JMLR: Workshop and Conference Proceedings vol 49:1–23, 2016

Benefits of depth in neural networks

Matus Telgarsky MTELGARS@CS.UCSD.EDU
University of Michigan, Ann Arbor

Abstract

For any positive integer k, there exist neural networks with $\Theta(k^3)$ layers, $\Theta(1)$ nodes per layer, and $\Theta(1)$ distinct parameters which can not be approximated by networks with $\mathcal{O}(k)$ layers unless they are exponentially large — they must possess $\Omega(2^k)$ nodes. This result is proved here for a class of nodes termed *semi-algebraic gates* which includes the common choices of ReLU, maximum, indicator, and piecewise polynomial functions, therefore establishing benefits of depth against not just standard networks with ReLU gates, but also convolutional networks with ReLU and maximization gates, sum-product networks, and boosted decision trees (in this last case with a stronger separation: $\Omega(2^{k^3})$ total tree nodes are required).

Keywords: Neural networks, representation, approximation, depth hierarchy.

1. Setting and main results

A neural network is a model of real-valued computation defined by a connected directed graph as follows. Nodes await real numbers on their incoming edges, thereafter computing a function of these reals and transmitting it along their outgoing edges. Root nodes apply their computation to a vector provided as input to the network, whereas internal nodes apply their computation to the output of other nodes. Different nodes may compute different functions, two common choices being the maximization gate $v \mapsto \max_i v_i$ (where v is the vector of values on incoming edges), and the *standard ReLU gate* $v \mapsto \sigma_{\mathrm{R}}(\langle a, v\rangle + b)$ where $\sigma_{\mathrm{R}}(z) := \max\{0, z\}$ is called the ReLU (rectified linear unit), and the parameters a and b may vary from node to node. Graphs in the present work are acyclic, and there is exactly one node with no outgoing edges whose computation is the output of the network.

Neural networks distinguish themselves from many other function classes used in machine learning by possessing multiple *layers*, meaning the output is the result of composing together an arbitrary number of (potentially complicated) nonlinear operations; by contrast, the functions computed by boosted decision stumps and SVMs can be written as neural networks with a constant number of layers.

The purpose of the present work is to show that standard types of networks always gain in representation power with the addition of layers. Concretely: it is shown that for every positive integer k, there exist neural networks with $\Theta(k^3)$ layers, $\Theta(1)$ nodes per layer, and $\Theta(1)$ distinct parameters which can not be approximated by networks with $\mathcal{O}(k)$ layers and $o(2^k)$ nodes.

1.1. Main result

Before stating the main result, a few choices and pieces of notation deserve explanation. First, the target many-layered function uses standard ReLU gates; this is by no means necessary, and a more general statement can be found in Theorem 3.12. Secondly, the notion of approximation is the L^1

distance: given two functions f and g, their pointwise disagreement $|f(x) - g(x)|$ is averaged over the cube $[0,1]^d$. Here as well, the same proofs allow flexibility (cf. Theorem 3.12). Lastly, the shallower networks used for approximation use *semi-algebraic gates*, which generalize the earlier maximization and standard ReLU gates, and allow for analysis of not just standard networks with ReLU gates, but convolutional networks with ReLU and maximization gates (Krizhevsky et al., 2012), sum-product networks (where nodes compute polynomials) (Poon and Domingos, 2011), and boosted decision trees; the full definition of semi-algebraic gates appears in Section 2.

Theorem 1.1 *Let any integer $k \geq 1$ and any dimension $d \geq 1$ be given. There exists $f : \mathbb{R}^d \to \mathbb{R}$ computed by a neural network with standard ReLU gates in $2k^3 + 8$ layers, $3k^3 + 12$ total nodes, and $4 + d$ distinct parameters so that*

$$\inf_{g \in \mathcal{C}} \int_{[0,1]^d} |f(x) - g(x)| dx \geq \frac{1}{64},$$

where $\mathcal{C}$ is the union of the following two sets of functions.

- *Functions computed by networks of (t, α, β)-semi-algebraic gates in $\leq k$ layers and $\leq 2^k/(t\alpha\beta)$ nodes. (E.g., as with standard ReLU networks or with convolutional neural networks with standard ReLU and maximization gates; cf. Section 2.)*
- *Functions computed by linear combinations of $\leq t$ decision trees each with $\leq 2^{k^3}/t$ nodes. (E.g., the function class used by boosted decision trees; cf. Section 2.)*

Analogs to Theorem 1.1 for boolean circuits — which have boolean inputs routed through $\{\text{and}, \text{or}, \text{not}\}$ gates — have been studied extensively by the circuit complexity community, where they are called *depth hierarchy theorems*. The seminal result, due to Håstad (1986), establishes the inapproximability of the parity function by shallow circuits (unless their size is exponential). Standard neural networks appear to have received less study; closest to the present work is an investigation by Eldan and Shamir (2015) analyzing the case $k = 2$ when the dimension d is large, showing an exponential separation between 2- and 3-layer networks, a regime not handled by Theorem 1.1. Further bibliographic notes and open problems may be found in Section 5.

The proof of Theorem 1.1 (and of the more general Theorem 3.12) occupies Section 3. The key idea is that just a few function compositions (layers) suffice to construct a highly oscillatory function, whereas function addition (adding nodes but keeping depth fixed) gives a function with few oscillations. Thereafter, an elementary counting argument suffices to show that low-oscillation functions can not approximate high-oscillation functions.

