals of Mathematics*, 199(2):795–821, 2024.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Marco Bressan, Nataly Brukhim, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A fine-grained characterization of PAC learnability. In *COLT*, Proceedings of Machine Learning Research, pp. 641–676. PMLR, 2025.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.*, 34:18–42, 2016.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Hunter Chase and James Freitag. Model theory and machine learning. *Bulletin of Symbolic Logic*, 25 (3):319–332, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6572–6583, 2018.

David Chiang. Transformers in Uniform TC0. *Transactions on Machine Learning Research*, 2025.

Y Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. *UCLA. URL: http://starai.cs.ucla.edu/papers/ProbCirc20.pdf*, 6, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by Exponential Linear Units (ELUs). In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Michel Coste. *An Introduction to O-Minimal Geometry*. Institut de Recherche Mathématique de Rennes (IRMAR), Rennes, France, 1999. Lecture notes.

Haitz Sáez de Ocáriz Borde and Michael Bronstein. Mathematical foundations of geometric deep learning. *arXiv preprint arXiv:2508.02723*, 2025.

DeepSeek-AI. DeepSeek-V3 technical report. *CoRR*, abs/2412.19437, 2024.

Valentino Delle Rose, Alexander Kozachinskiy, Cristóbal Rojas, and Tomasz Steifer. Find a witness or shatter: the landscape of computable PAC learning. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 511–524. PMLR, 2023.

Giuseppe Alessio D'Inverno, Monica Bianchini, and Franco Scarselli. VC dimension of graph neural networks with Pfaffian activation functions. *Neural Networks*, 182:106924, 2025.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts, 2022.

Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *ICML*, Proceedings of Machine Learning Research, pp. 5793–5831. PMLR, 2022.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

Andrei Gabrielov and Nicolai Vorobjov. Complexity of computations with Pfaffian and Noetherian functions. In *Normal Forms, Bifurcations and Finiteness Problems in Differential Equations*, volume 137 of *NATO Sci. Ser. II Math. Phys. Chem.*, pp. 211–250. Kluwer Acad. Publ., Dordrecht, 2004.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3419–3430. PMLR, 2020.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323. PMLR, 2011.

Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

Google. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026.

Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. *Journal of Machine Learning Research*, 22(273):1–29, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, Tao Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nat.*, 645(8081):633–638, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

H. R. Johnson and Michael C. Laskowski. Compression schemes, stable definable families, and o-minimal structures. *Discrete & Computational Geometry*, 43(4):914–926, 2010.

Stasys Jukna. *Boolean Function Complexity*, volume 27 of *Algorithms and Combinatorics*. Springer, Heidelberg, 2012. Advances and frontiers.

Stasys Jukna. Tropical circuit complexity. *Limits of Pure Dynamic Programming/by Stasys Jukna*, 2, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994.

A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. Translated from the Russian by Smilka Zdravkovska.

Anastasis Kratsios, Dennis Zvigelsky, and Bradd Hart. Quantifying the limits of AI reasoning: Systematic neural network representations of algorithms. *arXiv preprint arXiv:2508.18526*, 2025.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114, 2012.

Michael C. Laskowski. Vapnik-Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 45(2):377–384, 1992.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Wenhao Li, Anastasis Kratsios, Hrad Ghoukasian, and Dennis Zvigelsky. Certifiable Boolean reasoning is universal. *arXiv preprint arXiv:2602.05120*, 2026.

Yannick Limmer, Anastasis Kratsios, Xuwei Yang, Raeid Saqur, and Blanka Horvath. Higher-order transformer derivative estimates for explicit pathwise learning guarantees. *arXiv preprint arXiv:2405.16563*, 2024.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-Arnold Networks. In *ICLR*, 2025.

Antonio Lloris Ruiz, Encarnación Castillo Morales, Luis Parrilla Roure, and Antonio García Ríos. *Algebraic Circuits*, volume 66 of *Intelligent Systems Reference Library*. Springer, Heidelberg, 2014.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

Angus Macintyre and Eduardo D. Sontag. Finiteness results for sigmoidal "neural" networks. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pp. 325–334, 1993.

Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. WL meet VC. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25275–25302. PMLR, 2023.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

Philipp Petersen and Jakob Zech. Mathematical Theory of Deep Learning. *arXiv preprint arXiv:2407.18384*, 2024.

Anand Pillay and Charles Steinhorn. Definable sets in ordered structures. i. *Transactions of the American Mathematical Society*, 295(2):565–592, 1986.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *International Conference on Learning Representations Workshop*, 2018.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

Haitz Sáez De Ocáriz Borde. Elucidating graph neural networks, transformers, and graph transformers. ResearchGate, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, 1997.

Patrick Speissegger. The Pfaffian closure of an *o*-minimal structure. *Journal für die Reine und Angewandte Mathematik*, 1999(508):189–211, 1999.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2377–2385, Cambridge, MA, USA, 2015a. MIT Press.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015b.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Gemma Team. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025.

Llama Team. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Lou van den Dries. *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1998.

Lou van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

Benjie Wang, Denis Deratani Mauá, Guy Van den Broeck, and YooJung Choi. A compositional atlas for algebraic circuits. In *NeurIPS*, 2024.

A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020. URL https://arxiv.org/abs/2002.04745.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Honghua Zhang, Brendan Juba, and Guy Van den Broeck. Probabilistic generating circuits. In *International Conference on Machine Learning*, pp. 12447–12457. PMLR, 2021.

A Additional Background

In this appendix we provide additional mathematical background in Pfaffian functions, learning theory, and VC dimensions.

A.1 Pfaffian Functions: A Smooth Tame Class

Pfaffian functions, introduced by Khovanskii in his theory of fewnomials (Khovanskiĭ, 1991), form a robust class of real-analytic functions whose derivatives satisfy triangular polynomial differential equations. They provide a smooth analogue of semi-algebraic functions: their format controls their algebraic-differential complexity, while the main result of (Speissegger, 1999) ensures that adjoining Pfaffian solutions to an o-minimal structure preserves o-minimality. Thus Pfaffian functions are both analytically regular and model-theoretically tame.

Definition A.1 (Pfaffian Chain)**.** Let $U \subseteq \mathbb{R}^n$ be a nonempty open domain. A Pfaffian chain of length q and degree D on U is a sequence of real-analytic functions $f_1, \ldots, f_q : U \to \mathbb{R}$ such that, for every $1 \le i \le q$ and every $1 \le j \le n$, there exists a polynomial $P_{i,j} \in \mathbb{R}[X_1, \ldots, X_n, Y_1, \ldots, Y_i]$ of degree at most D satisfying

$$\frac{\partial f_i(\bar{x})}{\partial x_j} = P_{i,j}\big(\bar{x}, f_1(\bar{x}), \ldots, f_i(\bar{x})\big), \qquad \bar{x} \in U.$$

Definition A.2 (Pfaffian Function of Format (q, D, d))**.** Let $U \subseteq \mathbb{R}^n$ be a nonempty open domain. A function $g : U \to \mathbb{R}$ is Pfaffian of format (q, D, d) on U if there exist a Pfaffian chain $f_1, \ldots, f_q : U \to \mathbb{R}$ of length (or order) q and degree D, and a polynomial $Q \in \mathbb{R}[X_1, \ldots, X_n, Y_1, \ldots, Y_q]$ of degree at most d, such that

$$g(\bar{x}) = Q\big(\bar{x}, f_1(\bar{x}), \ldots, f_q(\bar{x})\big), \qquad \bar{x} \in U.$$

By convention, if $F : U \to \mathbb{R}^m$, then F is Pfaffian of format at most (q, D, d) on U if each coordinate function $F_1, \ldots, F_m$ is Pfaffian of format at most (q, D, d) on U.

Every Pfaffian function is real-analytic, and hence C^∞, on its domain. If its open domain U is definable in an o-minimal expansion $\mathcal{S}$ of the real field, then the function is definable in the Pfaffian closure $\mathcal{P}(\mathcal{S})$. This follows by applying the Pfaffian-closure theorem successively to the functions in its triangular chain (Speissegger, 1999). Without a definability assumption on U, the analytic definition alone does not imply o-minimal definability.

For illustrative purposes, following (Gabrielov & Vorobjov, 2004, Example 2.3), we record several elementary examples of Pfaffian functions.

Example 4 (Elementary functions). The functions $e^{\beta \cdot}$, for any $\beta \in \mathbb{R}$, $\log(|\cdot|)$, $\arctan(\cdot)$, and every real-analytic branch of an algebraic function are Pfaffian on their domains of definition. In particular, $x \mapsto x^\alpha$, for rational α, is Pfaffian on any interval on which it is real-valued and real-analytic.

Example 5 (Sinusoidal functions on bounded domains). The function $t \mapsto \sin(t)$ is not Pfaffian on all of $\mathbb{R}$, but its restriction to any bounded interval is Pfaffian. Consequently, the three-variable function $(x, w, b) \mapsto \sin(wx + b)$ is Pfaffian on every bounded open domain $U \subseteq \mathbb{R}^3$.

Indeed, set $u(x, w, b) \overset{\text{def.}}{=} wx + b$. Since U is bounded, $u(U)$ is bounded. Choose $k \in \mathbb{N}_+$ large enough so that $u(U)/k$ is contained in an interval of length strictly smaller than 2π. Then there exists $c \in \mathbb{R}$ such that

$$\frac{u(U)}{k} - c \subset (-\pi, \pi).$$

Define, for $(x, w, b) \in U$,

$$f_1(x, w, b) \overset{\text{def.}}{=} \tan\left(\frac{u(x, w, b)/k - c}{2}\right),$$
$$f_2(x, w, b) \overset{\text{def.}}{=} \frac{1}{1 + f_1(x, w, b)^2}.$$

The functions f_1 and f_2 are real-analytic on U. Moreover, for each $z \in \{x, w, b\}$,

$$\frac{\partial f_1}{\partial z} = \frac{1}{2k}\big(1 + f_1^2\big)\frac{\partial u}{\partial z} \in \mathbb{R}[x, w, b, f_1],$$

and

$$\frac{\partial f_2}{\partial z} = -2f_1 f_2^2 \frac{\partial f_1}{\partial z} = -\frac{1}{k} f_1 f_2^2 (1 + f_1^2) \frac{\partial u}{\partial z} \in \mathbb{R}[x, w, b, f_1, f_2].$$

Thus (f_1, f_2) is a Pfaffian chain on U; for instance, it has length 2 and degree at most 6.

Now let $\theta(x, w, b) \stackrel{\text{def.}}{=} u(x, w, b)/k - c$. By the half-angle identities,

$$\sin(\theta) = 2f_1 f_2, \qquad \cos(\theta) = (1 - f_1^2) f_2.$$

Therefore, by the angle-sum formula,

$$\sin\left(\frac{u(x, w, b)}{k}\right) = 2\cos(c) f_1 f_2 + \sin(c)(1 - f_1^2) f_2,$$

and

$$\cos\left(\frac{u(x, w, b)}{k}\right) = \cos(c)(1 - f_1^2) f_2 - 2\sin(c) f_1 f_2.$$

Finally, if U_{k-1} denotes the Chebyshev polynomial of the second kind, then

$$\sin(u(x, w, b)) = \sin\left(k \frac{u(x, w, b)}{k}\right) = \sin\left(\frac{u(x, w, b)}{k}\right) U_{k-1}\left(\cos\left(\frac{u(x, w, b)}{k}\right)\right).$$

Since both $\sin(u/k)$ and $\cos(u/k)$ are polynomial expressions in f_1 and f_2 of degree at most 3, it follows that $\sin(wx + b)$ is Pfaffian on U, with format at most $(2, 6, 3k)$.

As discussed in (Khovanskiĭ, 1991, page 111), Pfaffian functions are closed under sums, products, derivatives, composition, and reciprocals of nonvanishing Pfaffian functions. They are also stable under solving suitable Pfaffian equations. Quantitative bounds on the order and degree of Pfaffian functions constructed in this way are available; see, for instance, (Gabrielov & Vorobjov, 2004, Lemmas 2.4 and 2.5).

A.2 Learning Theory

Consider a uniform class of scalar-valued functions

$$\mathcal{F} = \{f(\bar{x}, \bar{\theta}) : \mathbb{R}^n \to \mathbb{R}^1, \bar{\theta} \in \mathbb{R}^p\}.$$

We say that a set $A = \{\bar{a}_1, \ldots, \bar{a}_d\}$ is *pseudo-shattered* by $\mathcal{F}$ if and only if there exists a set of real numbers $\{r_1, \ldots, r_d\}$ such that for every subset $A' \subseteq A$, there is $\bar{\theta}_{A'} \in \mathbb{R}^p$ such that

$$\bar{a}_i \in A' \Leftrightarrow f(\bar{a}_i, \bar{\theta}_{A'}) > r_i.$$

The *pseudo-dimension* of $\mathcal{F}$, $\text{P-}\dim(\mathcal{F})$, is the size of the largest set that can be pseudo-shattered by $\mathcal{F}$. If we allow non-scalar functions,

$$\mathcal{F} = \{f(\bar{x}, \bar{\theta}) : \mathbb{R}^n \to \mathbb{R}^m, \bar{\theta} \in \mathbb{R}^p\},$$

we will define the pseudo-dimension to be the component-wise minimum pseudo-dimension, and so we will restrict our attention to scalar-valued functions.

The pseudo-dimension is given with a strict inequality, but the bi-implication in this definition still gives the same encoding of the characteristic function of each subset $A' \subseteq A$, perhaps with a little less "wiggle room". This motivates the notion of fat-shattering dimension of Kearns & Schapire (1994), defined as follows.

Fix a tolerance $\gamma > 0$. We say that a set $A = \{\bar{a}_1, \ldots, \bar{a}_d\} \subseteq \mathbb{R}^n$ is *γ-fat shattered* by $\mathcal{F}$ if and only if there exists a set of real numbers $\{r_1, \ldots, r_d\}$ such that for every subset $A' \subseteq A$, there exists $\bar{\theta}_{A'} \in \mathbb{R}^p$ such that, for every $i \in \{1, \ldots, d\}$,

$\bar{a}_i \in A'$ implies that: $f(\bar{a}_i, \bar{\theta}_{A'}) \geq r_i + \gamma$ *and* $\bar{a}_i \notin A'$ implies that: $f(\bar{a}_i, \bar{\theta}_{A'}) \leq r_i - \gamma$.

The *γ-fat shattering dimension* of $\mathcal{F}$ is

$$\text{fat}_\gamma(\mathcal{F}) \stackrel{\text{def.}}{=} \sup\left\{ d \in \mathbb{N} \;:\; \exists A \subseteq \mathbb{R}^n,\ |A| = d,\ A \text{ is } \gamma\text{-fat shattered by } \mathcal{F} \right\},$$

with the convention that $\text{fat}_\gamma(\mathcal{F}) = \infty$ if sets of arbitrarily large size are γ-fat shattered by $\mathcal{F}$.

A.2.1 Reducing to the VC Dimension

We will work in $\mathbb{R}$ equipped with the usual language of ordered rings, with a symbol "$<$" for strict inequality. Consider a uniform family of scalar-valued functions

$$\mathcal{F} = \{f(\bar{x}, \bar{\theta}) : \mathbb{R}^n \to \mathbb{R}^1, \bar{\theta} \in \mathbb{R}^p\}.$$

Then P-$\dim(\mathcal{F}) \geq V$ if and only if there is a set $A = \{(\bar{a}_1, r_1), \ldots, (\bar{a}_V, r_V)\}$ such that for any subset $A' \subseteq A$, there is a tuple $\bar{\theta}_{A'} \in \mathbb{R}^p$ that encodes the characteristic function of A':

$$(\bar{a}_i, r_i) \in A' \Leftrightarrow f(\bar{a}_i, \bar{\theta}_{A'}) - r_i > 0.$$

In other words, the pseudo-dimension of $\mathcal{F}$ is at least V if and only if the VC dimension of the class of uniformly definable sets

$$\mathcal{C}_{\mathcal{F}} = \{f(\bar{x}, \bar{\theta}) - z > 0 : \bar{\theta} \in \mathbb{R}^p\}$$

is at least V. Thus, we may bound the pseudo-dimension of $\mathcal{F}$ by bounding the size V of a set that is shattered by $\mathcal{C}_{\mathcal{F}}$.

B Proofs of General Theorems

In this appendix we provide the proofs for the main theoretical results discussed in this paper, as well as for concrete neural network building blocks and architectures.

B.1 Technical Notes

We now make a few technical notes which we henceforth implicitly apply throughout our manuscripts.

In what follows, for any $s = (z_1, \ldots, z_N) \in \mathcal{Z}^N$, write

$$\widehat{R}_s(\theta) = \frac{1}{N} \sum_{i=1}^{N} h(z_i, \theta)$$

and for any Borel probability measure $\mathbb{P}$ on $\mathcal{Z}$, set

$$R(\theta) = \mathbb{E}_{Z \sim \mathbb{P}}[h(z, \theta)], \qquad \Delta_N(s) = \sup_{\theta \in \Theta} |R(\theta) - \widehat{R}_s(\theta)|.$$

Lemma B.1 (Measurability and Approximate ERM). *Let $\mathcal{Z} \subseteq \mathbb{R}^q$ and $\Theta \subseteq \mathbb{R}^p$ be nonempty Borel sets, and let $h : \mathcal{Z} \times \Theta \to [0, 1]$ be jointly Borel measurable. For amy Borel probability measure $\mathbb{P}$, R is Borel measurable and Δ_N is universally measurable, hence measurable for the completion of $\mathbb{P}^N$.*

Suppose additionally that $\mathcal{Z}$, Θ, and h are definable in the same o-minimal expansion of the real field. For every $N \in \mathbb{N}_+$ and $\eta > 0$, there exists a definable, hence Borel measurable, map $\widehat{\theta}_{N,\eta} : \mathcal{Z}^N \to \Theta$ such that

$$\widehat{R}_s(\widehat{\theta}_{N,\eta}(s)) \leq \inf_{\theta \in \Theta} \widehat{R}_s(\theta) + \eta, \qquad s \in \mathcal{Z}^N.$$

For every Borel probability measure $\mathbb{P}$ on $\mathcal{Z}$ and every $s \in \mathcal{Z}^N$, this map satisfies

$$R(\widehat{\theta}_{N,\eta}(s)) - \inf_{\theta \in \Theta} R(\theta) \leq 2\Delta_N(s) + \eta.$$

Proof. For Borel sets $A \subseteq \mathcal{Z}$ and $B \subseteq \Theta$,

$$\int_{\mathcal{Z}} I_A(z) I_B(\theta) \, \mathbb{P}(dz) = \mathbb{P}(A) I_B(\theta).$$

The bounded monotone-class theorem therefore implies that $\theta \mapsto \int g(z, \theta) \, \mathbb{P}(dz)$ is Borel measurable for every bounded jointly Borel function g. Applying this to $g = h$ proves that R is Borel measurable. The map $(s, \theta) \mapsto \widehat{R}_s(\theta)$ is also Borel measurable. Hence, for every $a \in \mathbb{R}$,

$$B_a \stackrel{\text{def.}}{=} \left\{(s, \theta) \in \mathcal{Z}^N \times \Theta : |R(\theta) - \widehat{R}_s(\theta)| > a\right\}$$

is Borel. Since $\mathcal{Z}^N$ and Θ are standard Borel spaces,

$$\{s : \Delta_N(s) > a\} = \mathrm{proj}_{\mathcal{Z}^N}(B_a)$$

is analytic, hence universally measurable. This proves the first assertion.