1.2. Companion results

Theorem 1.1 only provides the existence of *one* network (for each k) which can not be approximated by a network with many fewer layers. It is natural to wonder if there are *many* such special functions. The following bound indicates their population is in fact quite modest.

Specifically, the construction behind Theorem 1.1, as elaborated in Theorem 3.12, can be seen as exhibiting $\mathcal{O}(2^{k^3})$ points, and a fixed labeling of these points, upon which a shallow network hardly improves upon random guessing. The forthcoming Theorem 1.2 similarly shows that even on the more simpler task of fitting $\mathcal{O}(k^9)$ points, the earlier class of networks is useless on most random labellings.

In order to state the result, a few more definitions are in order. Firstly, for this result, the notion of neural network is more restrictive. Let a *neural net graph* $\mathfrak{G}$ denote not only the graph structure (nodes and edges), but also an assignment of gate functions to nodes, of edges to the inputs of gates, and an assignment of free parameters $w \in \mathbb{R}^p$ to the parameters of the gates. Let $\mathcal{N}(\mathfrak{G})$ denote the class of functions obtained by varying the free parameters; this definition is fairly standard, and is discussed in more detail in Section 2. As a final piece of notation, given a function $f : \mathbb{R}^d \to \mathbb{R}$, let $\tilde{f} : \mathbb{R}^d \to \{0, 1\}$ denote the corresponding classifier $\tilde{f}(x) := \mathbf{1}[f(x) \geq 1/2]$.

Theorem 1.2 *Let any neural net graph $\mathfrak{G}$ be given with $\leq p$ parameters in $\leq l$ layers and $\leq m$ total (t, α, β)-semi-algebraic nodes. Then for any $\delta > 0$ and any $n \geq 8pl^2 \ln(8emt\alpha\beta p(l+1)) + 4\ln(1/\delta)$ points $(x_i)_{i=1}^n$, with probability $\geq 1 - \delta$ over uniform random labels $(y_i)_{i=1}^n$,*

$$\inf_{f \in \mathcal{N}(\mathfrak{G})} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[\tilde{f}(x_i) \neq y_i] \geq \frac{1}{4}.$$

This proof is a direct corollary of the VC dimension of semi-algebraic networks, which in turn can be proved by a small modification of the VC dimension proof for piecewise polynomial networks (Anthony and Bartlett, 1999, Theorem 8.8). Moreover, the core methodology for VC dimension bounds of neural networks is due to Warren, whose goal was an analog of Theorem 1.2 for polynomials (Warren, 1968, Theorem 7).

Lemma 1.3 (Simplification of Lemma 4.2) *Let any neural net graph $\mathfrak{G}$ be given with $\leq p$ parameters in $\leq l$ layers and $\leq m$ total nodes, each of which is (t, α, β)-semi-algebraic. Then*

$$\mathrm{VC}(\mathcal{N}(\mathfrak{G})) \leq 6p(l+1)\big(\ln(2p(l+1)) + \ln(8emt\alpha) + l\ln(\beta)\big).$$

The proof of Theorem 1.2 and Lemma 1.3 may be found in Section 4. The argument for the VC dimension is very close to the argument for Theorem 1.1 that a network with few layers has few oscillations; see Section 4 for further discussion of this relationship.

2. Semi-algebraic gates and assorted network notation

The definition of a semi-algebraic gate is unfortunately complicated; it is designed to capture a few standard nodes in a single abstraction without degrading the bounds. Note that the name *semi-algebraic set* is standard (Bochnak et al., 1998, Definition 2.1.4), and refers to a set defined by unions and intersections of polynomial inequalities (and thus the name is somewhat abused here).

Definition 2.1 *A function $f : \mathbb{R}^k \to \mathbb{R}$ is (t, α, β)-*sa *((t, α, β)-*semi-algebraic*) if there exist t polynomials $(q_i)_{i=1}^t$ of degree $\leq \alpha$, and m triples $(U_j, L_j, p_j)_{j=1}^m$ where U_j and L_j are subsets of $[t]$ (where $[t] := \{1, \ldots, t\}$) and p_j is a polynomial of degree $\leq \beta$, such that*

$$f(v) = \sum_{j=1}^{m} p_j(v) \left(\prod_{i \in L_j} \mathbf{1}[q_i(v) < 0] \right) \left(\prod_{i \in U_j} \mathbf{1}[q_i(v) \geq 0] \right).$$

A notable trait of the definition is that the number of terms m does not need to enter the name as it does not affect any of the complexity estimates herein (e.g., Theorem 1.1 or Theorem 1.2).

Distinguished special cases of semi-algebraic gates are as follows in Lemma 2.3. The standard piecewise polynomial gates generalize the ReLU and have received a fair bit of attention in the theoretical community (Anthony and Bartlett, 1999, Chapter 8); here a function $\sigma : \mathbb{R} \to \mathbb{R}$ is (t, α)-*poly* if $\mathbb{R}$ can be partitioned into $\leq t$ intervals so that σ is a polynomial of degree $\leq \alpha$ within each piece. The maximization and minimization gates have become popular due to their use in convolutional networks (Krizhevsky et al., 2012), which will be discussed more in Section 2.1. Lastly, decision trees and boosted decision trees are practically successful classes usually viewed as competitors to neural networks (Caruana and Niculescu-Mizil, 2006), and have the following structure.

Definition 2.2 *A* k-dt (decision tree with k nodes) *is defined recursively as follows. If* $k = 1$*, it is a constant function. If* $k > 1$*, it first evaluates* $x \mapsto \mathbf{1}[\langle a, x\rangle - b \geq 0]$*, and thereafter conditionally evaluates either a left* l*-dt or a right* r*-dt where* $l + r < k$*. A* (t, k)-bdt *(boosted decision tree) evaluates* $x \mapsto \sum_{i=1}^{t} c_i g_i(x)$ *where each* $c_i \in \mathbb{R}$ *and each* g_i *is a* k*-dt.*

Lemma 2.3 (Example semi-algebraic gates)

1. *If* $\sigma : \mathbb{R} \to \mathbb{R}$ *is* (t, β)*-poly and* $q : \mathbb{R}^d \to \mathbb{R}$ *is a polynomial of degree* α*, then the standard piecewise polynomial gate* $\sigma \circ q$ *is* $(t, \alpha, \alpha\beta)$*-sa. In particular, the standard ReLU gate* $v \mapsto \sigma_{\mathrm{R}}(\langle a, v\rangle + b)$ *is* $(1, 1, 1)$*-sa.*

2. *Given polynomials* $(p_i)_{i=1}^{r}$ *of degree* $\leq \alpha$*, the standard* (r, α)*-min and -max gates* $\phi_{\min}(v) := \min_{i\in[r]} p_i(v)$ *and* $\phi_{\max}(v) := \max_{i\in[r]} q_i(v)$ *are* $(r(r-1), \alpha, \alpha)$*-sa.*

3. *Every* k*-dt is* $(k, 1, 0)$*-sa, and every* (t, k)*-bdt is* $(tk, 1, 0)$*.*

The proof of Lemma 2.3 is mostly a matter of unwrapping definitions, and is deferred to Appendix A. Perhaps the only interesting encoding is for the maximization gate (and similarly the minimization gate), which uses $\max_i v_i = \sum_i v_i(\prod_{j<i} \mathbf{1}[v_i > v_j])(\prod_{j>i} \mathbf{1}[v_i \geq v_j])$.

2.1. Notation for neural networks

A semi-algebraic gate is simply a function from some domain to $\mathbb{R}$, but its role in a neural network is more complicated as the domain of the function must be partitioned into arguments of three types: the input $x \in \mathbb{R}^d$ to the network, the parameter vector $w \in \mathbb{R}^p$, and a vector of real numbers coming from parent nodes.

As a convention, the input $x \in \mathbb{R}^d$ is only accessed by the root nodes (otherwise "layer" has no meaning). For convenience, let layer 0 denote the input itself: d nodes where node i is the map $x \mapsto x_i$. The parameter vector $w \in \mathbb{R}^p$ will be made available to all nodes in layers above 0, though they might only use a subset of it. Specifically, an internal node computes a function $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ using parents $(f_1, \ldots, f_k)$ and a semi-algebraic gate $\phi : \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}$, meaning $f(w, x) := \phi(w_1, \ldots, w_p, f_1(w, x), \ldots, f_k(w, x))$. Another common practice is to have nodes apply a univariate *activation function* to an affine mapping of their parents (as with piecewise polynomial gates in Lemma 2.3), where the weights in the affine combination are the parameters to the network, and additionally correspond to edges in the graph. It is permitted for the same

parameter to appear multiple times in a network, which explains how the number of parameters in Theorem 1.1 can be less than the number of edges and nodes. The entire network computes some function $F_{\mathfrak{G}} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$, which is equivalent to the function computed by the single node with no outgoing edges.

As stated previously, $\mathfrak{G}$ will denote not just the graph (nodes and edges) underlying a network, but also an assignment of gates to nodes, and how parameters and parent outputs are plugged into the gates (i.e., in the preceding paragraph, how to write f via ϕ). $\mathcal{N}(\mathfrak{G})$ is the set of functions obtained by varying $w \in \mathbb{R}^p$, and thus $\mathcal{N}(\mathfrak{G}) := \{F_{\mathfrak{G}}(w, \cdot) : w \in \mathbb{R}^p\}$ where $F_{\mathfrak{G}}$ is the function defined as above, corresponding to computation performed by $\mathfrak{G}$. The results related to VC dimension, meaning Theorem 1.2 and Lemma 1.3, will use the class $\mathcal{N}(\mathfrak{G})$.

Some of the results, for instance Theorem 1.1 and its generalization Theorem 3.12, will let not only the parameters but also network graph $\mathfrak{G}$ vary. Let $\mathcal{N}_d((m_i, t_i, \alpha_i, \beta_i)_{i=1}^l)$ denote a network where layer i has $\leq m_i$ nodes where each is (t_i, α_i, β_i)-sa and the input has dimension d. As a simplification, let $\mathcal{N}_d(m, l, t, \alpha, \beta)$ denote networks of (t, α, β)-sa gates in $\leq l$ layers (not including layer 0) each with $\leq m$ nodes. There are various empirical prescriptions on how to vary the number of nodes per layer; for instance, convolutional networks typically have an increase between layer 0 and layer 1, followed by exponential decrease for a few layers, and finally a few layers with the same number of nodes (Fukushima, 1980; LeCun et al., 1998; Krizhevsky et al., 2012).