Now impose the additional definability assumptions and set $m_N(s) \overset{\text{def.}}{=} \inf_{\theta\in\Theta} \widehat{R}_s(\theta) \in [0,1]$. Its graph is the definable set

$$\mathrm{graph}(m_N) = \big\{(s,a) \in \mathcal{Z}^N \times [0,1] : (\forall \theta \in \Theta)\ a \le \widehat{R}_s(\theta), \\ (\forall u > 0)(\exists \theta \in \Theta)\ \widehat{R}_s(\theta) < a + u\big\}.$$

For fixed $\eta > 0$, the definable set

$$E_{N,\eta} \overset{\text{def.}}{=} \big\{(s,\theta) \in \mathcal{Z}^N \times \Theta : \widehat{R}_s(\theta) < m_N(s) + \eta\big\}$$

has nonempty fibers over every $s \in \mathcal{Z}^N$. Definable choice, (Coste, 1999, Theorem 3.1), gives a definable map $\widehat{\theta}_{N,\eta}$ with $(s, \widehat{\theta}_{N,\eta}(s)) \in E_{N,\eta}$. By cell decomposition, this map is continuous on each member of a finite definable partition of its domain, and is therefore Borel measurable.

Finally, for every $s \in \mathcal{Z}^N$ and $\theta \in \Theta$,

$$R(\widehat{\theta}_{N,\eta}(s)) \le \widehat{R}_s(\widehat{\theta}_{N,\eta}(s)) + \Delta_N(s) \le \widehat{R}_s(\theta) + \eta + \Delta_N(s) \le R(\theta) + \eta + 2\Delta_N(s).$$

Taking the infimum over $\theta \in \Theta$ proves the final assertion. □

B.2 Proof of the Key Insight

Proof of Proposition 2.5. Let $f : \mathbb{R}^{d_{in}} \times \mathbb{R}^P \to \mathbb{R}^{d_{out}}$ be a $\mathbb{G}$-NN. By Definition 2.4, there is a finite DAG $G = (V, E)$, a binary lifting channel Π, gates $\mathbb{G} \ni g_v : \mathbb{R}^{M_v} \times \mathbb{R}^{p_v} \to \mathbb{R}^{m_v}$, $v \in \mathrm{comp}(G)$, and a linear readout A_{out} such that f is obtained by propagating states along G.

We prove, by induction along the partial order of G, that every node state is definable as a function of (x, θ). For the input nodes, this is immediate: the lifted input $z^0 = \Pi x$ is a linear, hence semi-algebraic, function of x. Since $\mathcal{S}$ expands the real field, every semi-algebraic map is definable in $\mathcal{S}$.

Now fix $v \in \mathrm{comp}(G)$ and assume that, for every parent $u \in \mathrm{Pa}_G(v)$, the state z_u is definable as a function of (x, θ). Then the concatenated parent state $(z_u)_{u\in\mathrm{Pa}_G(v)}$ is definable, since finite products and coordinate projections preserve definability. The parameter projection $\theta \mapsto \theta_v$ is also semi-algebraic, hence definable. Therefore the map $(x,\theta) \longmapsto \Big((z_u)_{u\in\mathrm{Pa}_G(v)}, \theta_v\Big)$ is definable. Since g_v is definable in $\mathcal{S}$, closure of definable maps under composition implies that $z_v(x,\theta) = (g_v)_{\theta_v}\big((z_u(x,\theta))_{u\in\mathrm{Pa}_G(v)}\big)$ is definable in $\mathcal{S}$.

Since G is finite and acyclic, this induction reaches every computation node. In particular, every output-node state z_v, $v \in \mathrm{out}(G)$, is definable in (x, θ). The final readout $f_\theta(x) = A_{\mathrm{out}}(z_v(x,\theta))_{v\in\mathrm{out}(G)}$ is obtained from definable maps using only finite products, coordinate projections, addition, and multiplication. Hence it is definable in $\mathcal{S}$. Thus f is jointly definable in its inputs and parameters, as claimed. □

Remark B.2. Moreover, if some gates are only defined on definable domains, then the same induction shows that $\mathrm{dom}(f)$ is definable, and f is definable on $\mathrm{dom}(f)$.

B.3 Proofs of General o-Minimal Bounds

The proof does not proceed through the usual metric-complexity route. Our main technical tool is a uniform cell decomposition over the full parameterization, exploiting the o-minimal tameness, or fixed-format Pfaffian regularity, of the parametrized learning rule. Consequently, we do not need to truncate the parameter space to $[-M, M]^p$, estimate covering numbers or Rademacher complexities there, and then control how these estimates propagate through the parameter-to-realization map. That approach is poorly matched to unbounded parameter spaces, since the metric size of $[-M, M]^p$

diverges as $M \to \infty$. The definable route captures a different phenomenon: uniformly over all parameters in $\mathbb{R}^p$, the model family cannot oscillate arbitrarily often. Equivalently, its realizable sign patterns admit a finite uniform cell decomposition. This is the geometric mechanism behind finite sample complexity with unbounded parameters.

Lemma B.3 (Jointly-Definable Families of Classifiers Have Tame Pre-Images)**.** *Let $m, n \in \mathbb{N}_+$, let $X \subseteq \mathbb{R}^m$ be non-empty and definable, and let $f : \mathbb{R}^n \times X \to \mathbb{R}$ be definable (in the same o-minimal setting). Then there exists a constant $K \in \mathbb{N}_+$ such that, for every $\theta \in X$, the set*

$$\big(I_{(0,\infty)} \circ f(\cdot, \theta)\big)^{-1}[\{0\}] = \{x \in \mathbb{R}^n : \ f(x, \theta) \le 0\}$$

has at most K connected components.

Proof of Lemma B.3. Define the jointly-definable classifier

$$g : \mathbb{R}^n \times X \to \{0, 1\}, \qquad g(x, \theta) \stackrel{\text{def.}}{=} I_{(0,\infty)}(f(x, \theta)).$$

Since $(0, \infty)$ is definable and definable maps are closed under composition (cf. (Coste, 1999, Exercise 1.11)), the map g is definable; hence so is

$$A \stackrel{\text{def.}}{=} g^{-1}[\{0\}] = \{(x, \theta) \in \mathbb{R}^n \times X : \ f(x, \theta) \le 0\}.$$

Let $\pi : \mathbb{R}^n \times X \to X$ denote the coordinate projection $(x, \theta) \mapsto \theta$. Apply the cell decomposition theorem to A *compatible with* π (cf. (Coste, 1999, Theorem 2.10)) to obtain finitely many pairwise disjoint definable cells $C_1, \ldots, C_N \subseteq \mathbb{R}^n \times X$ such that

$$A = \bigsqcup_{j=1}^{N} C_j,$$

and such that for each j and each $\theta \in X$ the fibre

$$(C_j)_\theta \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^n : \ (x, \theta) \in C_j\}$$

is either empty or a definable cell in $\mathbb{R}^n$. Now, fix $\theta \in X$ and set

$$A_\theta \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^n : \ (x, \theta) \in A\} = \big(I_{(0,\infty)} \circ f(\cdot, \theta)\big)^{-1}[\{0\}].$$

Then $A_\theta = \bigsqcup_{j=1}^{N} (C_j)_\theta$; whence, by (Coste, 1999, Proposition 2.5), every (non-empty) cell is definably homeomorphic to some Euclidean space, hence connected. Therefore each $(C_j)_\theta$ is either empty or connected, and so A_θ has at most N connected components. Taking $K \stackrel{\text{def.}}{=} N$ completes the proof. □

We next apply the polynomial growth theorem for uniformly definable families to bound the number of classifier traces.

Lemma B.4 (Polynomial Growth of O-Minimal Definable Families)**.** *Let $m, n \in \mathbb{N}_+$, let $Z \subseteq \mathbb{R}^n$ and $X \subseteq \mathbb{R}^m$ be nonempty definable sets in an o-minimal expansion of the real field, and let $f : Z \times X \to \mathbb{R}$ be definable in the same structure. Set*

$$\mathcal{F}^\star \stackrel{\text{def.}}{=} \big\{x \mapsto I_{(0,\infty)}(f(x, \theta)) : \theta \in X\big\}.$$

There exists $C \ge 2$, depending only on this fixed definable family, such that, for every $N \in \mathbb{N}_+$ and every N-point set $S \subseteq Z$,

$$\big|\mathcal{F}^\star|_S\big| \le C N^m. \tag{B.1}$$

Consequently,

$$\mathrm{VCdim}(\mathcal{F}^\star) \le 2 \log_2 C + 2m \log_2(2m). \tag{B.2}$$

Proof of Lemma B.4. Let $\varphi(x; \theta)$ define the relation

$$\big\{(x, \theta) \in Z \times X : f(x, \theta) > 0\big\},$$

with all fixed real parameters named as constants. By (Johnson & Laskowski, 2010, Theorem 4.4, p. 922), there exist finitely many formulas $\psi_1, \dots, \psi_J$, each with variables $(x; y_1, \dots, y_m)$ and $y_i \in \mathbb{R}^n$, such that

$$\forall S \subseteq Z \text{ finite and nonempty}, \quad \forall \theta \in X, \quad \exists j \in \{1, \dots, J\}, \quad \exists (a_1, \dots, a_m) \in S^m,$$
$$\forall x \in S, \qquad \varphi(x; \theta) \iff \psi_j(x; a_1, \dots, a_m).$$

In particular, writing $I_{\psi_j(\cdot; a_1, \dots, a_m)}$ for the indicator of the relation defined by ψ_j, we have

$$\mathcal{F}^\star|_S \subseteq \left\{ I_{\psi_j(\cdot; a_1, \dots, a_m)}|_S : 1 \le j \le J,\ (a_1, \dots, a_m) \in S^m \right\}.$$

Therefore

$$\left|\mathcal{F}^\star|_S\right| \le J|S|^m \le C|S|^m, \qquad C \stackrel{\text{def.}}{=} \max\{2, J\},$$

which proves (B.1). If an N-point set is shattered, then $2^N \le CN^m$. Using $\log_2 u \le u$ for $u > 0$, we obtain

$$\begin{aligned} N &\le \log_2 C + m \log_2 N \\ &= \log_2 C + m \log_2(2m) + m \log_2\left(\frac{N}{2m}\right) \\ &\le \log_2 C + m \log_2(2m) + \frac{N}{2}. \end{aligned}$$

Hence $N \le 2 \log_2 C + 2m \log_2(2m)$. Taking the supremum over the sizes of shattered sets proves (B.2); the conclusion also holds when no nonempty set is shattered. □

Remark B.5 (The One-Dimensional Refinement). In the special case where $Z \subseteq \mathbb{R}$, the preceding conclusion admits a sharper elementary bound. Indeed, suppose that there exists $K \in \mathbb{N}_+$ such that, for every $\theta \in X$, the set $\{x \in Z : f(x, \theta) \le 0\}$ has at most K connected components. Then

$$\mathrm{VCdim}(\mathcal{F}^\star) \le 2K.$$

To see this, define the associated set system

$$\tilde{\mathcal{F}} \stackrel{\text{def.}}{=} \left\{ \{x \in Z : f(x, \theta) \le 0\} \ : \ \theta \in X \right\}.$$

Since $x \mapsto I_{(0,\infty)}\big(f(x, \theta)\big) = 1$ if and only if $x \notin \{y \in Z : f(y, \theta) \le 0\}$, the classes $\mathcal{F}^\star$ and $\tilde{\mathcal{F}}$ induce the same labelings on every finite sample up to flipping 0 and 1. Hence

$$\mathrm{VCdim}(\mathcal{F}^\star) = \mathrm{VCdim}(\tilde{\mathcal{F}}).$$

Now let $S = \{x_1 < \cdots < x_{2K+1}\} \subseteq Z$. Since every set $\{x \in Z : f(x, \theta) \le 0\}$ has at most K connected components, and since connected components in $\mathbb{R}$ are intervals, its trace on S is a union of at most K blocks of consecutive points. Therefore it cannot realize the alternating subset $\{x_1, x_3, x_5, \dots, x_{2K+1}\}$, which consists of $K + 1$ singleton blocks. Thus no $(2K + 1)$-point subset of Z is shattered, and consequently $\mathrm{VCdim}(\mathcal{F}^\star) \le 2K$.

This now allows us to obtain classifier bounds via the core result in VC theory. We propose and prove the following more general formulation of Theorem 3.1.

Theorem B.6. *Let $m, n \in \mathbb{N}_+$, let $X \subseteq \mathbb{R}^m$ be non-empty and definable, and let $f : \mathbb{R}^n \times X \to \mathbb{R}$ be definable (in the same o-minimal setting). Set*

$$\mathcal{F}^\star \stackrel{\text{def.}}{=} \left\{ I_{(0,\infty)} \circ f(\cdot, \theta) \ : \ \theta \in X \right\},$$

so that, for each $\theta \in X$,

$$\big(I_{(0,\infty)} \circ f(\cdot, \theta)\big)^{-1}[\{0\}] = \{x \in \mathbb{R}^n : f(x, \theta) \le 0\}.$$

Let $K \stackrel{\text{def.}}{=} \max\{1, \mathrm{VCdim}(\mathcal{F}^\star)\} < \infty$, and let ℓ denote the 0–1 loss. Then, there exists an absolute constant $C > 0$ such that: for every Borel probability measure $\mathbb{P}$ on $\mathbb{R}^n \times \{0, 1\}$, every sequence of i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^\infty$ with law $\mathbb{P}$, every error $\varepsilon > 0$, and every failure probability $0 < \delta \le 1$, the following holds with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{F}^\star} \left| \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[\ell(h(X), Y)\big] - \frac{1}{N} \sum_{i=1}^{N} \ell(h(X_i), Y_i) \right| \le \varepsilon$$

provided that

$$N \ge \frac{C}{\varepsilon^2} \left[K + \ln\left(\frac{2}{\delta}\right) \right].$$

Proof of Theorem B.6. By Lemma B.4, $d \overset{\text{def.}}{=} \mathrm{VCdim}(\mathcal{F}^\star) < \infty$ and $K = \max\{1, d\}$. The map $((x, y), \theta) \mapsto \ell(I_{(0,\infty)}(f(x, \theta)), y)$, is jointly definable and takes values in $\{0, 1\}$. It is therefore jointly Borel measurable, and Lemma B.1 establishes measurability of the uniform-deviation event under $\overline{\mathbb{P}^N}$. If $d \geq 1$, apply (Shalev-Shwartz & Ben-David, 2014, Theorem 6.8(1)). If $d = 0$, every two members of $\mathcal{F}^\star$ agree at every input: otherwise a singleton would be shattered. Thus the loss class contains a single function, and Hoeffding's inequality gives

$$\overline{\mathbb{P}^N}\left(\sup_{h \in \mathcal{F}^\star} |\mathbb{P}\ell_h - \mathbb{P}_N \ell_h| > \varepsilon\right) \leq 2e^{-2N\varepsilon^2}, \qquad \ell_h(x, y) \overset{\text{def.}}{=} \ell(h(x), y).$$

Both cases yield the stated bound after choosing C universally. □

In a similar vein, pseudo-dimension bounds can be obtained from uniform cell decompositions of the type in Lemma B.4. Indeed, a standard lifting argument applied to the hypothesis class reduces the pseudo-dimension estimate to a VC dimension bound for an associated class of subgraphs. Once this reduction is in place, the proof is immediate: we invoke the preceding lemma that converts uniform cell decompositions into VC bounds, and then apply the main generalization theorem of Alon et al. (1997).

Lemma B.7 (Standard Lifting Trick)**.** *Let $m, n \in \mathbb{N}_+$, let $Z \subseteq \mathbb{R}^n$ and $X \subseteq \mathbb{R}^m$ be non-empty, and let $f : Z \times X \to \mathbb{R}$. Define the real-valued class*

$$\mathcal{F} \overset{\text{def.}}{=} \Big\{ x \mapsto f(x, \theta) \ : \ \theta \in X \Big\}.$$

For each $\gamma \geq 0$, define the lifted classifier class on $Z \times \mathbb{R}$ by

$$\mathcal{H}_\gamma \overset{\text{def.}}{=} \Big\{ (x, r) \mapsto I_{(0,\infty)}\big(f(x, \theta) - r - \gamma\big) \ : \ \theta \in X \Big\}.$$

Then, for every $\gamma > 0$, we have $\mathrm{fat}_\gamma(\mathcal{F}) \leq \mathrm{Pdim}(\mathcal{F}) \leq \mathrm{VCdim}(\mathcal{H}_0)$.