3. Benefits of depth

The purpose of this section is to prove Theorem 1.1 and its generalization Theorem 3.12 in the following three steps.

1. Functions with few oscillations poorly approximate functions with many oscillations.
2. Functions computed by networks with few layers must have few oscillations.
3. Functions computed by networks with many layers can have many oscillations.

3.1. Approximation via oscillation counting

The idea behind this first step is depicted at right. Given functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ (the multivariate case will come soon), let $\mathcal{I}_f$ and $\mathcal{I}_g$ denote partitions of $\mathbb{R}$ into intervals so that the classifiers $\tilde{f}(x) = \mathbf{1}[f(x) \geq 1/2]$ and $\tilde{g}$ are constant within each interval. To formally count oscillations, define *the crossing number* $\mathrm{Cr}(f)$ *of* f as $\mathrm{Cr}(f) = |\mathcal{I}_f|$ (thus $\mathrm{Cr}(\sigma_{\mathrm{R}}) = 2$). If $\mathrm{Cr}(f)$ is much larger than $\mathrm{Cr}(g)$, then most piecewise constant regions of $\tilde{g}$ will exhibit many oscillations of f, and thus g poorly approximates f.

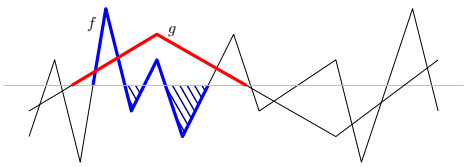

Figure 1: f crosses more than g.

Lemma 3.1 *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be given, and take $\mathcal{I}_f$ to denote the partition of $\mathbb{R}$ given by the pieces of $\tilde{f}$ (meaning $|\mathcal{I}_f| = \mathrm{Cr}(f)$). Then*

$$\frac{1}{\mathrm{Cr}(f)} \sum_{U \in \mathcal{I}_f} \mathbf{1}[\forall x \in U \,.\, \tilde{f}(x) \neq \tilde{g}(x)] \geq \frac{1}{2}\left(1 - 2\left(\frac{\mathrm{Cr}(g)}{\mathrm{Cr}(f)}\right)\right).$$

The arguably strange form of the left hand side of the bound in Lemma 3.1 is to accommodate different notions of distance. For the L^1 distance with the Lebesgue measure as in Theorem 1.1, it does not suffice for f to cross 1/2: it must be *regular*, meaning it must cross by an appreciable distance, and the crossings must be evenly spaced. (It is worth highlighting that the ReLU easily gives rise to a regular f.) However, to merely show that f and g give very different classifiers $\tilde{f}$ and $\tilde{g}$ over an arbitrary measure (as in part of Theorem 3.12), no additional regularity is needed.

Proof (of Lemma 3.1) Let $\mathcal{I}_f$ and $\mathcal{I}_g$ respectively denote the sets of intervals corresponding to $\tilde{f}$ and $\tilde{g}$, and set $s_f := \mathrm{Cr}(f) = |\mathcal{I}_f|$ and $s_g := \mathrm{Cr}(g) = |\mathcal{I}_g|$.

For every $J \in \mathcal{I}_g$, set $X_J := \{U \in \mathcal{I}_f : U \subseteq J\}$. Fixing any $J \in \mathcal{I}_g$, since $\tilde{g}$ is constant on J whereas $\tilde{f}$ alternates, the number of elements in X_J where $\tilde{g}$ disagrees everywhere with $\tilde{f}$ is $|X_J|/2$ when $|X_J|$ is even and at least $(|X_J|-1)/2$ when $|X_J|$ is odd, thus at least $(|X_J|-1)/2$ in general. As such,

$$\frac{1}{s_f}\sum_{U\in\mathcal{I}_f} \mathbf{1}[\forall x \in U \centerdot \tilde{f}(x) \neq \tilde{g}(x)] \geq \frac{1}{s_f}\sum_{J\in\mathcal{I}_g}\sum_{U\in X_J} \mathbf{1}[\forall x \in U \centerdot \tilde{f}(x) \neq \tilde{g}(x)] \geq \frac{1}{s_f}\sum_{J\in\mathcal{I}_g} \frac{|X_J|-1}{2}. \tag{3.1}$$

To control this expression, note that every X_J is disjoint, however $X := \cup_{J\in\mathcal{I}_j} X_j$ can be smaller than $\mathcal{I}_f$: in particular, it misses intervals $U \in \mathcal{I}_f$ whose interior intersects with the boundary of an interval in $\mathcal{I}_g$. Since there are at most $s_g - 1$ such boundaries,

$$s_f = |\mathcal{I}_f| \leq s_g - 1 + |X| \leq s_g + \sum_{J\in\mathcal{I}_g} |X_J|,$$

which rearranges to gives $\sum_{J\in\mathcal{I}_g} |X_J| \geq s_f - s_g$. Combining this with eq. (3.1),

$$\frac{1}{s_f}\sum_{U\in\mathcal{I}_f} \mathbf{1}[\forall x \in U \centerdot \tilde{f}(x) \neq \tilde{g}(x)] \geq \frac{1}{2s_f}(s_f - s_g - s_g) = \frac{1}{2}\left(1 - \frac{2s_g}{s_f}\right).$$

■

3.2. Few layers, few oscillations

As in the preceding section, oscillations of a function f will be counted via the crossing number $\mathrm{Cr}(f)$. Since $\mathrm{Cr}(\cdot)$ only handles univariate functions, the multivariate case is handled by first choosing an affine map $h : \mathbb{R} \to \mathbb{R}^d$ (meaning $h(z) = az + b$) and considering $\mathrm{Cr}(f \circ h)$.

Before giving the central upper bounds and sketching their proofs, notice by analogy to polynomials how compositions and additions vary in their impact upon oscillations. By adding together two polynomials, the resulting polynomial has at most twice as many terms and does not exceed the maximum degree of either polynomial. On the other hand, composing polynomials, the result has the product of the degrees and can have more than the product of the terms. As both of these can impact the number of roots or crossings (e.g., by the Bezout Theorem or Descartes' Rule of Signs), composition wins the race to higher oscillations.

Lemma 3.2 *Let $h : \mathbb{R} \to \mathbb{R}^d$ be affine.*

1. *Suppose* $f \in \mathcal{N}_d((m_i, t_i, \alpha_i, \beta_i)_{i=1}^l)$ *with* $\min_i \min\{\alpha_i, \beta_i\} \geq 1$. *Setting* $\alpha := \max_i \alpha_i, \beta := \max_i \beta_i$, $t := \max_i t_i$, $m := \sum_i m_i$, *then* $\mathrm{Cr}(f \circ h) \leq 2(2tm\alpha/l)^l \beta^{l^2}$.
2. *Let* k*-dt* $f : \mathbb{R}^d \to \mathbb{R}$ *and* (t, k)*-bdt* $g : \mathbb{R}^d \to \mathbb{R}$ *be given. Then* $\mathrm{Cr}(f \circ h) \leq k$ *and* $\mathrm{Cr}(g \circ h) \leq 2tk$.

Lemma 3.2 shows the key tradeoff: the number of layers is in the exponent, while the number of nodes is in the base.

Rather than directly controlling $\mathrm{Cr}(f \circ h)$, the proofs will first show $f \circ h$ is (t, α)-poly, which immediately bounds $\mathrm{Cr}(f \circ h)$ as follows.

Lemma 3.3 *If* $f : \mathbb{R} \to \mathbb{R}$ *is* (t, α)*-poly, then* $\mathrm{Cr}(f) \leq t(1 + \alpha)$.

Proof The polynomial in each piece has at most α roots, which thus divides each piece into $\leq 1 + \alpha$ further pieces within which $\tilde{f}$ is constant. ■

A second technical lemma is needed to reason about combinations of partitions defined by (t, α, β)-sa and (t, α)-poly functions.

Lemma 3.4 *Let* k *partitions* $(A_i)_{i=1}^k$ *of* $\mathbb{R}$ *each into at most* t *intervals be given, and set* $A := \cup_i A_i$. *Then there exists a partition* B *of* $\mathbb{R}$ *of size at most* kt *so that every interval expressible as a union of intersections of elements of* A *is a union of elements of* B.

The proof is somewhat painful owing to the fact that there is no convention on the structure of the intervals in the partitions, namely which ends are closed and which are open, and is thus deferred to Appendix A. The principle of the proof is elementary, and is depicted at right: given a collection of partitions, an intersection of constituent intervals must share endpoints with intervals in in the intersection, thus the total number of intervals bounds the total number of possible intersections. Arguably, this failure to increase complexity in the face of arbitrary intersections is why semi-algebraic gates do not care about the number of terms in their definition.

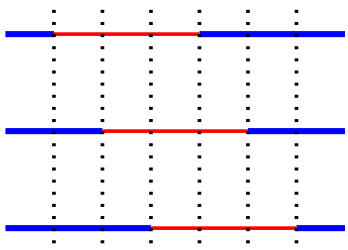

Figure 2: Three partitions.

Recall that (t, α, β)-sa means there is a set of t polynomials of degree at most α which form the regions defining the function by intersecting simpler regions $x \mapsto \mathbf{1}[q(x) \geq 0]$ and $x \mapsto \mathbf{1}[q(x) < 0]$. As such, in order to analyze semi-algebraic gates composed with piecewise polynomial gates, consider first the behavior of these predicate polynomials.

Lemma 3.5 *Suppose* $f : \mathbb{R}^k \to \mathbb{R}$ *is polynomial with degree* $\leq \alpha$ *and* $(g_i)_{i=1}^k$ *are each* (t, γ)*-poly. Then* $h(x) := f(g_1(x), \ldots, g_k(x))$ *is* $(tk, \alpha\gamma)$*-poly, and the partition defining* h *is a refinement of the partitions for each* g_i *(in particular, each* g_i *is a fixed polynomial (of degree* $\leq \gamma$*) within the* $\leq tk$ *pieces defining* h*).*

Proof By Lemma 3.4, there exists a partition of $\mathbb{R}$ into $\leq tk$ intervals which refines the partitions defining each g_i. Since f is a polynomial with degree $\leq \alpha$, then within each of these intervals, its composition with $(g_1, \ldots, g_k)$ gives a polynomial of degree $\leq \alpha\gamma$. ■

This gives the following complexity bound for composing (s, α, β)-sa and (t, γ)-poly gates.