Proof of Lemma B.7. For $\gamma \geq 0$ and $\theta \in X$, define $h_\theta^{(\gamma)} : Z \times \mathbb{R} \to \{0, 1\}$ by $h_\theta^{(\gamma)}(x, r) \overset{\text{def.}}{=} I_{(0,\infty)}\big(f(x, \theta) - r - \gamma\big)$, so that $\mathcal{H}_\gamma = \{h_\theta^{(\gamma)} : \theta \in X\}$.
Fix $d \in \mathbb{N}_+$. By definition, $\{x_1, \ldots, x_d\} \subseteq Z$ is pseudo-shattered by $\mathcal{F}$ if and only if there exist $r_1, \ldots, r_d \in \mathbb{R}$ such that, for every $b \in \{0, 1\}^d$, one can choose $\theta \in X$ with

$$I_{(0,\infty)}\big(f(x_i, \theta) - r_i\big) = b_i$$

for each $i = 1, \ldots, d$. Setting $z_i \overset{\text{def.}}{=} (x_i, r_i) \in Z \times \mathbb{R}$, this implies that, for every $b \in \{0, 1\}^d$, there exists $\theta \in X$ such that $h_\theta^{(0)}(z_i) = b_i$ for each $i = 1, \ldots, d$. Hence $\{z_1, \ldots, z_d\}$ is shattered by $\mathcal{H}_0$. Taking suprema over d yields $\mathrm{Pdim}(\mathcal{F}) \leq \mathrm{VCdim}(\mathcal{H}_0)$.
Now fix $\gamma > 0$ and suppose that $\{x_1, \ldots, x_d\} \subseteq Z$ is γ-fat shattered by $\mathcal{F}$, witnessed by $r_1, \ldots, r_d \in \mathbb{R}$. Thus, for every labeling $b \in \{0, 1\}^d$, one can choose $\theta \in X$ so that, for each $i = 1, \ldots, d$, $b_i = 1$ implies $f(x_i, \theta) \geq r_i + \gamma$ and $b_i = 0$ implies $f(x_i, \theta) \leq r_i - \gamma$. Since $\gamma > 0$, this implies

$$I_{(0,\infty)}\big(f(x_i, \theta) - r_i\big) = b_i$$

for each $i = 1, \ldots, d$. Hence $\{x_1, \ldots, x_d\}$ is pseudo-shattered by $\mathcal{F}$ with witnesses $r_1, \ldots, r_d$. Therefore $d \leq \mathrm{Pdim}(\mathcal{F})$, and taking suprema over all such d yields $\mathrm{fat}_\gamma(\mathcal{F}) \leq \mathrm{Pdim}(\mathcal{F})$. □

In order to obtain our generalization bound from our uniform cell decomposition, guaranteed for definable families of parametric functions, within an o-minimal structure, we only require the following repackaged result of Alon et al. (1997).

Lemma B.8 (Finite Pseudo-Dimension Implies Statistical Guarantees)**.** *Let $p, n, m \in \mathbb{N}_+$, let $\Theta \subseteq \mathbb{R}^p$ be nonempty and Borel, let $f : \mathbb{R}^n \times \Theta \to \mathbb{R}^m$ be jointly Borel measurable, and let $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0, 1]$ be Borel measurable. Define*

$$h_\theta(x, y) \overset{\text{def.}}{=} \ell(f(x, \theta), y), \qquad \mathcal{F}_\ell \overset{\text{def.}}{=} \{h_\theta : \theta \in \Theta\}, \qquad d \overset{\text{def.}}{=} \max\{1, \mathrm{Pdim}(\mathcal{F}_\ell)\}, \tag{B.3}$$

and suppose that $d < \infty$. There exists a universal constant $C > 0$ such that, for every Borel probability measure $\mathbb{P}$ on $\mathbb{R}^{n+m}$ and every $0 < \varepsilon, \delta < 1$,

$$\overline{\mathbb{P}^N}\left(\sup_{\theta\in\Theta}\left|\int h_\theta \, d\mathbb{P} - \frac{1}{N}\sum_{i=1}^{N} h_\theta(Z_i)\right| \le \varepsilon\right) \ge 1 - \delta \tag{B.4}$$

whenever

$$N \ge \frac{C}{\varepsilon^2}\left[d\ln^2\left(\frac{ed}{\varepsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right]. \tag{B.5}$$

Here $Z_i = (X_i, Y_i)$ are the coordinate samples under $\mathbb{P}^N$.

Proof of Lemma B.8. The map $(z, \theta) \mapsto h_\theta(z)$ is jointly Borel measurable. Since Θ is standard Borel, it admits a Polish topology generating the same Borel σ-algebra. Thus $\mathcal{F}_\ell$ has a jointly Borel parametrization by a Polish space and satisfies the image-admissible Suslin hypothesis used in (Alon et al., 1997, Section 1). Lemma B.1 also shows that the event in (B.4) is measurable under $\overline{\mathbb{P}^N}$.

Set

$$q_\varepsilon \overset{\text{def.}}{=} \operatorname{fat}_{\varepsilon/24}(\mathcal{F}_\ell) \le \operatorname{Pdim}(\mathcal{F}_\ell) \le d, \tag{B.6}$$

where the first inequality follows from Lemma B.7. If $q_\varepsilon \ge 1$, then (Alon et al., 1997, Theorem 3.6) gives (B.4) whenever

$$N \ge \frac{C_0}{\varepsilon^2}\left[q_\varepsilon \ln^2\left(\frac{eq_\varepsilon}{\varepsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right] \tag{B.7}$$

for a universal $C_0 > 0$. For $u \ge 1$,

$$\frac{d}{du}\left[u\ln^2\left(\frac{eu}{\varepsilon}\right)\right] = \ln^2\left(\frac{eu}{\varepsilon}\right) + 2\ln\left(\frac{eu}{\varepsilon}\right) > 0.$$

Consequently, (B.5) implies (B.7) whenever $C \ge C_0$.

It remains to treat $q_\varepsilon = 0$. Fix $h_0 \in \mathcal{F}_\ell$. If $|h(z) - h_0(z)| > \varepsilon/12$ for some $h \in \mathcal{F}_\ell$ and z, then the threshold $(h(z) + h_0(z))/2$ witnesses that $\{z\}$ is $\varepsilon/24$-fat shattered, a contradiction. Hence

$$\sup_{h\in\mathcal{F}_\ell}\sup_{z\in\mathbb{R}^{n+m}}|h(z) - h_0(z)| \le \frac{\varepsilon}{12}.$$

Writing $\mathbb{P}_N h \overset{\text{def.}}{=} N^{-1}\sum_{i=1}^{N} h(Z_i)$, we obtain

$$\sup_{h\in\mathcal{F}_\ell}|\mathbb{P}h - \mathbb{P}_N h| \le |\mathbb{P}h_0 - \mathbb{P}_N h_0| + \frac{\varepsilon}{6}.$$

Since $0 \le h_0 \le 1$, Hoeffding's inequality yields

$$\overline{\mathbb{P}^N}\left(\sup_{h\in\mathcal{F}_\ell}|\mathbb{P}h - \mathbb{P}_N h| > \varepsilon\right) \le 2\exp\left(-\frac{25}{18}N\varepsilon^2\right).$$

The right-hand side is at most δ under (B.5) after enlarging C universally. □

Remark B.9 (Lipschitz Losses May Destroy Definability). Now, if we only knew that the loss function ℓ was Lipschitz, without assuming definability in an o-minimal structure, then there would be no guarantee that the pseudo-dimension of the composite class $\mathcal{F}_\ell$ would remain finite even if that of $\mathcal{F}$ is finite. For instance, if $n = m = 1$, $\mathcal{F} = \{ax : a \in [0, 2\pi]\}$ and $\ell(y_1, y_2) \overset{\text{def.}}{=} \frac{\sin(y_1)+1}{2}$, then $\mathcal{F}$ has finite pseudo-dimension, but the composite class need not have finite pseudo-dimension.

Remark B.10. Note also that we did not pass from the fat-shattering dimension to covering number bounds, e.g., using a result such as (Bartlett et al., 1997, Theorem 2), which may unnecessarily loosen our bounds.

We now prove the regression variant of Theorem 3.2 for definable regressors.

Theorem B.11 (Definability Implies Learning Is Possible—Regression Case)**.** *Let $m, n, p \in \mathbb{N}_+$, let $X \subseteq \mathbb{R}^p$ be non-empty and definable, let $f : \mathbb{R}^n \times X \to \mathbb{R}^m$ be definable, and let $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0, 1]$ be definable. Set*

$$\mathcal{F} \overset{\text{def.}}{=} \{x \mapsto f(x, \theta) : \theta \in X\}.$$

Then there exist constants $C, K > 0$ such that, for any Borel probability measure $\mathbb{P}$ on $\mathbb{R}^{n+m}$, any i.i.d. random variables $(X_1, Y_1), \ldots, (X_N, Y_N)$ with law $\mathbb{P}$, any accuracy parameter $\varepsilon > 0$, and any failure probability $0 < \delta \leq 1$, the following bound holds with probability at least $1 - \delta$:

$$\sup_{g \in \mathcal{F}} \left| \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[\ell(g(X), Y)\big] - \frac{1}{N} \sum_{i=1}^{N} \ell(g(X_i), Y_i) \right| \leq \varepsilon$$

provided that

$$N \geq \frac{C}{\varepsilon^2} \left[K \ln^2 \left(\frac{eK}{\varepsilon} \right) + \ln \left(\frac{2}{\delta} \right) \right].$$

Proof of Theorem B.11. Define $h((x, y), \theta) \overset{\text{def.}}{=} \ell(f(x, \theta), y)$ and $H((x, y, r), \theta) \overset{\text{def.}}{=} h((x, y), \theta) - r$. Both maps are jointly definable in the given o-minimal structure. Applying Lemma B.4 to $H : \mathbb{R}^{n+m+1} \times X \to \mathbb{R}$ gives

$$\operatorname{VCdim}\big(\{(x, y, r) \mapsto I_{(0,\infty)}(H((x, y, r), \theta)) : \theta \in X\}\big) < \infty.$$

Therefore Lemma B.7 implies

$$\operatorname{Pdim}\big(\{h(\cdot, \theta) : \theta \in X\}\big) < \infty.$$

Set $K \overset{\text{def.}}{=} \max\big\{1, \operatorname{Pdim}\big(\{h(\cdot, \theta) : \theta \in X\}\big)\big\}$. Definable sets and maps are Borel measurable, so Lemma B.8 applies with $\Theta = X$ and $d = K$. □

B.4 Proofs of Quantitative Estimates via Pfaffian Nonlinearities

We now obtain quantitative bounds for fixed architectures whose selected gates from the dictionary $\mathbb{G}$ in (2.1) are globally Pfaffian jointly in their inputs and parameters. We retain the network notation $f(x, \theta)$ and the total parameter count P from Definition 2.4, writing $n = d_{in}$ and $m = d_{out}$. The Pfaffian complexity is controlled by the formats of a layered representation of the architecture.

Definition B.12 (Pfaffian Neural Networks and Pfaffian Layers)**.** Fix $n, m, P, L \in \mathbb{N}_+$ and layer dimensions $n_0, \ldots, n_L \in \mathbb{N}_+$, with $n_0 = n$ and $n_L = m$. For $1 \leq k \leq L$, a *trainable Pfaffian layer* is a map $g_k : \mathbb{R}^{n_{k-1}} \times \mathbb{R}^P \to \mathbb{R}^{n_k}$ whose coordinates are Pfaffian jointly in their input and parameter variables over a common chain. More precisely, write $g_k = (g_{k,1}, \ldots, g_{k,n_k})$ and suppose that there is a Pfaffian chain $f_1^k, \ldots, f_{q_k}^k$ of degree at most D_k on all of $\mathbb{R}^{n_{k-1}+P}$ such that

$$g_{k,i}(u, \theta) = Q_{k,i}\big(u, \theta, f_1^k(u, \theta), \ldots, f_{q_k}^k(u, \theta)\big), \qquad 1 \leq i \leq n_k,$$

where $u \in \mathbb{R}^{n_{k-1}}$, $\theta \in \mathbb{R}^P$, and $Q_{k,i}$ is a polynomial of total degree at most $d_{k,i}$. The chain and format conventions are those of Definitions A.1 and A.2. Here $q_k, D_k \in \mathbb{N}_0$ and $d_{k,i} \in \mathbb{N}_+$. We allow an empty chain, with $q_k = D_k = 0$, for polynomial layers.

A $\mathbb{G}$-neural network $f : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}^m$ is called a *Pfaffian neural network* (PNN) if its fixed architecture admits a representation by such layers,

$$f(x, \theta) \overset{\text{def.}}{=} \big((g_L)_\theta \circ \cdots \circ (g_1)_\theta\big)(x), \qquad (x, \theta) \in \mathbb{R}^n \times \mathbb{R}^P, \tag{B.8}$$

using the parameter-specialization notation from (2.1). The layer maps $g_1, \ldots, g_L$ and their order are fixed, and the realization family is $\mathcal{F} \overset{\text{def.}}{=} \{f_\theta : \theta \in \mathbb{R}^P\}$.

The layered representation is compatible with the DAG in Definition 2.4. Order its computation nodes topologically and keep a state containing the input and all previously computed node states. At each step, a layer appends the next value in (2.2), using fixed coordinate projections to select its parent states and its parameter block, while copying the existing state through identity coordinates. Each update is globally jointly Pfaffian whenever its selected gate is; its coordinates share the pullback of

the gate's chain under the fixed coordinate projections, since the copied coordinates are polynomials. Include the fixed input lift and the trainable readout (2.3) in the representation. This gives (B.8), possibly with additional state coordinates. Thus a layer may be a state update assembled from the original gates; the primitive dictionary $\mathbb{G}$ need not contain the assembled update as a separate gate.

Each layer may use only a subtuple of the full parameter vector θ. The same format calculations also apply after parameters are fixed or shared: express the resulting subfamily in its independent free coordinates and use their number for P, computing the layer formats after this affine substitution. If no free parameter remains, the realization family is a singleton and has pseudo-dimension zero. The common-chain requirement is explicit: separate coordinate chains can be concatenated, but their lengths must then be included in q_k.

Remark B.13. All degree specifications are upper bounds. We take $d_{k,i} \geq 1$ without excluding constant nodes: a polynomial of degree zero also has degree at most one. A trainable constant, such as $g_{k,i}(u, \theta) = \theta_j$, has joint degree one. A trainable affine map $(u, A, b) \mapsto Au + b$ has joint degree at most two. In the unrestricted parameterization of Definition 2.4, the entries of the readout A_{out} are included in θ and in P. Its joint degree must be included in the format of the final layer.

Notation. *The architecture and layer formats of a PNN are summarized by*

$$((n, m, P), (q_k, D_k, d_{k,i}), 1 \leq i \leq n_k, 1 \leq k \leq L).$$

Consistently with the latent-state notation in (2.2), write $z_{k,i}$ for the accumulated output at coordinate i of layer k. Set $z_{0,i}(x, \theta) \stackrel{\text{def.}}{=} x_i$ for $1 \leq i \leq n$, and define

$$z_{k,i}(x, \theta) \stackrel{\text{def.}}{=} g_{k,i}\big(z_{k-1,1}(x, \theta), \ldots, z_{k-1,n_{k-1}}(x, \theta), \theta\big).$$

Here $1 \leq k \leq L$ and $1 \leq i \leq n_k$, so $f_i = z_{L,i}$ for $1 \leq i \leq m$.

B.5 Bounded Constants Under Pfaffianity

The preceding results give finite statistical complexity over the full, possibly unbounded parameter space. Their regression conclusion concerns the bounded loss in Theorem 3.2. We now obtain explicit pseudo-dimension bounds for globally Pfaffian networks, followed by a quantitative guarantee for clipped squared loss. The bounds depend on the joint Pfaffian formats in the inputs and trainable parameters.

Theorem B.14 (Bounding the Pseudo-Dimension of Pfaffian Neural Networks)**.** *Suppose that a PNN $f : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}^m$ is summarized by $((n, m, P), (q_k, D_k, d_{k,j}), 1 \leq j \leq n_k, 1 \leq k \leq L)$, as in Definition B.12. Fix an output coordinate $i \in \{1, \ldots, m\}$ and let $\mathcal{F}_i \stackrel{\text{def.}}{=} \{f_i(\cdot, \theta) : \theta \in \mathbb{R}^P\}$. Set $\delta_k \stackrel{\text{def.}}{=} \max_{1 \leq j \leq n_k} d_{k,j}$ and*

$$q \stackrel{\text{def.}}{=} \sum_{k=1}^{L} q_k, \qquad D \stackrel{\text{def.}}{=} \sum_{k=1}^{L} (D_k + 1) \prod_{t=1}^{k-1} \delta_t - L, \qquad d \stackrel{\text{def.}}{=} d_{L,i} \prod_{k=1}^{L-1} \delta_k,$$

where an empty product equals 1. Writing $A \stackrel{\text{def.}}{=} P(D + d - 1) + 1$, we have

$$\text{P-dim}(\mathcal{F}_i) \leq 16P + Pq(Pq - 1) + 2P \log_2 d + 2P(q + 1) \log_2 A + 2Pq \log_2(P + 1).$$

In particular, when $m = 1$, this bounds the pseudo-dimension of the realization family $\mathcal{F}$ from Definition B.12.

The Quantitative Agnostic PAC-Learning Bound For a scalar output in Theorem B.14, let q, D, d be its format bounds and, for $b \geq 1$, write

$$\begin{aligned} K(b) \stackrel{\text{def.}}{=}\ & 16P + Pq(Pq - 1) + 2P \log_2 b \\ & + 2P(q + 1) \log_2\big(P(D + b - 1) + 1\big) + 2Pq \log_2(P + 1). \end{aligned} \tag{B.9}$$

The binary class $\{I_{(0,\infty)} \circ f_\theta : \theta \in \mathbb{R}^P\}$ has VC dimension at most $K(d)$, since its shattering is a special case of pseudo-shattering with all thresholds zero. For regression, the complexity of the loss must also be accounted for. The following gives an explicit instance in the bounded-loss setting of Theorem 3.2.

Theorem B.15 (Quantitative Learning with Clipped Squared Loss)**.** *Suppose that $f : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}$ satisfies the hypotheses of Theorem B.14. Set $\ell(u, y) \overset{\text{def.}}{=} \min\{1, (u-y)^2\}$ and $K_{\text{sq}} \overset{\text{def.}}{=} K(2d)$. Then $\text{P-dim}(\mathcal{F}_\ell) \le K_{\text{sq}}$, with $\mathcal{F}_\ell$ as in (B.3) and $\Theta = \mathbb{R}^P$. Use the risk notation R, $\widehat{R}_s$, and Δ_N from Subsection B.1, with $\mathcal{Z} = \mathbb{R}^{n+1}$ and $h((x, y), \theta) = \ell(f_\theta(x), y)$. For every Borel probability measure $\mathbb{P}$ on $\mathcal{Z}$, every i.i.d. sample $s = ((X_i, Y_i))_{i=1}^N$ with law $\mathbb{P}^N$, and $0 < \varepsilon, \delta < 1$, we have*

$$\overline{\mathbb{P}^N}\big(\Delta_N(s) \le \varepsilon\big) \ge 1 - \delta$$

provided that

$$N \ge \frac{C}{\varepsilon^2}\left[K_{\text{sq}} \ln^2\left(\frac{eK_{\text{sq}}}{\varepsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right], \tag{B.10}$$

where $C > 0$ is universal. For each $\eta > 0$, a Borel measurable empirical η-minimizer $\widehat{\theta}_{N,\eta}$ exists, and on the same event it satisfies

$$R\big(\widehat{\theta}_{N,\eta}(s)\big) - \inf_{\theta \in \mathbb{R}^P} R(\theta) \le 2\varepsilon + \eta.$$

Proof of Theorem B.15. Define the "effective dimension" quantity set $C_{\text{gate}} \overset{\text{def.}}{=} q_0^2 + 12(q_0 + 1)\left[2 + \log_2(d_0(D_0 + 2))\right]$, $K_{\text{sq}} \overset{\text{def.}}{=} C_{\text{gate}} P^2 U^2$, and note that $K_{\text{sq}} \in \mathcal{O}(P^2 U^2 L^2)$. Consider the sub-graphs of $(x, y) \mapsto \ell(f_\theta(x), y)$. Any shattered collection (x_j, y_j, t_j) must have $0 \le t_j < 1$: a threshold below zero is always exceeded, and a threshold at least one is never exceeded. For $0 \le t < 1$, the inequality $\ell(f_\theta(x), y) > t$ is equivalent to $(f(x, \theta) - y)^2 - t > 0$. The latter expression is globally Pfaffian of format at most $(q, D, 2d)$: extend the chain for f constantly in y, t, and use $d \ge 1$ and the polynomial representation of f. The proof of Theorem B.14, applied with (x, y, t) as data and θ as the parameter vector, therefore bounds the VC dimension of these subgraphs by $K(2d)$. This gives $\text{P-dim}(\mathcal{F}_\ell) \le K_{\text{sq}}$ without requiring the clipped loss to be analytic.

Since $K_{\text{sq}} \ge 16$, the quantity $\max\{1, \text{P-dim}(\mathcal{F}_\ell)\}$ in Lemma B.8 is at most K_{sq}. That lemma gives (B.10), using the monotonicity of $v \mapsto v \ln^2(ev/\varepsilon)$ for $v \ge 1$ and $0 < \varepsilon < 1$. Moreover, f is globally Pfaffian and hence jointly definable in the Pfaffian closure of the real field, as recalled in Subsection A.1. Since ℓ is semi-algebraic, the composite map h is jointly definable in the same structure. Lemma B.1, with the stated $\mathcal{Z}$ and Θ, supplies a Borel measurable empirical η-minimizer for every $\eta > 0$ and gives

$$R\big(\widehat{\theta}_{N,\eta}(s)\big) - \inf_{\theta \in \mathbb{R}^P} R(\theta) \le 2\Delta_N(s) + \eta.$$

On $\{\Delta_N \le \varepsilon\}$ this is the claimed excess-risk bound. The same comparison holds with $\eta = 0$ whenever a measurable exact empirical minimizer is available. □

Remark B.16. This corollary bounds clipped squared risk. For another bounded definable loss, Theorem 3.2 gives qualitative learnability, while an explicit constant requires a corresponding complexity estimate for its composite loss class.

Lemma B.17 (Pfaffian Complexity of $\mathbb{G}$-Neural Networks)**.** *Let f be a $\mathbb{G}$-neural network as in Definition 2.4. Suppose that, for every $v \in \text{comp}(G)$ and $1 \le j \le m_v$, the coordinate $(u, \theta_v) \mapsto g_{v,j}(u, \theta_v)$ is globally Pfaffian on $\mathbb{R}^{M_v + p_v}$ of format at most (q_0, D_0, d_0), where $q_0, D_0 \in \mathbb{N}_0$ and $d_0 \in \mathbb{N}_+$. No common chain between different coordinates is assumed. Fix $L \in \mathbb{N}_+$ and a partition*

$$\text{comp}(G) = \bigsqcup_{k=1}^{L} V_k, \qquad V_k \ne \varnothing, \qquad \text{Pa}_G(v) \cap \text{comp}(G) \subseteq \bigsqcup_{j<k} V_j \quad (v \in V_k).$$

Set $W_k \overset{\text{def.}}{=} \sum_{v \in V_k} m_v$, $U \overset{\text{def.}}{=} \sum_{k=1}^{L} W_k$, and $W \overset{\text{def.}}{=} \max_{1 \le k \le L} W_k$. Then $L \le U \le LW$, and f admits a PNN representation with $L + 1$ layers and the same parameter vector $\theta \in \mathbb{R}^P$. The first L layers have common-chain format bounds $(q_0 W_k, D_0, d_0)$, while the last layer has format bound $(0, 0, 2)$.

All coordinates of $(x, \theta) \mapsto f(x, \theta)$ are globally Pfaffian over a common chain, with format at most (q, D, d), where

$$q \overset{\text{def.}}{=} q_0 U, \qquad D \overset{\text{def.}}{=} (D_0 + 1)\sum_{j=0}^{L-1} d_0^j - L, \qquad d \overset{\text{def.}}{=} d_0^L + 1. \tag{B.11}$$

For the coordinate realization classes $\mathcal{F}_i$ of Theorem B.14, put

$$H \stackrel{\text{def.}}{=} \log_2\big((P+1)(L+1)(D_0+2)\big) + L\log_2 d_0.$$

Then, for every $1 \le i \le d_{out}$,

$$\text{P-dim}(\mathcal{F}_i) \le P^2 q_0^2 U^2 + 12P(q_0 U + 1)H. \tag{B.12}$$

In particular, with

$$C_{\text{gate}} \stackrel{\text{def.}}{=} q_0^2 + 12(q_0+1)\left[2 + \log_2\big(d_0(D_0+2)\big)\right],$$

we have

$$\text{P-dim}(\mathcal{F}_i) \le C_{\text{gate}} P^2 U^2 \le C_{\text{gate}} P^2 L^2 W^2. \tag{B.13}$$

Thus, for fixed (q_0, D_0, d_0), the bound is $O(P^2U^2)$, and hence $O(P^2L^2W^2)$.

If, in addition, the coordinates of each entire gate g_v admit a common chain of length at most q_0 and degree at most D_0, with representing polynomial degrees at most d_0, set $T \stackrel{\text{def.}}{=} |\operatorname{comp}(G)|$. One may then take $q = q_0 T$ in (B.11), and $\text{P-dim}(\mathcal{F}_i) \le C_{\text{gate}} P^2 T^2$.

Proof. Fix an ordering within each V_k, preserving the coordinate order of every gate. Define

$$s_0 \stackrel{\text{def.}}{=} x, \qquad s_k \stackrel{\text{def.}}{=} \big(x, (z_v)_{v \in V_1 \cup \cdots \cup V_k}\big), \qquad n_k \stackrel{\text{def.}}{=} d_{in} + \sum_{j=1}^{k} W_j \quad (0 \le k \le L),$$

where the sum is empty for $k = 0$, and put $n_{L+1} \stackrel{\text{def.}}{=} d_{out}$. The condition on the parent sets yields fixed linear maps

$$C_v : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{M_v}, \qquad E_v : \mathbb{R}^P \to \mathbb{R}^{p_v}, \qquad v \in V_k,$$
$$C_v s_{k-1} = (z_u)_{u \in \text{Pa}_G(v)}, \qquad E_v \theta = \theta_v.$$

Here C_v recovers input-node coordinates from Πx, and E_v is the coordinate projection onto the parameter block of g_v. Likewise, a fixed linear map $C_{\text{out}} : \mathbb{R}^{n_L} \to \mathbb{R}^{M_{\text{out}}}$ satisfies $C_{\text{out}} s_L = (z_v)_{v \in \text{out}(G)}$.

For every coordinate $g_{v,j}$, choose a witnessing chain on $\mathbb{R}^{M_v + p_v}$ of length at most q_0 and degree at most D_0. Concatenate these chains over $1 \le j \le m_v$, obtaining a common chain $(\varphi_{v,1}, \dots, \varphi_{v,r_v})$ for g_v, with $r_v \le m_v q_0$. Its differential degree remains at most D_0: each differential polynomial for an appended chain is independent of all earlier appended chains. Thus there are polynomials $T_{v,j,b}$ and $Q_{v,c}$ such that, for $y \in \mathbb{R}^{M_v + p_v}$,

$$\partial_{y_b} \varphi_{v,j}(y) = T_{v,j,b}\big(y, \varphi_{v,1}(y), \dots, \varphi_{v,j}(y)\big), \qquad \deg T_{v,j,b} \le D_0,$$
$$g_{v,c}(y) = Q_{v,c}\big(y, \varphi_{v,1}(y), \dots, \varphi_{v,r_v}(y)\big), \qquad \deg Q_{v,c} \le d_0,$$

where $1 \le j \le r_v$, $1 \le b \le M_v + p_v$, and $1 \le c \le m_v$. For $v \in V_k$, set

$$J_v(u, \theta) \stackrel{\text{def.}}{=} (C_v u, E_v \theta), \qquad \psi_{v,j} \stackrel{\text{def.}}{=} \varphi_{v,j} \circ J_v.$$

For every coordinate ξ of (u, θ),

$$\partial_\xi \psi_{v,j} = \sum_{b=1}^{M_v + p_v} (\partial_\xi J_{v,b}) T_{v,j,b}\big(J_v, \psi_{v,1}, \dots, \psi_{v,j}\big).$$

Since J_v is fixed linear, the coefficients $\partial_\xi J_{v,b}$ are constant; hence the right-hand side has polynomial degree at most D_0. The representing polynomial of $g_{v,c} \circ J_v$ has degree at most d_0. Concatenating these pulled-back chains over $v \in V_k$ therefore gives one global chain of length at most $q_0 W_k$ and degree at most D_0 for all coordinates of $(g_v(C_v u, E_v \theta))_{v \in V_k}$.

Define

$$\Gamma_k(u, \theta) \stackrel{\text{def.}}{=} \big(u, (g_v(C_v u, E_v \theta))_{v \in V_k}\big) \quad (1 \le k \le L), \qquad \Gamma_{L+1}(u, \theta) \stackrel{\text{def.}}{=} A_{\text{out}} C_{\text{out}} u.$$

The copied coordinates have degree one, which is at most d_0. Thus Γ_k has the asserted common-chain format bounds, while Γ_{L+1} is polynomial of joint degree at most two. Equations (2.2) and (2.3) imply

$$s_k = (\Gamma_k)_\theta(s_{k-1}), \qquad f(x,\theta) = \big((\Gamma_{L+1})_\theta \circ \cdots \circ (\Gamma_1)_\theta\big)(x).$$

The maps $C_v, E_v, C_{\mathrm{out}}$ and the copied coordinates introduce no parameters. Consequently, this is a PNN representation in the sense of Definition B.12, with precisely the original parameter vector θ.

Set

$$\widehat{q}_k \overset{\text{def.}}{=} q_0 \sum_{j=1}^{k} W_j, \qquad \widehat{D}_k \overset{\text{def.}}{=} (D_0+1)\sum_{j=0}^{k-1} d_0^j - k, \qquad \widehat{d}_k \overset{\text{def.}}{=} d_0^k \quad (0 \le k \le L).$$

We prove by induction that the coordinates of $s_k(x,\theta)$ are globally Pfaffian over a common chain of length at most $\widehat{q}_k$ and degree at most $\widehat{D}_k$, with representing polynomial degrees at most $\widehat{d}_k$. At $k=0$, the empty chain suffices. At step k, the coordinates of $(s_{k-1}(x,\theta),\theta)$ have representing polynomial degrees at most $d_0^{k-1} \ge 1$ over the preceding chain. Lemmas B.25 and B.26, equivalently the induction in Proposition B.29, give

$$\begin{aligned}
\widehat{q}_k &= \widehat{q}_{k-1} + q_0 W_k, \\
\widehat{D}_k &= \widehat{D}_{k-1} + (D_0+1)d_0^{k-1} - 1, \\
\widehat{d}_k &= d_0 \widehat{d}_{k-1} = d_0^k.
\end{aligned}$$

The witnessing chain is common to all coordinates, completing the induction. Denote its value at layer L by $\zeta(x,\theta) \in \mathbb{R}^\ell$, where $\ell \le q$. Each coordinate of $C_{\mathrm{out}} s_L$ has the form $R_j(x,\theta,\zeta(x,\theta))$, with $\deg R_j \le d_0^L$. Therefore

$$f_i(x,\theta) = \sum_{j=1}^{M_{\mathrm{out}}} (A_{\mathrm{out}})_{ij} R_j(x,\theta,\zeta(x,\theta)), \qquad \deg\left(\sum_j (A_{\mathrm{out}})_{ij} R_j\right) \le d_0^L + 1.$$

The readout changes neither the chain length nor its differential degree, proving (B.11).

We apply the counting argument in the proof of Theorem B.14 to this final joint format. Put $A \overset{\text{def.}}{=} P(D+d-1)+1$ and fix i. The function $\widetilde{f}_i(x,t,\theta) \overset{\text{def.}}{=} f_i(x,\theta) - t$ has format at most (q, D, d), with (x,t) as data and $\theta \in \mathbb{R}^P$ as parameters. Fix $1 \le r \le P$, data points $(x_j, t_j) \in \mathbb{R}^{d_{in}+1}$, $1 \le j \le r$, and a regular value $\varepsilon \in \mathbb{R}^r$ of

$$\Xi(\theta) \overset{\text{def.}}{=} \big(\widetilde{f}_i(x_j,t_j,\theta)\big)_{j=1}^{r}.$$

Put $\chi_j(\theta) \overset{\text{def.}}{=} \widetilde{f}_i(x_j,t_j,\theta) - \varepsilon_j$. Corollary B.28 supplies a common global chain for $\chi_1,\dots,\chi_r$ of length at most rq and degree at most D, with output degrees at most d. Since 0 is a regular value of $(\chi_1,\dots,\chi_r)$, Lemma B.34 gives

$$b_0\big(\Xi^{-1}(\varepsilon)\big) \le 2^{rq(rq-1)/2} d^r S_r^{P-r}\big[(P-r+1)S_r - (P-r)\big]^{rq}, \qquad S_r \overset{\text{def.}}{=} PD + 1 + r(d-1).$$

Here b_0 denotes the number of connected components, with $b_0(\varnothing) = 0$. Since $1 \le S_r \le A$ and $1 \le (P-r+1)S_r - (P-r) \le (P+1)A$, this is at most $B \overset{\text{def.}}{=} 2^{Pq(Pq-1)/2} d^P A^P ((P+1)A)^{Pq}$. The subgraph characterization of pseudo-dimension and Lemma B.33, with one predicate, imply

$$\begin{aligned}
\text{P-}\dim(\mathcal{F}_i) \le 16P + 2\log_2 B = 16P + Pq(Pq-1) + 2P\log_2 d \\
+ 2P(q+1)\log_2 A + 2Pq\log_2(P+1).
\end{aligned} \tag{B.14}$$

The inequalities $d_0 \ge 1$ and $P, L \ge 1$ yield $H \ge 3$ and

$$\begin{aligned}
\log_2 d &\le 1 + L\log_2 d_0 \le H, \\
D + d - 1 &\le \big((D_0+1)L + 1\big)d_0^L \le (L+1)(D_0+2)d_0^L, \\
A &\le (P+1)(L+1)(D_0+2)d_0^L, \\
\log_2 A &\le H, \qquad \log_2(P+1) \le H.
\end{aligned}$$

Using $Pq(Pq-1) \le P^2q^2$ in (B.14), we obtain

$$\begin{aligned}\text{P-dim}(\mathcal{F}_i) &\le P^2q^2 + \left(\tfrac{16}{3} + 2 + 2 + 2\right) P(q+1)H \\ &\le P^2q_0^2U^2 + 12P(q_0U+1)H.\end{aligned}$$

This proves (B.12). Moreover, $\log_2(P+1) \le P$, $\log_2(L+1) \le L \le U$, and $P, U \ge 1$ imply

$$H \le P + U\left[1 + \log_2\big(d_0(D_0+2)\big)\right] \le \left[2 + \log_2\big(d_0(D_0+2)\big)\right] PU.$$

Together with $q_0U + 1 \le (q_0+1)U$ and $U \le LW$, this proves (B.13).

Under the additional common-chain hypothesis for each entire gate, the chain for Γ_k has length at most $q_0|V_k|$ instead of q_0W_k. The preceding construction therefore gives $q = q_0T$, with D, d, H unchanged. Repeating the last two estimates with T in place of U, using $L \le T$ and $q_0T + 1 \le (q_0+1)T$, proves the final assertion. □

Remark B.18. If $\text{comp}(G) = \varnothing$, there is a fixed linear map C such that $f(x,\theta) = A_{\text{out}}Cx$. Its joint coordinate format is at most $(0,0,2)$, and the same counting argument gives $\text{P-dim}(\mathcal{F}_i) \le 18P + 2P\log_2(P+1)$.