Lemma 3.6 *Suppose $f : \mathbb{R}^k \to \mathbb{R}$ is (s, α, β)-sa and $(g_1, \ldots, g_k)$ are (t, γ)-poly. Then $h(x) := f(g_1(x), \ldots, g_k(x))$ is $(stk(1 + \alpha\gamma), \beta\gamma)$-poly.*

Proof By definition, f is polynomial in regions defined by intersections of the predicates $U_i(x) = \mathbf{1}[q_i(x) \geq 0]$ and $L_i(x) = \mathbf{1}[q_i(x) < 0]$. By Lemma 3.5, $q_i(g_1, \ldots, g_k)$ is $(tk, \alpha\gamma)$-poly, thus U_i and L_i together define a partition of $\mathbb{R}$ which has $\mathrm{Cr}(x \mapsto q_i(g_1(x), \ldots, g_k(x)))$ pieces, which by Lemma 3.3 has cardinality at most $tk(1+\alpha\gamma)$ and refines the partitions for each g_i. By Lemma 3.4, these partitions across all predicate polynomials $(q_i)_{i=1}^s$ can be refined into a single partition of size $\leq stk(1 + \alpha\gamma)$, and which thus also refines the partitions defined by $(g_1, \ldots, g_k)$. Thanks to these refinements, h over any element U of this final partition is a fixed polynomial $p_U(g_1, \ldots, g_k)$ of degree $\leq \beta\gamma$, meaning h is $(stk(1 + \alpha\gamma), \beta\gamma)$-poly. ■

The proof of Lemma 3.2 now follows by Lemma 3.6. In particular, for semi-algebraic networks, the proof is an induction over layers, establishing node j is (t_j, α_j)-poly (for appropriate (t_j, α_j)).

3.3. Many layers, many oscillations

The idea behind this construction is as follows. Consider any continuous function $f : [0, 1] \to [0, 1]$ which is a generalization of a triangle wave with a single peak: $f(0) = f(1) = 0$, and there is some $a \in (0, 1)$ with $f(a) = 1$, and additionally f strictly increases along $[0, a]$ and strictly decreases along $[a, 1]$.

Now consider the effect of the composition $f \circ f = f^2$. Along $[0, a]$, this is a stretched copy of f, since $f(f(a)) = f(1) = 0 = f(0) = f(f(0))$ and moreover f is a bijection between $[0, a]$ and $[0, 1]$ (when restricted to $[0, a]$). The same reasoning applies to f^2 along $[a, 1]$, meaning f^2 is a function with two peaks. Iterating this argument implies f^k is a function with 2^{k-1} peaks; the following definition and lemmas formalize this reasoning.

Definition 3.7 *f is $(t, [a, b])$-*triangle *when it is continuous along $[a, b]$, and $[a, b]$ may be divided into $2t$ intervals $[a_i, a_{i+1}]$ with $a_1 = a$ and $a_{2t+1} = b$, $f(a_i) = f(a_{i+2})$ whenever $1 \leq i \leq 2t - 1$, $f(a_1) = 0$, $f(a_2) = 1$, f is strictly increasing along odd-numbered intervals (those starting from a_i with i odd), and strictly decreasing along even-numbered intervals.*

Lemma 3.8 *If f is $(s, [0, 1])$-triangle and g is $(t, [0, 1])$-triangle, then $f \circ g$ is $(2st, [0, 1])$-triangle.*

Proof Since $g([0, 1]) = [0, 1]$ and f and g are continuous along $[0, 1]$, then $f \circ g$ is continuous along $[0, 1]$. In the remaining analysis, let $(a_1, \ldots, a_{2s+1})$ and $(c_1, \ldots, c_{2t+1})$ respectively denote the interval boundaries for f and g.

Now consider any interval $[c_j, c_{j+1}]$ where j is odd, meaning the restriction $g_j : [c_j, c_{j+1}] \to [0, 1]$ of g to $[c_j, c_{j+1}]$ is strictly increasing. It will be shown that $f \circ g_j$ is $(s, [c_j, c_{j+1}])$-triangle, and an analogous proof holds for the strictly decreasing restriction $g_{j+1} : [c_{j+1}, c_{j+2}] \to [0, 1]$, whereby it follows that $f \circ g$ is $(2st, [0, 1])$ by considering all choices of j.

To this end, note for any $i \in \{1, \ldots, 2s + 1\}$ that $g_j^{-1}(a_i)$ exists and is unique, thus set $a'_i := g_j^{-1}(a_i)$. By this choice, for odd i it holds that $f(g_j(a'_i)) = f(g_j(g_j^{-1}(a_i))) = f(a_i) = f(a_1) = 0$ and $f \circ g_j$ is strictly increasing along $[a'_i, a'_{i+1}]$ (since g_j is strictly increasing everywhere and f is strictly increasing along $[g_j(a'_i), g_j(a'_{i+1})] = [a_i, a_{i+1}]$), and similarly even i has $f(g_j(a'_i)) = f(a_2) = 1$ and $f \circ g_j$ is strictly decreasing along $[a'_i, a'_{i+1}]$. ■

Corollary 3.9 *If $f \in \mathcal{N}_1(m, l, t, \alpha, \beta)$ is $(t, [0,1])$-triangle with p distinct parameters, then $f^k \in \mathcal{N}_1(m, kl, t, \alpha, \beta)$ is $(2^{k-1}t^k, [0,1])$-triangle with p distinct parameters and* $\mathrm{Cr}(f^k) = (2t)^k + 1$.