B.6 Proof of Statistical Guarantees for Pfaffian Neural Networks

Properties of Pfaffian Functions We use the chain and format conventions of Definition A.2. Format entries are upper bounds; in particular, a constant polynomial may always be assigned an output-degree bound of one. The local calculus lemmas below hold on open domains. The counting results are stated for chains defined on the whole Euclidean space, so no unrecorded domain complexity enters their bounds.

Lemma B.19 (Khovanskiĭ (1991); (Gabrielov & Vorobjov, 2004, Theorem 3.1))**.** *Let $f_1, \dots, f_n$ be Pfaffian on $\mathbb{R}^n$ over a common chain of length q and degree at most D, with output-degree bounds $d_1, \dots, d_n \ge 1$. Then the number of nondegenerate solutions of $f_1 = \cdots = f_n = 0$ in $\mathbb{R}^n$ is at most*

$$\mathcal{M}(n,q,D,d_1,\dots,d_n) \overset{\text{def.}}{=} 2^{q(q-1)/2} \prod_{i=1}^{n} d_i \left(\min\{n,q\}D + \sum_{i=1}^{n} d_i - n + 1\right)^q.$$

Lemma B.20. *Suppose that $g(x) = Q(x, f_1(x), \dots, f_q(x))$ is Pfaffian of format at most (q, D, d) on an open domain $U \subseteq \mathbb{R}^a$. For every $1 \le i \le a$, the derivative $\partial g/\partial x_i$ is Pfaffian over the same chain, of format at most $(q, D, \max\{0, D+d-1\})$.*

Proof. If $d = 0$, then g is constant and its derivative is represented by the zero polynomial. If $d \ge 1$, the chain rule gives

$$\frac{\partial g}{\partial x_i} = \frac{\partial Q}{\partial x_i}(x, f_1, \dots, f_q) + \sum_{j=1}^{q} \frac{\partial Q}{\partial y_j}(x, f_1, \dots, f_q)\frac{\partial f_j}{\partial x_i}.$$

The first term has polynomial degree at most $d-1$, and each summand has degree at most $d-1+D$. Since $D \ge 0$, this proves the stated bound. □

Corollary B.21. *Under the assumptions of Lemma B.20, the sequence $(f_1, \dots, f_q, g)$ is a Pfaffian chain on U of length $q+1$ and degree at most $\max\{D, D+d-1\}$.*

Proof. The original chain derivatives have degree at most D. Lemma B.20 expresses each derivative of g as a polynomial in $(x, f_1, \dots, f_q)$ of degree at most $\max\{0, D+d-1\}$. Taking the maximum of these bounds proves the assertion, since $D \ge 0$. □

Lemma B.22. *Let f and g be Pfaffian on the same open domain U, of formats at most (q_f, D_f, d_f) and (q_g, D_g, d_g) over chains $(f_1, \dots, f_{q_f})$ and $(g_1, \dots, g_{q_g})$, respectively. Then fg is Pfaffian on U of format at most $(q_f + q_g, \max\{D_f, D_g\}, d_f + d_g)$.*

Proof. Concatenating the two chains gives a chain of length q_f+q_g and degree at most $\max\{D_f, D_g\}$. Indeed, the differential polynomials for the second chain do not depend on the first chain. Over the concatenated chain, the product of the representing polynomials has degree at most $d_f + d_g$. □

Lemma B.23. *Under the assumptions of Lemma B.22, the sum $f + g$ is Pfaffian on U of format at most $(q_f + q_g, \max\{D_f, D_g\}, \max\{d_f, d_g\})$.*

Proof. Use the same concatenated chain and add the two representing polynomials. Their sum has degree at most $\max\{d_f, d_g\}$. □

Lemma B.24. *Let f and g be Pfaffian over the same chain of length q and degree at most D on U, with representing polynomial degrees at most d_f and d_g. Then fg and $f + g$ are Pfaffian over that chain of formats at most $(q, D, d_f + d_g)$ and $(q, D, \max\{d_f, d_g\})$, respectively.*

Proof. Multiply or add the representing polynomials; no enlargement of the common chain is needed. □

Example 6. Composition can require additional chain elements. The function $f_1(x) = e^x$ forms a Pfaffian chain of length one and degree one. The composition $f_2(x) = e^{e^x}$ belongs to the chain (f_1, f_2), since $f_1' = f_1$ and $f_2' = f_1 f_2$; hence f_2 has format at most $(2, 2, 1)$. It is not a polynomial in (x, e^x), because every such polynomial grows at most exponentially times a polynomial as $x \to +\infty$, whereas e^{e^x} grows faster.

Lemma B.25. *Let $\mathcal{F} = (f_1, \ldots, f_q)$ be a Pfaffian chain of degree at most D on an open domain $U_{\mathcal{F}} \subseteq \mathbb{R}^a$. For $1 \le j \le n$, let $h_j = R_j(x, f_1, \ldots, f_q)$ have format at most (q, D, d_j) over $\mathcal{F}$. Let $\mathcal{G} = (g_1, \ldots, g_{q'})$ be a Pfaffian chain of degree at most D' on an open domain $U_{\mathcal{G}} \subseteq \mathbb{R}^n$, and assume that $h = (h_1, \ldots, h_n)$ maps $U_{\mathcal{F}}$ into $U_{\mathcal{G}}$. Set $\delta \stackrel{\text{def.}}{=} \max\{1, d_1, \ldots, d_n\}$ and $s_i \stackrel{\text{def.}}{=} g_i \circ h$. Then $(f_1, \ldots, f_q, s_1, \ldots, s_{q'})$ is a Pfaffian chain on $U_{\mathcal{F}}$ of length $q+q'$ and degree at most $D+(D'+1)\delta-1$.*

Proof. All functions in the proposed chain are real-analytic. Let $P_{i,j}$ be differential polynomials for $\mathcal{G}$, of degree at most D'. For $1 \le i \le q'$ and $1 \le k \le a$, the chain rule gives

$$\frac{\partial s_i}{\partial x_k} = \sum_{j=1}^{n} P_{i,j}(h_1, \ldots, h_n, s_1, \ldots, s_i) \frac{\partial h_j}{\partial x_k}.$$

Replacing each h_j by R_j in the first factor gives a polynomial in $(x, f_1, \ldots, f_q, s_1, \ldots, s_i)$ of degree at most $D'\delta$: the substituted input coordinates have degree at most δ, and each s_j is a chain variable of degree one. By Lemma B.20, the second factor has a polynomial representation of degree at most $\max\{0, D + d_j - 1\} \le D + \delta - 1$. Thus each product has degree at most $D + (D' + 1)\delta - 1$ and depends only on the chain elements up to s_i. This bound is at least D, since $\delta \ge 1$, so it also bounds the differential polynomials of the original chain. □

Lemma B.26. *In the setting of Lemma B.25, let $\alpha(y) = Q(y, g_1(y), \ldots, g_{q'}(y))$ have format at most (q', D', d') over $\mathcal{G}$. Then $w \stackrel{\text{def.}}{=} \alpha \circ h$ is Pfaffian over the concatenated chain from that lemma, with format at most $(q + q', D + (D' + 1)\delta - 1, d'\delta)$ on $U_{\mathcal{F}}$.*

Proof. The function w is represented over the concatenated chain by

$$Q\big(R_1(x, f_1, \ldots, f_q), \ldots, R_n(x, f_1, \ldots, f_q), s_1, \ldots, s_{q'}\big).$$

A monomial $y^\alpha z^\beta$ in Q has $|\alpha| + |\beta| \le d'$. After substitution, its degree is at most $\sum_j \alpha_j d_j + |\beta| \le \delta(|\alpha| + |\beta|) \le d'\delta$. The chain bounds follow from Lemma B.25. □

Lemma B.27. *Let $h_1, \dots, h_n$ have formats at most $(q, D, d_1), \dots, (q, D, d_n)$ over a common Pfaffian chain on an open domain $U \subseteq \mathbb{R}^a$. Let α have format at most (q', D', d') on an open domain $V \subseteq \mathbb{R}^n$, and suppose that $h = (h_1, \dots, h_n)$ maps U into V. With $\delta \stackrel{\text{def.}}{=} \max\{1, d_1, \dots, d_n\}$, the composition $\alpha \circ h$ has format at most*

$$\big(q + q', D + (D' + 1)\delta - 1, d'\delta\big).$$

The witnessing chain is obtained by adjoining to the chain of h the compositions with h of the chain elements for α.

Proof. Apply Lemmas B.25 and B.26. □

Corollary B.28. *Let $(f_1(x, y), \dots, f_q(x, y))$ be a Pfaffian chain of degree at most D on an open domain $U \subseteq \mathbb{R}^{k+l}$, and let g have format at most (q, D, d) over this chain. Fix $\alpha_1, \dots, \alpha_n \in \mathbb{R}^k$ and suppose that the open set*

$$V \stackrel{\text{def.}}{=} \bigcap_{j=1}^{n} \{y \in \mathbb{R}^l : (\alpha_j, y) \in U\}$$

is nonempty. Then, on V, the chain obtained by concatenating $(f_1(\alpha_j, y), \dots, f_q(\alpha_j, y))$ for $j = 1, \dots, n$ has length nq and degree at most D. Every $g(\alpha_j, y)$ has format at most (nq, D, d) over this common chain. In particular, if $U = \mathbb{R}^{k+l}$, then $V = \mathbb{R}^l$.

Proof. Substituting constants for x in each differential polynomial and in the representing polynomial of g does not increase their degrees. Each specialized sequence remains triangular in the original chain order. Concatenating these sequences preserves their differential degree bounds and provides a common chain for all the specialized functions. □

Bounding the Format of Pfaffian Neural Networks We next quantify the Pfaffian formats of the jointly parametrized network computations.

Proposition B.29. *Suppose that a PNN is summarized by $((n, m, P), (q_k, D_k, d_{k,i}), 1 \le i \le n_k, 1 \le k \le L)$, and set $\delta_k \stackrel{\text{def.}}{=} \max_{1 \le i \le n_k} d_{k,i} \ge 1$. For each $1 \le k \le L$, the accumulated computations $z_{k,i}(x, \theta)$, $1 \le i \le n_k$, are Pfaffian on all of $\mathbb{R}^{n+P}$ over a common chain, with respective formats at most*

$$\left(\sum_{\ell=1}^{k} q_\ell,\ \sum_{\ell=1}^{k} (D_\ell + 1) \prod_{t=1}^{\ell-1} \delta_t - k,\ d_{k,i} \prod_{t=1}^{k-1} \delta_t \right).$$

Empty products equal one.

Proof. Write q'_k, D'_k, and $d'_{k,i}$ for the three displayed bounds. For $k = 1$, the claim is the assumed common-chain representation of the first layer, since $z_{1,i} = g_{1,i}$. Suppose it holds at layer $k < L$. To retain joint dependence on the parameters, compose the next layer with the augmented map

$$H_k(x, \theta) \stackrel{\text{def.}}{=} \big(z_{k,1}(x, \theta), \dots, z_{k,n_k}(x, \theta), \theta\big).$$

Its coordinates are represented over the common chain of layer k by polynomials of degree at most $\prod_{t=1}^{k} \delta_t$: this bounds $\max_i d'_{k,i}$, and also the degree-one parameter coordinates. Lemma B.27, applied to the common chain of the next layer, gives one chain for all $z_{k+1,i} = g_{k+1,i} \circ H_k$. Its length and degree are at most $q'_k + q_{k+1}$ and

$$D'_k + (D_{k+1} + 1) \prod_{t=1}^{k} \delta_t - 1 = \sum_{\ell=1}^{k+1} (D_\ell + 1) \prod_{t=1}^{\ell-1} \delta_t - (k + 1),$$

respectively. The representing polynomial of $z_{k+1,i}$ has degree at most $d_{k+1,i} \prod_{t=1}^{k} \delta_t$. These are the required bounds at layer $k + 1$, completing the induction. □

Corollary B.30. *Under the assumptions and notation of Proposition B.29, all output coordinates share a Pfaffian chain of length at most* $\sum_{\ell=1}^{L} q_\ell$ *and degree at most*

$$D_1 + \sum_{\ell=2}^{L} (D_\ell + 1) \prod_{t=1}^{\ell-1} \delta_t - L + 1.$$

The representing polynomial of output i *has degree at most* $d_{L,i} \prod_{t=1}^{L-1} \delta_t$.

Proof. Set $k = L$ in Proposition B.29. □

Remark B.31 (Simplification). If the layer inputs $u_1, \ldots, u_L$ are regarded as independent variables, the layer chains can be concatenated into a chain of length $\sum_{k=1}^{L} q_k$ and degree at most $\max_{1 \le k \le L} D_k$ in $(u_1, \ldots, u_L, \theta)$. Substituting the accumulated computations for these independent layer inputs changes the differential degree; Proposition B.29 accounts for this change.

Corollary B.32 (Simplified Bound)**.** *Suppose that a PNN is summarized as in Proposition B.29. Let* $q \stackrel{\text{def.}}{=} \sum_{k=1}^{L} q_k$, $D \stackrel{\text{def.}}{=} \max_{1 \le k \le L} D_k$, *and* $d \stackrel{\text{def.}}{=} \max_{1 \le k \le L,\, 1 \le i \le n_k} d_{k,i} \ge 1$. *Its output coordinates have formats at most*

$$\left(q, (D+1) \sum_{j=0}^{L-1} d^j - L, d^L \right).$$

In particular, $(q, (D+1)Ld^{L-1}, d^L)$ *is a valid format bound. When* $d = 1$, *the sharper chain-degree bound is* LD.

Proof. The length and output-degree bounds follow from Proposition B.29. Its chain-degree bound satisfies

$$\sum_{\ell=1}^{L} (D_\ell + 1) \prod_{t=1}^{\ell-1} \delta_t - L \le (D+1) \sum_{\ell=1}^{L} d^{\ell-1} - L \le (D+1) L d^{L-1}.$$

For $d = 1$, the middle expression is LD. □

Bounding the Pseudo-Dimension of Pfaffian Neural Networks We combine the preceding format bounds with the theorem of Marek Karpinski and Angus Macintyre Karpinski & Macintyre (1997) and Khovanskii's bound for regular Pfaffian level sets Khovanskiĭ (1991). The parameter space remains all of $\mathbb{R}^P$ throughout.

Lemma B.33 ((Karpinski & Macintyre, 1997, Theorem 2))**.** *Let* $f_1, \ldots, f_s : \mathbb{R}^{a+P} \to \mathbb{R}$ *be* C^∞ *functions, with* $P, s \ge 1$, *and let* $\varphi(x, \theta)$ *be a Boolean combination of the predicates* $f_j(x, \theta) > 0$ *and* $f_j(x, \theta) = 0$. *Write* $\mathcal{C}_\varphi \stackrel{\text{def.}}{=} \{\{x \in \mathbb{R}^a : \varphi(x, \theta)\} : \theta \in \mathbb{R}^P\}$. *Suppose that* $B \ge 1$ *uniformly bounds the number of connected components of* $\Xi^{-1}(\varepsilon)$ *whenever* $1 \le r \le P$, $\Xi = (\tau_1, \ldots, \tau_r) : \mathbb{R}^P \to \mathbb{R}^r$, *each* τ_j *is obtained by fixing the first* a *variables in one of the* f_i, *and* ε *is a regular value of* Ξ. *Then*

$$\operatorname{VCdim}(\mathcal{C}_\varphi) \le 2 \log_2 B + (16 + 2 \log_2 s) P.$$

Here a value is regular if $\operatorname{rank} D\Xi(\theta) = r$ *at every* $\theta \in \Xi^{-1}(\varepsilon)$; *an empty fiber is permitted.*

Lemma B.34 ((Khovanskiĭ, 1991, p. 91, Example 3); (Karpinski & Macintyre, 1997, Section 1.4(d)))**.** *Let* $1 \le r \le b$, *and let* $Q_1, \ldots, Q_r : \mathbb{R}^b \to \mathbb{R}$ *be Pfaffian functions over a common chain defined on all of* $\mathbb{R}^b$, *of length* ℓ *and degree at most* D. *Suppose that their polynomial representations have degrees at most* $d \ge 1$ *and that* 0 *is a regular value of* $(Q_1, \ldots, Q_r)$. *Then their common zero set has at most*

$$2^{\ell(\ell-1)/2} d^r S^{b-r} \big[(b - r + 1)S - (b - r) \big]^\ell, \qquad S \stackrel{\text{def.}}{=} bD + 1 + r(d-1),$$

connected components.

Proof of Theorem B.14. Fix an output coordinate i. By Proposition B.29, f_i has format at most (q, D, d) over a chain defined on all of $\mathbb{R}^{n+P}$. Introduce $H(x, t, \theta) \overset{\text{def.}}{=} f_i(x, \theta) - t$ and let

$$\mathcal{C}_i \overset{\text{def.}}{=} \big\{\{(x, t) \in \mathbb{R}^{n+1} : f_i(x, \theta) > t\} : \theta \in \mathbb{R}^P\big\}.$$

By the definition of pseudo-dimension, $\text{P-}\dim(\mathcal{F}_i) = \text{VCdim}(\mathcal{C}_i)$. The threshold t is part of the data variable, so the number of trainable parameters remains P. Extend the chain for f_i to $\mathbb{R}^{n+P+1}$ by making every chain element independent of t. Its derivatives in t are zero, and subtracting t from the output polynomial gives degree at most $\max\{d, 1\} = d$. Thus H is Pfaffian of format at most (q, D, d); in particular, $\partial H/\partial t = -1$.

To verify Lemma B.33, fix $1 \le r \le P$, data points $(a_1, c_1), \dots, (a_r, c_r) \in \mathbb{R}^{n+1}$, and the functions $\tau_j(\theta) \overset{\text{def.}}{=} H(a_j, c_j, \theta)$. These points may repeat. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_r)$ be a regular value of $\Xi = (\tau_1, \dots, \tau_r)$. Substituting each a_j into the chain for f_i and concatenating the resulting r chains gives, by Corollary B.28, a common chain on all of $\mathbb{R}^P$ of length at most rq and degree at most D. Each $\chi_j \overset{\text{def.}}{=} \tau_j - \varepsilon_j$ has polynomial degree at most d over this chain. Since 0 is a regular value of $(\chi_1, \dots, \chi_r)$, Lemma B.34 yields

$$b_0\big(\Xi^{-1}(\varepsilon)\big) \le 2^{rq(rq-1)/2} d^r S_r^{P-r} \big[(P - r + 1)S_r - (P - r)\big]^{rq}, \qquad S_r \overset{\text{def.}}{=} PD + 1 + r(d - 1),$$

where b_0 denotes the number of connected components. For an empty fiber the left-hand side is zero. Because $d \ge 1$, $D \ge 0$, and $r \le P$, we have $1 \le S_r \le A$ and $1 \le (P{-}r{+}1)S_r{-}(P{-}r) \le (P{+}1)A$. All bases are therefore at least 1, and the displayed bound is at most

$$B \overset{\text{def.}}{=} 2^{Pq(Pq-1)/2} d^P A^P \big[(P + 1)A\big]^{Pq}.$$

This bound is independent of the data points, r, and the regular value ε. Lemma B.33, applied to the single predicate $H > 0$, gives $\text{P-}\dim(\mathcal{F}_i) \le 16P + 2\log_2 B$. Finally,

$$2\log_2 B = Pq(Pq - 1) + 2P\log_2 d + 2P(q + 1)\log_2 A + 2Pq\log_2(P + 1),$$

which is the asserted bound. □

Proof of Theorem B.15. Lemma B.17 supplies a common global chain for all coordinates of f, of format at most

$$q = q_0 U, \qquad D = (D_0 + 1)\sum_{j=0}^{L-1} d_0^j - L, \qquad d = d_0^L + 1.$$

Every point (x, y, t) in a set shattered by the strict subgraphs of $\mathcal{F}_\ell$ satisfies $0 \le t < 1$: for $t < 0$ the subgraph predicate is always true, and for $t \ge 1$ it is always false. For $0 \le t < 1$,

$$\ell(f_\theta(x), y) > t \quad \Longleftrightarrow \quad \Psi(x, y, t, \theta) \overset{\text{def.}}{=} \sum_{j=1}^{d_{out}} \big(f_j(x, \theta) - y_j\big)^2 - t > 0.$$

Extend the common chain of f independently of (y, t). Its length and differential degree remain at most q and D, and Ψ has representing polynomial degree at most $2d$. Thus Ψ is globally Pfaffian of format at most $(q, D, 2d)$, with (x, y, t) as data and exactly P trainable parameters. The counting argument giving (B.14), applied to the predicate $\Psi > 0$, yields

$$\begin{aligned}\text{P-}\dim(\mathcal{F}_\ell) \le\ & 16P + Pq(Pq - 1) + 2P\log_2(2d) \\ & + 2P(q + 1)\log_2 A_{\text{sq}} + 2Pq\log_2(P + 1),\end{aligned} \qquad A_{\text{sq}} \overset{\text{def.}}{=} P(D + 2d - 1) + 1.$$

Since $d_0 \ge 1$ and $L \ge 1$,

$$\begin{aligned} D + 2d - 1 &= (D_0 + 1)\sum_{j=0}^{L-1} d_0^j - L + 2d_0^L + 1 \\ &\le \big((D_0 + 1)L + 2\big)d_0^L \le (L + 1)(D_0 + 2)d_0^L. \end{aligned}$$

Consequently,

$$\log_2 A_{\text{sq}} \le H, \qquad \log_2(2d) \le 2 + L\log_2 d_0 \le H, \qquad \log_2(P + 1) \le H, \qquad H \ge 3.$$

Therefore

$$\text{P-dim}(\mathcal{F}_\ell) \le P^2q^2 + \left(\tfrac{16}{3} + 2 + 2 + 2\right) P(q+1)H \le K_{\text{sq}}.$$

The estimate $K_{\text{sq}} \le C_{\text{gate}} P^2U^2$ is the final algebraic estimate in the proof of Lemma B.17; $U \le LW$ gives the remaining inequality in (**??**).

The map h is jointly Borel measurable and takes values in $[0,1]$. Since $K_{\text{sq}} \ge 36$, we have $\max\{1, \text{P-dim}(\mathcal{F}_\ell)\} \le K_{\text{sq}}$. Lemma B.8 and the monotonicity of $v \mapsto v \ln^2(ev/\varepsilon)$ on $[1,\infty)$ give (B.10) and (**??**).

Moreover, the global Pfaffian map f is jointly definable in the Pfaffian closure of the real field, as recalled in Subsection A.1. Since ℓ is semi-algebraic, $\mathcal{Z}$, Θ, and h are definable in the same structure. Lemma B.1 supplies $\widehat{\theta}_{N,\eta}$ for every $N \in \mathbb{N}_+$ and $\eta > 0$. For any empirical η-minimizer and every $\theta \in \mathbb{R}^P$,

$$\begin{aligned} R\big(\widehat{\theta}_{N,\eta}(s)\big) &\le \widehat{R}_s\big(\widehat{\theta}_{N,\eta}(s)\big) + \Delta_N(s) \\ &\le \widehat{R}_s(\theta) + \eta + \Delta_N(s) \\ &\le R(\theta) + \eta + 2\Delta_N(s). \end{aligned}$$

Taking the infimum over θ proves the asserted excess-risk bound on $\{\Delta_N \le \varepsilon\}$.

Under the additional common-chain hypothesis for each entire gate, the architecture lemma gives $q = q_0 T$, with D, d, H unchanged. The preceding argument applies with T in place of U. □

C Proofs regarding Definable Neural Network Building Blocks and Architectures

Proof of Proposition 4.1. Write $P = (P_1, \dots, P_m)$, where each $P_j : \mathbb{R}^n \to \mathbb{R}$ is a polynomial. Then

$$\text{graph}(P) = \Big\{(x,y) \in \mathbb{R}^{n+m} : y_j - P_j(x) = 0,\ j = 1, \dots, m\Big\}.$$

Hence $\text{graph}(P)$ is algebraic, and therefore semi-algebraic. Thus P is semi-algebraic, so it is definable in $(\mathbb{R}, +, \cdot, <)$, and consequently in every o-minimal expansion of the real field.

Finally, polynomials are Pfaffian functions relative to the empty Pfaffian chain. Therefore each coordinate P_j is Pfaffian, and hence P is Pfaffian. □

Corollary C.1 (Definability of Linear Layers in MLPs)**.** *Fix $n, m \in \mathbb{N}_+$. The affine layer $L : \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^m \to \mathbb{R}^m, L(x,A,b) \overset{\text{def.}}{=} Ax + b$, is a polynomial map jointly in (x,A,b). In particular, it is semi-algebraic and definable in the real field $(\mathbb{R}, +, \cdot, <)$, and hence in every o-minimal expansion of the real field.*

Proof of Corollary C.1. For each output coordinate $i \in \{1, \dots, m\}$, we have

$$L_i(x,A,b) = \sum_{j=1}^{n} A_{ij}x_j + b_i.$$

This is a polynomial function of the coordinates of (x,A,b). Hence L is a polynomial map. The conclusion follows from Proposition 4.1. □

Corollary C.2 (Definability of Convolutional Layers in CNNs)**.** *Fix $n, m, p \in \mathbb{N}_+$. Let $T : \mathbb{R}^p \to \mathbb{R}^{m\times n}$ be the linear map that sends a vector of convolutional kernel parameters to the corresponding Toeplitz-type (or doubly block Toeplitz) matrix. The layer $\text{Conv} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^m, \text{Conv}(x,\theta,b) \overset{\text{def.}}{=} T(\theta)x + b$, is a polynomial map jointly in (x,θ,b). It is semi-algebraic and definable in the real field $(\mathbb{R}, +, \cdot, <)$, and in every o-minimal expansion of the real field.*

Proof of Corollary C.2. For each output coordinate $i \in \{1, \dots, m\}$, we have

$$\mathrm{Conv}(x, \theta, b)_i = \sum_{j=1}^{n} T(\theta)_{ij} x_j + b_i.$$

Since $T : \mathbb{R}^p \to \mathbb{R}^{m \times n}$ is linear, each entry $T(\theta)_{ij}$ is a linear function of θ. Hence each term $T(\theta)_{ij} x_j$ is polynomial in the coordinates of (x, θ, b). Therefore each output coordinate is polynomial in (x, θ, b), and so Conv is a polynomial map.

The conclusion follows from Proposition 4.1. □

Proof of Corollary 4.2. Write

$$F = (F_1, \dots, F_n) \qquad \text{and} \qquad G = (G_1, \dots, G_n),$$

where each $F_i, G_i : \mathbb{R}^n \to \mathbb{R}$ is a polynomial function.

The residual block

$$R(x) \stackrel{\text{def.}}{=} x + F(x)$$

has coordinates

$$R_i(x) = x_i + F_i(x), \qquad i = 1, \dots, n.$$

Since sums of polynomial functions are polynomial, each coordinate R_i is polynomial. Hence R is a polynomial map.

Similarly, the gated residual block

$$H(x) \stackrel{\text{def.}}{=} x + G(x) \odot F(x)$$

has coordinates

$$H_i(x) = x_i + G_i(x) F_i(x), \qquad i = 1, \dots, n.$$

Since products and sums of polynomial functions are polynomial, each coordinate H_i is polynomial. Hence H is also a polynomial map.

Therefore both residual and gated residual polynomial blocks are polynomial maps. By Proposition 4.1, they are semi-algebraic and definable in the real field. □

The following proposition is an extended version of Proposition 4.3.

Proposition C.3 (Definability of Modern Activation Functions)**.** *The following activation functions are definable.*

(i) The logistic sigmoid activation

$$\sigma(x) \stackrel{\text{def.}}{=} \frac{1}{1 + e^{-x}}$$

is Pfaffian on $\mathbb{R}$ of format at most $(1, 2, 1)$. In particular, it is definable in the Pfaffian closure of the real field, and also in $\mathbb{R}_{\exp}$.

(ii) The hyperbolic tangent activation

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

is Pfaffian on $\mathbb{R}$ of format at most $(1, 2, 1)$. In particular, it is definable in the Pfaffian closure of the real field, and also in $\mathbb{R}_{\exp}$.

(iii) The softplus activation

$$\mathrm{Softplus}(x) \stackrel{\text{def.}}{=} \log(1 + e^x)$$

is Pfaffian on $\mathbb{R}$ of format at most $(2, 2, 1)$. In particular, it is definable in the Pfaffian closure of the real field, and also in $\mathbb{R}_{\exp}$.

(iv) Every piecewise-polynomial map $f : \mathbb{R}^n \to \mathbb{R}^m$ of the form $f = \sum_{k=1}^{K} p_k I_{A_k}$, where $p_1, \ldots, p_K : \mathbb{R}^n \to \mathbb{R}^m$ are polynomial maps and $A_1, \ldots, A_K \subseteq \mathbb{R}^n$ form a finite semi-algebraic partition of $\mathbb{R}^n$, is semi-algebraic. In particular, it is definable in the real field $(\mathbb{R}, +, \cdot, <)$.

In particular, the following standard activations are semi-algebraic and hence definable in the real field:

$$\mathrm{ReLU}(x) = \max\{x, 0\}, \qquad \mathrm{LeakyReLU}_\alpha(x) = \max\{x, \alpha x\},$$

for fixed $\alpha \in \mathbb{R}$, the parametric ReLU $\mathrm{PReLU}(x, \alpha) = \max\{x, \alpha x\}$ jointly in (x, α), ReLU^p for $p \in \mathbb{N}_+$, hard-threshold activations, hard-tanh, hard-sigmoid, and finite-knot spline activations, such as those used in KANs.

(v) The exponential linear unit

$$\mathrm{ELU}_\alpha(x) \overset{\text{def.}}{=} \begin{cases} x, & x \geq 0, \\ \alpha(e^x - 1), & x < 0, \end{cases}$$

for fixed $\alpha \in \mathbb{R}$, is definable in $\mathbb{R}_{\exp}$. Moreover, the parametrized map

$$(x, \alpha) \longmapsto \mathrm{ELU}_\alpha(x)$$

is definable in $\mathbb{R}_{\exp}$ jointly in (x, α). Consequently, for fixed constants $\lambda, \alpha \in \mathbb{R}$, the SELU activation

$$\mathrm{SELU}_{\lambda,\alpha}(x) \overset{\text{def.}}{=} \lambda\ \mathrm{ELU}_\alpha(x)$$

is also definable in $\mathbb{R}_{\exp}$.

(vi) The Gaussian Error Linear Unit

$$\mathrm{GELU}(x) \overset{\text{def.}}{=} x\Phi(x) = \frac{x}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$

is Pfaffian on $\mathbb{R}$ of format at most $(2, 2, 2)$. In particular, GELU is definable in the Pfaffian closure of the real field.

The commonly used tanh approximation, implemented for example by PyTorch when `approximate='tanh'`, is

$$\mathrm{GELU}_{\tanh}(x) \overset{\text{def.}}{=} \frac{x}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right).$$

This approximation is also Pfaffian on $\mathbb{R}$, and is definable in $\mathbb{R}_{\exp}$, hence also in $\mathbb{R}_{\mathrm{an,exp}}$.

(vii) The Swish activation $\mathrm{Swish}(x, \beta) \overset{\text{def.}}{=} \frac{x}{1+e^{-\beta x}}$ on $(x, \beta) \in \mathbb{R} \times (0, \infty)$ is Pfaffian of format at most $(2, 4, 2)$. In particular, it is definable in the Pfaffian closure of the real field, and also in $\mathbb{R}_{\exp}$.

(viii) The SwiGLU activation

$$\mathrm{SwiGLU}(x_1, x_2, \beta) \overset{\text{def.}}{=} x_1\ \mathrm{Swish}(x_2, \beta) = x_1 \frac{x_2}{1 + e^{-\beta x_2}}, \qquad (x_1, x_2, \beta) \in \mathbb{R}^2 \times \mathbb{R}_{>0},$$

is Pfaffian of format at most $(2, 4, 3)$. In particular, it is definable in the Pfaffian closure of the real field, and also in $\mathbb{R}_{\exp}$.

(ix) The softsign activation

$$\mathrm{Softsign}(x) \overset{\text{def.}}{=} \frac{x}{1 + |x|}$$

is semi-algebraic. In particular, it is definable in the real field $(\mathbb{R}, +, \cdot, <)$, and hence in $\mathbb{R}_{\mathrm{an}}$ and $\mathbb{R}_{\mathrm{an,exp}}$.

(x) The Mish activation

$$\mathrm{Mish}(x) \overset{\text{def.}}{=} x\,\tanh\big(\mathrm{Softplus}(x)\big), \qquad \mathrm{Softplus}(x) \overset{\text{def.}}{=} \log(1+e^x),$$

is definable in $\mathbb{R}_{\exp}$, and hence in $\mathbb{R}_{\mathrm{an,exp}}$.

Proof of Proposition C.3.(i). Let

$$f_1(x) \overset{\text{def.}}{=} \sigma(x) = \frac{1}{1+e^{-x}}, \qquad x \in \mathbb{R}.$$

Then f_1 is real-analytic on $\mathbb{R}$. Moreover,

$$\frac{df_1}{dx}(x) = \frac{e^{-x}}{(1+e^{-x})^2} = f_1(x)\big(1-f_1(x)\big).$$

Equivalently,

$$\frac{df_1}{dx} = P(f_1), \qquad P(Y_1) \overset{\text{def.}}{=} Y_1 - Y_1^2.$$

The polynomial $P \in \mathbb{R}[Y_1] \subseteq \mathbb{R}[X, Y_1]$ has degree 2. Hence (f_1) is a Pfaffian chain on $\mathbb{R}$ of length 1 and degree 2.

Finally,

$$\sigma(x) = f_1(x)$$

is obtained from this chain by the polynomial $Q(Y_1) = Y_1$, of degree 1. Therefore σ is Pfaffian on $\mathbb{R}$, of format at most $(1, 2, 1)$. In particular, σ is definable in the Pfaffian closure of the real field. Since it is also obtained from field operations and the global exponential map, it is definable in $\mathbb{R}_{\exp}$, and hence in $\mathbb{R}_{\mathrm{an,exp}}$. □

Proof of Proposition C.3.(ii). Let

$$f_1(x) \overset{\text{def.}}{=} \tanh(x), \qquad x \in \mathbb{R}.$$

Then f_1 is real-analytic on $\mathbb{R}$. Moreover,

$$\frac{df_1}{dx}(x) = 1 - \tanh^2(x) = 1 - f_1(x)^2.$$

Equivalently,

$$\frac{df_1}{dx} = P(f_1), \qquad P(Y_1) \overset{\text{def.}}{=} 1 - Y_1^2.$$

The polynomial $P \in \mathbb{R}[Y_1] \subseteq \mathbb{R}[X, Y_1]$ has degree 2. Hence (f_1) is a Pfaffian chain on $\mathbb{R}$ of length 1 and degree 2.

Finally,

$$\tanh(x) = f_1(x)$$

is obtained from this chain by the polynomial $Q(Y_1) = Y_1$, of degree 1. Therefore $\tanh$ is Pfaffian on $\mathbb{R}$, of format at most $(1, 2, 1)$. In particular, $\tanh$ is definable in the Pfaffian closure of the real field. Since

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1},$$

and the denominator $e^{2x}+1$ is strictly positive, $\tanh$ is also definable in $\mathbb{R}_{\exp}$. □

Proof of Proposition C.3.(iii). Let

$$f_1(x) \overset{\text{def.}}{=} \frac{1}{1+e^{-x}}, \qquad f_2(x) \overset{\text{def.}}{=} \mathrm{Softplus}(x) = \log(1+e^x).$$

Both f_1 and f_2 are real-analytic on $\mathbb{R}$. Moreover,

$$\frac{df_1}{dx}(x) = f_1(x)\big(1-f_1(x)\big),$$

and

$$\frac{df_2}{dx}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} = f_1(x).$$

Equivalently,

$$\frac{df_1}{dx} = P_1(f_1), \qquad P_1(Y_1) \overset{\text{def.}}{=} Y_1 - Y_1^2,$$

and

$$\frac{df_2}{dx} = P_2(f_1, f_2), \qquad P_2(Y_1, Y_2) \overset{\text{def.}}{=} Y_1.$$

The polynomials $P_1 \in \mathbb{R}[Y_1]$ and $P_2 \in \mathbb{R}[Y_1, Y_2]$ have degrees at most 2. Hence (f_1, f_2) is a Pfaffian chain on $\mathbb{R}$ of length 2 and degree 2.