Proof It suffices to perform $k-1$ applications of Lemma 3.8. ■

Next, note the following examples of triangle functions.

Lemma 3.10 *The following functions are $(1, [0,1])$-triangle.*

1. $f(z) := \sigma_{\mathrm{R}}(2\sigma_{\mathrm{R}}(z) - 4\sigma_{\mathrm{R}}(z - 1/2)) \in \mathcal{N}_1(2, 1, 1, 1, 1)$.
2. $g(z) := \min\{\sigma_{\mathrm{R}}(2z), \sigma_{\mathrm{R}}(2 - 2z)\} \in \mathcal{N}_1(2, 1, 2, 1, 1)$.
3. $h(z) := 4z(1-z) \in \mathcal{N}_1(1, 1, 0, 2, 0)$. *Cf. Schmitt (2000).*

Lastly, consider the first example $f(z) = \sigma_{\mathrm{R}}(2\sigma_{\mathrm{R}}(z) - 4(\sigma_{\mathrm{R}}(z - 1/2))) = \min\{\sigma_{\mathrm{R}}(2z), \sigma_{\mathrm{R}}(2 - 2z)\}$, whose graph linearly interpolates (in $\mathbb{R}^2$) between $(0,0)$, $(1/2, 1)$, and $(1, 0)$. Consequently, $f \circ f$ along $[0, 1/2]$ linear interpolates between $(0,0)$, $(1/4, 1)$, and $(1/2, 1)$, and $f \circ f$ is analogous on $[1/2, 1]$, meaning it has produced two copies of f and then shrunken them horizontally by a factor of 2. This process repeats, meaning f^k has 2^{k-1} copies of f, and grants the regularity needed to use the Lebesgue measure in Theorem 1.1.

Lemma 3.11 *Set $f(z) := \sigma_{\mathrm{R}}(2\sigma_{\mathrm{R}}(z) - 4\sigma_{\mathrm{R}}(z - 1/2)) \in \mathcal{N}_1(2, 1, 1, 1, 1)$ (cf. Lemma 3.10). Let real $z \in [0,1]$ and positive integer k be given, and choose the unique nonnegative integer $i_k \in \{0, \dots, 2^{k-1}\}$ and real $z_k \in [0, 1)$ so that $z = (i_k + z_k)2^{1-k}$. Then*

$$f^k(z) = \begin{cases} 2z_k & \text{when } 0 \le z_k \le 1/2, \\ 2(1 - z_k) & \text{when } 1/2 < z_k < 1. \end{cases}$$

3.4. Proof of Theorem 1.1

The proof of Theorem 1.1 now follows: Lemma 3.11 shows that a many-layered ReLU network can give rise to a highly oscillatory and regular function f^k, Lemma 3.2 shows that few-layered networks and (boosted) decision trees give rise to functions with few oscillations, and lastly Lemma 3.1 shows how to combine these into an inapproximability result.

In this last piece, the proof averages over the possible offsets $y \in \mathbb{R}^{d-1}$ and considers univariate problems after composing networks with the affine map $h_y(z) := (z, y)$. In this way, the result carries some resemblance to the random projection technique used in depth hierarchy theorems for boolean functions (Håstad, 1986; Rossman et al., 2015), as well as earlier techniques on complexities of multivariate sets (Vitushkin, 1955, 1959), albeit in an extremely primitive form (considering variations along only one dimension).

Proof (of Theorem 1.1) Set $h(z) := \sigma_{\mathrm{R}}(2\sigma_{\mathrm{R}}(z) - 4\sigma_{\mathrm{R}}(z - 1/2))$ (cf. Lemma 3.10), and define $f_0(z) := h^{k^3+4}(z)$ and $f : \mathbb{R}^d \to \mathbb{R}$ as $f(x) = f_0(x_1)$. Let $\mathcal{I}_f$ denote the pieces of $\tilde{f}_0$, meaning $|\mathcal{I}_f| = \mathrm{Cr}(f_0)$, and Corollary 3.9 grants $\mathrm{Cr}(f_0) = 2^{k^3+4} + 1$. Moreover, by Lemma 3.11, for any $U \in \mathcal{I}_f$, $f_0 - 1/2$ is a triangle with height 1/2 and base either 2^{-k-1} (when $0 \in U$ or $1 \in U$) or

2^{-k}, whereby $\int_U |f_0(x) - 1/2|dx \geq 2^{-k-1}/4 \geq |\mathcal{I}_f|/16$ (which has thus made use of the special regularity of h).

Now for any $y \in \mathbb{R}^{d-1}$ define the map $p_y : \mathbb{R} \to \mathbb{R}^d$ as $p_y(z) := (z, y)$. If g is a semi-algebraic network with $\leq k$ layers and $m \leq 2^k/(t\alpha\beta)$ total nodes, then Lemma 3.2 grants $\mathrm{Cr}(g \circ p_y) \leq 2(2tm\alpha/k)^k \beta^{k^2} \leq 4(tm\alpha\beta)^{k^2} \leq 2^{k^3+2}$. Otherwise, g is $(t, 2^{k^3}/t)$-bdt, whereby Lemma 3.2 gives $\mathrm{Cr}(g \circ p_y) \leq 2t2^{k^3}/t \leq 2^{k^3+2}$ once again.