Finally,

$$\mathrm{Softplus}(x) = f_2(x)$$

is obtained from this chain by the polynomial $Q(Y_1, Y_2) = Y_2$, of degree 1. Therefore $\mathrm{Softplus}$ is Pfaffian on $\mathbb{R}$, of format at most $(2, 2, 1)$. In particular, it is definable in the Pfaffian closure of the real field.

It is also definable in $\mathbb{R}_{\exp}$: the map $x \mapsto 1 + e^x$ is $\mathbb{R}_{\exp}$-definable and strictly positive, and the logarithm is definable in $\mathbb{R}_{\exp}$ as the inverse of the global exponential function. □

Proof of Proposition C.3.(iv). Since each A_k is semi-algebraic and each p_k is polynomial, the graph of the restriction $p_k|_{A_k}$ is

$$\{(x, y) \in \mathbb{R}^{n+m} : x \in A_k,\ y = p_k(x)\},$$

which is semi-algebraic. Since the sets $A_1, \ldots, A_K$ form a finite partition of $\mathbb{R}^n$, the graph of f is

$$\mathrm{Graph}(f) = \bigcup_{k=1}^{K} \{(x, y) \in \mathbb{R}^{n+m} : x \in A_k,\ y = p_k(x)\}.$$

This is a finite union of semi-algebraic sets, hence is semi-algebraic. Therefore f is semi-algebraic, and thus definable in the real field $(\mathbb{R}, +, \cdot, <)$.

For the activations listed above, this framework applies directly. Indeed,

$$\mathrm{ReLU}(x) = \begin{cases} 0, & x < 0, \\ x, & x \geq 0, \end{cases}$$

and, for fixed $\alpha \in \mathbb{R}$,

$$\mathrm{LeakyReLU}_\alpha(x) = \begin{cases} \alpha x, & x < 0, \\ x, & x \geq 0. \end{cases}$$

Both are therefore piecewise-polynomial with respect to the finite semi-algebraic partition

$$\mathbb{R} = (-\infty, 0) \cup [0, \infty).$$

Similarly, this directly applies to the parametric ReLU (PReLU), where α is learned rather than fixed: the map $(x, \alpha) \mapsto \mathrm{PReLU}(x, \alpha)$ is semi-algebraic. □

Proof of Proposition C.3.(v). Consider the semi-algebraic partition

$$\mathbb{R} = (-\infty, 0) \cup [0, \infty).$$

On $[0, \infty)$, the map is $x \mapsto x$, which is polynomial. On $(-\infty, 0)$, the map is

$$x \mapsto \alpha(e^x - 1),$$

which is definable in $\mathbb{R}_{\exp}$. Hence, for fixed $\alpha \in \mathbb{R}$, the graph of ELU_α is

$$\begin{aligned} \mathrm{Graph}(\mathrm{ELU}_\alpha) = &\{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y = x\} \\ &\cup \{(x, y) \in \mathbb{R}^2 : x < 0,\ y = \alpha(e^x - 1)\}. \end{aligned}$$

This is a finite union of $\mathbb{R}_{\exp}$-definable sets, and is therefore $\mathbb{R}_{\exp}$-definable.

The joint statement follows similarly. Define

$$E : \mathbb{R}^2 \to \mathbb{R}, \qquad E(x, \alpha) \stackrel{\text{def.}}{=} \mathrm{ELU}_\alpha(x).$$

Then

$$\begin{aligned}\mathrm{Graph}(E) = &\{(x, \alpha, y) \in \mathbb{R}^3 : x \geq 0,\ y = x\} \\ &\cup \{(x, \alpha, y) \in \mathbb{R}^3 : x < 0,\ y = \alpha(e^x - 1)\}.\end{aligned}$$

Again this is a finite union of $\mathbb{R}_{\exp}$-definable sets, so E is definable in $\mathbb{R}_{\exp}$.

Finally, SELU is obtained from ELU by multiplication by a fixed scalar λ. Since definable maps are closed under scalar multiplication, $\mathrm{SELU}_{\lambda,\alpha}$ is definable in $\mathbb{R}_{\exp}$. □

Proof of Proposition C.3.(vi). We first treat the exact GELU. Define, on $U = \mathbb{R}$,

$$f_1(x) \stackrel{\text{def.}}{=} e^{-x^2/2}, \qquad f_2(x) \stackrel{\text{def.}}{=} \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right).$$

Both functions are real-analytic on $\mathbb{R}$. Moreover,

$$\frac{df_1}{dx}(x) = -x f_1(x),$$

and, using

$$\frac{d}{dt}\,\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} e^{-t^2},$$

we obtain

$$\frac{df_2}{dx}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2}} = \sqrt{\frac{2}{\pi}}\, f_1(x).$$

Thus

$$\frac{df_1}{dx} = P_1(x, f_1), \qquad P_1(X, Y_1) \stackrel{\text{def.}}{=} -XY_1,$$

and

$$\frac{df_2}{dx} = P_2(x, f_1, f_2), \qquad P_2(X, Y_1, Y_2) \stackrel{\text{def.}}{=} \sqrt{\frac{2}{\pi}}\, Y_1.$$

The polynomials P_1 and P_2 have degrees at most 2. Hence (f_1, f_2) is a Pfaffian chain on $\mathbb{R}$ of length 2 and degree 2. Finally,

$$\mathrm{GELU}(x) = \frac{x}{2}\big(1 + f_2(x)\big)$$

is obtained from this chain by the polynomial

$$Q(X, Y_1, Y_2) \stackrel{\text{def.}}{=} \frac{X}{2}(1 + Y_2),$$

which has degree 2. Therefore the exact GELU is Pfaffian on $\mathbb{R}$ of format at most $(2, 2, 2)$. In particular, it is definable in the Pfaffian closure of the real field.

We next treat the tanh approximation. Let

$$c \stackrel{\text{def.}}{=} \sqrt{\frac{2}{\pi}}, \qquad a \stackrel{\text{def.}}{=} 0.044715, \qquad h(x) \stackrel{\text{def.}}{=} c(x + ax^3),$$

and define

$$g_1(x) \stackrel{\text{def.}}{=} \tanh(h(x)).$$

Then g_1 is real-analytic on $\mathbb{R}$, and

$$\frac{dg_1}{dx}(x) = h'(x)\big(1 - \tanh^2(h(x))\big) = c(1 + 3ax^2)\big(1 - g_1(x)^2\big).$$

Hence

$$\frac{dg_1}{dx} = R(x, g_1), \qquad R(X, Y_1) \stackrel{\text{def.}}{=} c(1 + 3aX^2)(1 - Y_1^2),$$

where $R \in \mathbb{R}[X, Y_1]$. Thus (g_1) is a Pfaffian chain on $\mathbb{R}$. Moreover,

$$\mathrm{GELU}_{\tanh}(x) = \frac{x}{2}\big(1 + g_1(x)\big),$$

so $\mathrm{GELU}_{\tanh}$ is Pfaffian.

Finally, since $\tanh$ is definable in $\mathbb{R}_{\exp}$, and $\mathrm{GELU}_{\tanh}$ is obtained from $\tanh$ by polynomial operations and composition with a polynomial, it is definable in $\mathbb{R}_{\exp}$. Hence it is also definable in $\mathbb{R}_{\mathrm{an,exp}}$. □

Proof of Proposition C.3.(vii). Consider Swish on $U = \mathbb{R} \times \mathbb{R}_{>0}$. Define

$$f_1(x, \beta) = e^{-\beta x},$$
$$f_2(x, \beta) = \frac{1}{1 + f_1(x, \beta)}.$$

Then $\frac{\partial f_1}{\partial x} = -\beta f_1$, $\frac{\partial f_1}{\partial \beta} = -x f_1$,

$$\frac{\partial f_2}{\partial x} = -\frac{\frac{\partial f_1}{\partial x}}{(1 + f_1)^2} = \beta f_1 f_2^2,$$

and $\frac{\partial f_2}{\partial \beta} = x f_1 f_2^2$. Thus (f_1, f_2) is a Pfaffian chain of order 2 and degree 4. Since $\mathrm{Swish}(x, \beta) = x f_2(x, \beta)$, Swish is Pfaffian of format at most $(2, 4, 2)$. Pfaffian functions are definable in the Pfaffian closure of the real field. Moreover, Swish is obtained from polynomials, $\exp$, addition, multiplication, and division by the strictly positive term $1 + e^{-\beta x}$; hence it is also definable in $\mathbb{R}_{\exp}$. □

Proof of Proposition C.3.(viii). Consider SwiGLU on $U = \mathbb{R}^2 \times \mathbb{R}_{>0}$. Define

$$f_1(x_2, \beta) = e^{-\beta x_2},$$
$$f_2(x_2, \beta) = \frac{1}{1 + f_1(x_2, \beta)}.$$

The same computation as for Swish gives

$$\frac{\partial f_1}{\partial x_2} = -\beta f_1, \qquad \frac{\partial f_1}{\partial \beta} = -x_2 f_1,$$
$$\frac{\partial f_2}{\partial x_2} = \beta f_1 f_2^2, \qquad \frac{\partial f_2}{\partial \beta} = x_2 f_1 f_2^2,$$

while the derivatives with respect to x_1 vanish. Thus (f_1, f_2) is again a Pfaffian chain of order 2 and degree 4. Since $\mathrm{SwiGLU}(x_1, x_2, \beta) = x_1 x_2 f_2(x_2, \beta)$, SwiGLU is Pfaffian of format at most $(2, 4, 3)$. Pfaffian functions are definable in the Pfaffian closure of the real field. Moreover, SwiGLU is obtained from polynomials, $\exp$, addition, multiplication, and division by the strictly positive term $1 + e^{-\beta x_2}$; hence it is also definable in $\mathbb{R}_{\exp}$. □

Proof of Proposition C.3.(ix). Consider the semi-algebraic partition

$$\mathbb{R} = (-\infty, 0) \cup [0, \infty).$$

On $[0, \infty)$, we have $|x| = x$, and hence

$$\mathrm{Softsign}(x) = \frac{x}{1 + x}.$$

On $(-\infty, 0)$, we have $|x| = -x$, and hence

$$\mathrm{Softsign}(x) = \frac{x}{1 - x}.$$

The denominators $1 + x$ on $[0, \infty)$ and $1 - x$ on $(-\infty, 0)$ are strictly positive. Therefore the graph of Softsign is

$$\mathrm{Graph}(\mathrm{Softsign}) = \{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y(1 + x) = x\}$$
$$\cup \{(x, y) \in \mathbb{R}^2 : x < 0,\ y(1 - x) = x\}.$$

This is a finite union of sets defined by polynomial equalities and inequalities, hence is semi-algebraic. Consequently, Softsign is definable in the real field $(\mathbb{R}, +, \cdot, <)$. □

Proof of Proposition C.3.(x). The composition

$$x \mapsto \tanh(\mathrm{Softplus}(x))$$

is definable in $\mathbb{R}_{\exp}$. Multiplication by x preserves definability, so

$$\mathrm{Mish}(x) = x\ \tanh(\mathrm{Softplus}(x))$$

is definable in $\mathbb{R}_{\exp}$. Since $\mathbb{R}_{\mathrm{an},\exp}$ expands $\mathbb{R}_{\exp}$, Mish is also definable in $\mathbb{R}_{\mathrm{an},\exp}$. □

Proof of Proposition 4.4. For the maxout map, write $\ell_k(x) \overset{\text{def.}}{=} a_k^\top x + b_k$. Its graph is

$$\mathrm{Graph}(\sigma) = \bigcup_{k=1}^{K} \Big\{ (x, y, (a_j, b_j)_{j=1}^K) : y = \ell_k(x)\ \wedge\ \bigwedge_{j=1}^{K} \ell_k(x) \geq \ell_j(x) \Big\}.$$

Each set in the union is cut out by finitely many polynomial equalities and inequalities in $(x, y, (a_j, b_j)_{j=1}^K)$. Hence the graph is semi-algebraic.

For winner-take-all gating, let $R_1, \ldots, R_K$ be the given finite polyhedral partition of $\mathbb{R}^d$, and suppose that the affine map $x \mapsto A_k x + b_k$ is applied on R_k. Since each R_k is polyhedral, it is semi-algebraic. The graph of the corresponding gating map is

$$\mathrm{Graph}(\sigma) = \bigcup_{k=1}^{K} \Big\{ (x, y) : x \in R_k\ \wedge\ y = A_k x + b_k \Big\}.$$

Each set in the union is defined by finitely many polynomial equalities and inequalities, and hence is semi-algebraic. Since finite unions of semi-algebraic sets are semi-algebraic, the graph is semi-algebraic.

Therefore both maps are semi-algebraic, hence definable in the real field $(\mathbb{R}, +, \cdot, <)$ and in every o-minimal expansion of the real field. □

Proposition C.4 (Definability of Softmax)**.** *Fix $K \in \mathbb{N}_+$. The softmax map*

$$\mathrm{softmax} : \mathbb{R}^K \to \mathbb{R}^K, \qquad \mathrm{softmax}(z)_k \overset{\text{def.}}{=} \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}},$$

is definable in $\mathbb{R}_{\exp}$. Consequently, any finite softmax-weighted aggregation map whose scores and values are definable in $\mathbb{R}_{\exp}$ is again definable in $\mathbb{R}_{\exp}$.

Proof. Each coordinate $z \mapsto e^{z_k}$ is definable in $\mathbb{R}_{\exp}$, and the denominator

$$\sum_{j=1}^{K} e^{z_j}$$

is a finite sum of strictly positive $\mathbb{R}_{\exp}$-definable functions. Hence the denominator is strictly positive and definable. Division by a strictly positive definable function preserves definability, so each coordinate of $\mathrm{softmax}$ is definable in $\mathbb{R}_{\exp}$.

If the score maps $s_1, \ldots, s_K$ and value maps $v_1, \ldots, v_K$ are definable, then

$$x \mapsto \sum_{k=1}^{K} \frac{e^{s_k(x)}}{\sum_{j=1}^{K} e^{s_j(x)}}\, v_k(x)$$

is obtained from definable maps by composition, finite summation, multiplication, and division by a strictly positive definable denominator. Thus it is definable in $\mathbb{R}_{\exp}$. □

This directly applies to the attention mechanism.

Proof of Proposition 4.5. For each index h, n, m, the score map $(X, W_Q^h, W_K^h) \mapsto \lambda\langle W_Q^h X_n, W_K^h X_m\rangle/\sqrt{d_k}$ is polynomial in the coordinates of (X, W_Q^h, W_K^h). Since the exponential map is definable in $\mathbb{R}_{\exp}$, the numerator of each attention weight is $\mathbb{R}_{\exp}$-definable. The denominator is a finite sum of strictly positive $\mathbb{R}_{\exp}$-definable functions, hence is strictly positive and definable. Therefore each attention weight $\alpha_{n,m}^{(h)}$ is definable in $\mathbb{R}_{\exp}$, jointly in X, W_Q^h, W_K^h.

Next, the value map $(X, W_V^h) \mapsto W_V^h X_m$ is polynomial, hence semi-algebraic. Therefore each head output $(X, W_Q^h, W_K^h, W_V^h) \mapsto \mathrm{Attn}_h(X)_n = \sum_{m=1}^N \alpha_{n,m}^{(h)}(X)\, W_V^h X_m$ is obtained from definable maps by finite summation and multiplication, and is therefore definable in $\mathbb{R}_{\exp}$.

Finally, concatenation over heads is definable, and the output projection $((z_1, \ldots, z_H), W_O) \mapsto W_O(z_1, \ldots, z_H)$ is polynomial in the concatenated head outputs and in W_O. Hence MHA is definable in $\mathbb{R}_{\exp}$, jointly in X and the parameters $(W_Q^h, W_K^h, W_V^h)_{h=1}^H, W_O$. □

Corollary C.5 (Definability of Multi-Head Cross-Attention)**.** *Fix $N_Q, N_C, H, d_Q, d_C, d_k, d_v, d_{out} \in \mathbb{N}_+$ and $\lambda > 0$. For each $h \in \{1, \ldots, H\}$, let $W_Q^h \in \mathbb{R}^{d_k \times d_Q}$, $W_K^h \in \mathbb{R}^{d_k \times d_C}$, and $W_V^h \in \mathbb{R}^{d_v \times d_C}$, and let $W_O \in \mathbb{R}^{d_{out} \times H d_v}$. For $X \in \mathbb{R}^{N_Q \times d_Q}$ and $Y \in \mathbb{R}^{N_C \times d_C}$, define*

$$\mathrm{CrossAttn}_h(X, Y)_n \overset{\text{def.}}{=} \sum_{m=1}^{N_C} \frac{\exp\big(\lambda\langle W_Q^h X_n, W_K^h Y_m\rangle/\sqrt{d_k}\big)}{\sum_{\ell=1}^{N_C} \exp\big(\lambda\langle W_Q^h X_n, W_K^h Y_\ell\rangle/\sqrt{d_k}\big)} W_V^h Y_m.$$

Then the parametrized multi-head cross-attention map

$$(X, Y, (W_Q^h, W_K^h, W_V^h)_{h=1}^H, W_O) \longmapsto \mathrm{CrossMHA}(X, Y),$$

where $\mathrm{CrossMHA}(X, Y)_n \overset{\text{def.}}{=} W_O(\mathrm{CrossAttn}_1(X, Y)_n, \ldots, \mathrm{CrossAttn}_H(X, Y)_n)$*, is definable in $\mathbb{R}_{\exp}$, jointly in (X, Y) and the parameters.*

Proof. The proof is identical to Proposition 4.5: the scores are polynomial, the softmax weights are definable in $\mathbb{R}_{\exp}$, the value maps and output projection are polynomial, and only finite sums, products, concatenations, and divisions by strictly positive definable denominators are used. □

Corollary C.6 (Definability of Sliding-Window Self-Attention)**.** *Fix $N, H, d_{in}, d_k, d_v, d_{out} \in \mathbb{N}_+$ and $\lambda > 0$. For each $n \in \{1, \ldots, N\}$, let $\mathcal{W}(n) \subseteq \{1, \ldots, N\}$ be a fixed nonempty finite window. With the notation of Proposition 4.5, define*

$$\mathrm{SWAttn}_h(X)_n \overset{\text{def.}}{=} \sum_{m \in \mathcal{W}(n)} \frac{\exp\big(\lambda\langle W_Q^h X_n, W_K^h X_m\rangle/\sqrt{d_k}\big)}{\sum_{\ell \in \mathcal{W}(n)} \exp\big(\lambda\langle W_Q^h X_n, W_K^h X_\ell\rangle/\sqrt{d_k}\big)} W_V^h X_m.$$

Then the corresponding parametrized multi-head sliding-window self-attention map is definable in $\mathbb{R}_{\exp}$, jointly in X and the parameters.

Proof. The proof is identical to Proposition 4.5, except that the finite sums are taken over the fixed windows $\mathcal{W}(n)$ rather than over all tokens. Fixed finite restrictions, finite sums, products, exponentials, concatenations, and divisions by strictly positive definable denominators preserve definability. □

Proof of Proposition 4.6. For each $G \in \mathcal{G}$, the map $x \mapsto \mu_G(x)$ is affine, and the map $x \mapsto \sigma_G^2(x)$ is polynomial. Since $\varepsilon > 0$, one has $\sigma_G^2(x) + \varepsilon > 0$ for every $x \in \mathbb{R}^d$. Moreover, the graph of $x \mapsto \sqrt{\sigma_G^2(x) + \varepsilon}$ is

$$\Big\{(x, y) \in \mathbb{R}^{d+1} : y \geq 0,\ y^2 = \sigma_G^2(x) + \varepsilon\Big\},$$

which is semi-algebraic. Hence $x \mapsto \sqrt{\sigma_G^2(x) + \varepsilon}$ is semi-algebraic.

Since the denominator is strictly positive, division by this term is also semi-algebraic. Thus each coordinate map

$$x \mapsto \gamma_i \frac{x_i - \mu_{G(i)}(x)}{\sqrt{\sigma^2_{G(i)}(x) + \varepsilon}} + \beta_i$$

is semi-algebraic. Therefore $\mathrm{Norm}_{\gamma,\beta,\varepsilon}$ is semi-algebraic.

Layer normalization is obtained by taking a single block $G = \{1, \dots, d\}$. Group normalization and instance normalization correspond to fixed partitions of the coordinates into groups. Finite-batch batch normalization is the same construction after viewing the batch and feature coordinates as one finite-dimensional input and choosing the blocks over which the batch statistics are computed. At inference time, batch normalization uses fixed population statistics, and is therefore affine. Finally, RMS normalization follows identically, since $x \mapsto |G|^{-1} \sum_{j \in G} x_j^2$ is polynomial. This proves the claim. □

Proof of Proposition 4.7. The graph of Emb is

$$\mathrm{Graph}(\mathrm{Emb}) = \bigcup_{r=1}^{N} \Big\{ (i, E, y) \in \mathbb{R} \times \mathbb{R}^{N \times d} \times \mathbb{R}^d : i = r,\ y = E_{r,:} \Big\}.$$

For each $r \in \{1, \dots, N\}$, the set

$$\Big\{ (i, E, y) : i = r,\ y = E_{r,:} \Big\}$$

is defined by the polynomial equalities $i - r = 0$ and $y_j - E_{rj} = 0$ for $j = 1, \dots, d$. Hence each set in the union is semi-algebraic. Since finite unions of semi-algebraic sets are semi-algebraic, $\mathrm{Graph}(\mathrm{Emb})$ is semi-algebraic. Therefore Emb is semi-algebraic, and hence definable in the real field and in every o-minimal expansion of the real field. □

Proof of Proposition 4.9. For $j \in \{1, \dots, M\}$, set $a_j(t) \overset{\text{def.}}{=} 2\pi\langle \omega_j, t \rangle + \varphi_j$. Since D is bounded and the frequencies and phases are fixed, there exists $R > 0$ such that

$$|a_j(t)| \le R, \qquad t \in D, \quad j \in \{1, \dots, M\}.$$

Define the restricted analytic functions

$$s_R(u) \overset{\text{def.}}{=} \begin{cases} \sin(Ru), & |u| \le 1, \\ 0, & |u| > 1, \end{cases} \qquad c_R(u) \overset{\text{def.}}{=} \begin{cases} \cos(Ru), & |u| \le 1, \\ 0, & |u| > 1. \end{cases}$$

Then

$$\mathrm{PE}_{\Omega,\varphi}(t) = \big(s_R(a_j(t)/R), c_R(a_j(t)/R) \big)_{j=1}^{M}, \qquad t \in D.$$

The maps a_j/R are affine, s_R and c_R are $\mathbb{R}_{\mathrm{an}}$-definable, and D is $\mathbb{R}_{\mathrm{an}}$-definable. Closure under composition and restriction proves the first assertion. If $D = \{1, \dots, N\}$, then

$$\mathrm{graph}(\mathrm{PE}_{\Omega,\varphi}) = \bigcup_{i=1}^{N} \big\{ (i, \mathrm{PE}_{\Omega,\varphi}(i)) \big\}.$$

This is a finite union of singletons with fixed real coordinates, hence a semi-algebraic set. □

Remark C.7 (Bounded Phases and Trainable Frequencies). Let $A \subseteq \mathbb{R}^r \times (\mathbb{R}^r)^M \times \mathbb{R}^M$ be an $\mathbb{R}_{\mathrm{an}}$-definable joint input–parameter domain. The preceding proof applies whenever

$$\sup_{(t,\Omega,\varphi) \in A} \max_{1 \le j \le M} \big| 2\pi\langle \omega_j, t \rangle + \varphi_j \big| < \infty.$$

Under this condition, $(t, \Omega, \varphi) \mapsto \mathrm{PE}_{\Omega,\varphi}(t)$ is jointly $\mathbb{R}_{\mathrm{an}}$-definable. In particular, bounded definable domains for positions, frequencies, and phases suffice.

Unrestricted frequencies can violate definability even at a fixed nonzero position. Indeed, for $t_0 \neq 0$ and fixed φ_1, define

$$\omega_1(u) \overset{\text{def.}}{=} \frac{u - \varphi_1}{2\pi \|t_0\|_2^2}\, t_0, \qquad u \in \mathbb{R}.$$

Then

$$\sin\left(2\pi \langle \omega_1(u), t_0 \rangle + \varphi_1\right) = \sin u.$$

Consequently, definability on a domain containing this entire parameter slice would imply definability of global sine. This is impossible in an o-minimal structure because $\{u : \sin u = 0\} = \pi\mathbb{Z}$ is infinite and discrete. Boundedness of the position domain itself is not necessary: for example, $\omega_j = 0$ for every j makes the encoding constant in t.

Pooling Layers. Pooling layers are among the classical mechanisms used to build local invariance in hierarchical vision architectures, going back at least to the local aggregation mechanisms of the Neocognitron (Fukushima, 1980), the trainable sub-sampling layers of LeNet-style convolutional neural networks (LeCun et al., 1998), and the MAX-like operations used in hierarchical object-recognition models (Riesenhuber & Poggio, 1999). Their widespread use in large-scale convolutional neural networks was further reinforced by architectures such as AlexNet (Krizhevsky et al., 2012). Additionally, they are also common in GNNs (Bronstein et al., 2021). From the o-minimal viewpoint, these layers are harmless: average pooling is affine, while max-pooling is semi-algebraic.

Proposition C.8 (Definability of Pooling Layers)**.** *Fix $d \in \mathbb{N}_+$. The following are definable: (i) Average Pooling: The average-pooling map $P_{\text{avg}} : \mathbb{R}^d \to \mathbb{R}$ defined by $P_{\text{avg}}(x_1, \ldots, x_d) \overset{\text{def.}}{=} \frac{1}{d}\sum_{i=1}^{d} x_i$ is definable in any o-minimal structure. (ii) Max Pooling: The max-pooling map $P_{\max} : \mathbb{R}^d \to \mathbb{R}$ defined by $P_{\max}(x_1, \ldots, x_d) \overset{\text{def.}}{=} \max_{1 \leq i \leq d} x_i$ is definable in any o-minimal structure.*

Proof of Proposition C.8.(i). The map P_{avg} is affine, hence polynomial. Therefore its graph is algebraic, and in particular semi-algebraic. Since polynomial maps are Pfaffian functions, P_{avg} is Pfaffian. □

Proof of Proposition C.8.(ii). We first treat the case $d = 2$. For $a, b \in \mathbb{R}$,

$$\max\{a, b\} = \frac{a + b + |a - b|}{2}.$$

The graph of $t \mapsto |t|$ is semi-algebraic, since

$$\text{graph}(|\cdot|) = \left\{(t, s) \in \mathbb{R}^2 : s \geq 0,\ s^2 = t^2\right\}.$$

Thus $(a, b) \mapsto \max\{a, b\}$ is semi-algebraic. For general $d \in \mathbb{N}_+$, one writes

$$P_{\max}(x_1, \ldots, x_d) = \max\left\{x_1, \max\{x_2, \ldots, x_d\}\right\}$$

and iterates the two-variable construction finitely many times. Since finite compositions of semi-algebraic maps are semi-algebraic, $P_{\max}$ is semi-algebraic. Hence it is definable in the real field $(\mathbb{R}, +, \cdot, <)$, and therefore in every o-minimal expansion of the real field. □

Proof of Proposition C.8; final claim. A pooling layer acting on finitely many fixed windows is obtained by applying P_{avg} or $P_{\max}$ to each window, followed by coordinate projection and finite concatenation. Coordinate projections, finite products, and finite concatenations preserve definability. Therefore any finite pooling layer built from fixed average-pooling and max-pooling windows is definable. □

Deep Equilibrium Layers. Deep equilibrium models (Bai et al., 2019) belong to the broader class of implicit-depth models, where the output of a layer is not obtained by applying a prescribed finite composition of maps, but rather by solving an equation whose solution defines the hidden representation. This viewpoint also underlies several influential architectures and implicit layers: including neural ordinary differential equations (ODEs) (Chen et al., 2018) and differentiable optimization layers (Amos & Kolter, 2017).

Proposition C.9 (Definability of Deep Equilibrium Layers)**.** *Let $\mathfrak{S}$ be an o-minimal expansion of the real field. Let $X \subseteq \mathbb{R}^{d_x}$, $Z \subseteq \mathbb{R}^{d_z}$, and $\Theta \subseteq \mathbb{R}^p$ be nonempty $\mathfrak{S}$-definable sets, and let*

$$F : X \times Z \times \Theta \to Z, \qquad G : X \times Z \times \Theta \to \mathbb{R}^{d_y}$$

be $\mathfrak{S}$-definable maps. Assume that, for every $(x, \theta) \in X \times \Theta$, there exists a unique $z(x, \theta) \in Z$ satisfying

$$z(x, \theta) = F(x, z(x, \theta), \theta).$$

Then $\mathrm{DEQ} : X \times \Theta \to \mathbb{R}^{d_y}$ given by $\mathrm{DEQ}(x, \theta) = G(x, z(x, \theta), \theta)$, is $\mathfrak{S}$-definable jointly in its inputs and parameters. In particular, if all the preceding data are definable in $\mathbb{R}_{\exp}$, then so is DEQ.

Proof of Proposition C.9. Define

$$\Gamma \overset{\text{def.}}{=} \big\{(x, \theta, z) \in X \times \Theta \times Z : z - F(x, z, \theta) = 0\big\}.$$

The map $(x, \theta, z) \mapsto z - F(x, z, \theta)$ is $\mathfrak{S}$-definable; hence Γ is $\mathfrak{S}$-definable. By assumption,

$$\Gamma_{(x,\theta)} \overset{\text{def.}}{=} \{z \in Z : (x, \theta, z) \in \Gamma\} = \{z(x, \theta)\}, \qquad (x, \theta) \in X \times \Theta.$$

Thus Γ is the graph of $(x, \theta) \mapsto z(x, \theta)$, proving its definability. Moreover,

$$\mathrm{graph}(\mathrm{DEQ}) = \big\{(x, \theta, y) \in X \times \Theta \times \mathbb{R}^{d_y} : \exists z \in Z,\ (x, \theta, z) \in \Gamma,\ y = G(x, z, \theta)\big\}.$$

This set is $\mathfrak{S}$-definable by closure under finite intersections and coordinate projections. The specialization to $\mathbb{R}_{\exp}$ follows by applying the same argument in that structure. □

Proof of Corollary 4.10. Applying Proposition 2.5 to Propositions 4.1 and 4.3 implies the joint definability of f_θ over all of $\mathbb{R}^{d_0} \times \mathbb{R}^P$. Thus, Theorems 3.1 and 3.2 apply. □

Proof of Corollary 4.11. Applying Proposition 2.5 to Proposition 4.7, 4.9, 4.6, 4.3, we deduce that f_θ is jointly definable in $(x, \theta) \in \mathbb{R}^{H_1 \times d_1} \times \mathbb{R}^P$. Thus, Theorems 3.1 and 3.2 apply. □

C.1 How Pfaffian Properties Preserve Complexity under Common Network Transformations

Residual connections of the form $x \mapsto \varphi(x) + x$ occur in ResNets He et al. (2016) and Transformers Vaswani et al. (2017). Their formats are controlled by the sum rule.

Corollary C.10 (Residual connections)**.** *Let $U \subseteq \mathbb{R}^n$ be open, and let $\varphi = (\varphi_1, \ldots, \varphi_n) : U \to \mathbb{R}^n$ have coordinate functions of format at most (q, D, d) over a common Pfaffian chain. Then $x \mapsto \varphi(x) + x$ has coordinate functions of format at most $(q, D, \max\{d, 1\})$ over the same chain. In particular, the format bound is unchanged when $d \geq 1$.*

Proof. If Q_i represents φ_i over the common chain, then $Q_i + x_i$ represents the i-th residual coordinate and has degree at most $\max\{d, 1\}$. The same argument applies jointly in additional parameter variables. □

Remark C.11 (Residuals with trainable scaling). For a trainable scalar γ, the skip term γx has total degree two jointly in (x, γ). Therefore $(x, \gamma) \mapsto \varphi(x) + \gamma x$ has coordinate formats at most $(q, D, \max\{d, 2\})$, after extending the original chain to be independent of γ. More generally, the same bound holds if φ already depends jointly on γ and has the stated format in all its input and parameter variables. For a fixed scalar γ, the skip term has degree at most one.

Corollary C.12 (Pfaffian functions under fixed linear or affine maps)**.** *Let φ be Pfaffian of format at most (q, D, d) on an open domain $U \subseteq \mathbb{R}^n$, and let $L : \mathbb{R}^a \to \mathbb{R}^n$ be a fixed affine map $L(x) = Ax + b$. If $L^{-1}(U)$ is nonempty, then $\varphi \circ L$ is Pfaffian there of format at most (q, D, d).*

Proof. Let $(f_1, \ldots, f_q)$ be a chain for φ, with differential polynomials $P_{i,j}$ of degree at most D. For $s_i \overset{\text{def.}}{=} f_i \circ L$, one has

$$\frac{\partial s_i}{\partial x_k} = \sum_{j=1}^{n} A_{jk} P_{i,j}\big(L(x), s_1, \ldots, s_i\big).$$

Since A and b are fixed, each right-hand side is a polynomial of degree at most D in $(x, s_1, \ldots, s_i)$. Thus $(s_1, \ldots, s_q)$ is a chain of the same length and degree on the open set $L^{-1}(U)$. Affine substitution into the representing polynomial of φ preserves its degree bound d. □

Remark C.13 (Linear projections in neural networks). Corollary C.12 applies to fixed projection coefficients. When A and b are trainable, the map $(x, A, b) \mapsto Ax + b$ has joint degree at most two. Lemma B.27 then gives format at most $(q, 2D+1, 2d)$ for $(x, A, b) \mapsto \varphi(Ax+b)$ on its open domain of definition. In the PNN format estimates, a trainable linear projection or readout must therefore be included in the layer formats and parameter count. For example, the trainable readout in (2.3) can be treated as an additional polynomial layer with $q_k = D_k = 0$ and $d_{k,i} = 2$.

D Future Work

We conclude by outlining promising directions for future research.

Quantitative Versions via $\sharp$o-Minimality. Our theory shows how o-minimality implies finite sample complexity in the PAC-learning model. We expect that precise quantitative versions of our qualitative results can be derived by relying on the emerging notion of $\sharp$o-minimality Binyamini & Novikov (2023); Binyamini et al. (2024; 2022), since this new model-theoretic tool allows one to control the number of cells/osculations of a jointly definable function through notions of degree and format, akin to the corresponding complexity theory for Pfaffian functions; cf. Khovanskiĭ (1991).

Definability Beyond PAC Learning. It would be interesting to apply our proof strategy to obtain general learning guarantees in other models, where bounds on the number of definable cells, or osculations, yield uniform control on the relevant learning-theoretic dimensions; e.g., the Littlestone dimension for online learning Littlestone (1988), the γ-Natarajan dimension Simon (1997) for multi-class classification, or other extensions for robust PAC Gourdeau et al. (2021) and computable PAC learning; cf. Agarwal et al. (2021); Delle Rose et al. (2023).

AI use statement

We used ChatGPT to assist with manuscript editing, LaTeX formatting, and the development and critical examination of mathematical arguments, including suggestions for proofs and the identification of potential gaps. The authors reviewed all AI-assisted material and independently checked the mathematical claims, proofs, and references. The authors take full responsibility for the final manuscript.