By Lemma 3.1, for any $y \in \mathbb{R}^{d-1}$, $\mathrm{Cr}(f \circ p_y) = \mathrm{Cr}(f_0)$, and

$$\begin{aligned}
\int_{[0,1]} |f(p_y(z)) - g(p_y(z))|dz &= \sum_{U \in \mathcal{I}_f} \int_U |(f \circ p_y)(z) - (g \circ p_y)(z)|dz \\
&\geq \sum_{U \in \mathcal{I}_f} \int_U |(f \circ p_y)(z) - 1/2| \mathbf{1}[\forall z \in U \centerdot \widetilde{(f \circ p_y)}(z) \neq \widetilde{(g \circ p_y)}(z)]dz \\
&\geq \frac{1}{16|\mathcal{I}_f|} \sum_{U \in \mathcal{I}_f} \mathbf{1}[\forall z \in U \centerdot \widetilde{(f \circ p_y)}(z) \neq \widetilde{(g \circ p_y)}(z)]dz \\
&\geq \frac{1}{32}\left(1 - \frac{2\mathrm{Cr}(g \circ p_y)}{\mathrm{Cr}(f \circ p_y)}\right) \geq \frac{1}{32}\left(1 - \frac{2(2^{k^3+2})}{2^{k^3+4}}\right) \geq \frac{1}{64}.
\end{aligned}$$

To finish,

$$\int_{[0,1]^d} |f(x) - g(x)|dx = \int_{[0,1]^{d-1}} \int_{[0,1]} |(f \circ p_y)(z) - (g \circ p_y)(z)|dzdy \geq \frac{1}{64}.$$

■

Using nearly the same proof, but giving up on continuous uniform measure, it is possible to handle other distances and more flexible target functions.

Theorem 3.12 *Let integer $k \geq 1$ and function $f : \mathbb{R} \to \mathbb{R}$ be given where f is $(1, [0,1])$-triangle, and define $h : \mathbb{R}^d \to \mathbb{R}$ as $h(x) := f^k(x_1)$. For every $y \in \mathbb{R}^{d-1}$, define the affine function $p_y(z) := (z, y)$. Then there exist Borel probability measures μ and ν over $[0,1]^d$ where ν is discrete uniform on 2^k+1 points and μ is continuous and positive on exactly $[0,1]^d$ so that every $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathrm{Cr}(g \circ p_y) \leq 2^{k-2}$ for every $y \in \mathbb{R}^{d-1}$ satisfies*

$$\int |h - g|d\mu \geq \frac{1}{32}, \qquad \int |\tilde{h} - \tilde{g}|d\mu \geq \frac{1}{8}, \qquad \int |h - g|d\nu \geq \frac{1}{8}, \qquad \int |\tilde{h} - \tilde{g}|d\nu \geq \frac{1}{4}.$$

4. Limitations of depth

Theorem 3.12 can be taken to say: there exists a labeling of $\Theta(2^{k^3})$ points which is realizable by a network of depth and size $\Theta(k^3)$, but can not be approximated by networks with depth k and size $o(2^k)$. On the other hand, this section will sketch the proof of Theorem 1.2, which implies that these $\Theta(k^3)$ depth networks realize relatively few different labellings. The proof is a quick consequence of the VC dimension of semi-algebraic networks (cf. Lemma 1.3) and the following fact, where $\mathrm{Sh}(\cdot)$ is used to denote the *growth function* (Anthony and Bartlett, 1999, Chapter 3).

Lemma 4.1 *Let any function class $\mathcal{F}$ and any distinct points $(x_i)_{i=1}^n$ be given. Then with probability at least $1-\delta$ over a uniform random draw of labels $(y_i)_{i=1}^n$ (with $y_i \in \{-1,+1\}$),*

$$\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}[\tilde{f}(x_i) \neq y_i] \geq \frac{1}{2}\left(1 - \sqrt{\frac{\ln(\mathrm{Sh}(\mathcal{F};n)) + \ln(1/\delta)}{2n}}\right).$$

The proof of the preceding result is similar to proofs of the Gilbert-Varshamov packing bound via Hoeffding's inequality (Duchi, 2016, Lemma 13.5). Note that a similar result was used by Warren to prove rates of approximation of continuous functions by polynomials, but without invoking Hoeffding's inequality (Warren, 1968, Theorem 7).

The remaining task is to control the VC dimension of semi-algebraic networks. To this end, note the following generalization of Lemma 1.3, which further provides that semi-algebraic networks compute functions which are polynomial when restricted to certain polynomial regions.

Lemma 4.2 *Let neural network graph $\mathfrak{G}$ be given with $\leq p$ parameters, $\leq l$ layers, and $\leq m$ total nodes, and suppose every gate is (t,α,β)-sa. Then*

$$\mathrm{VC}(\mathcal{N}(\mathfrak{G})) \leq 6p(l+1)\big(\ln(2p(l+1)) + \ln(8emt\alpha) + l\ln(\beta)\big).$$

Additionally, given any $n \geq p$ data points, there exists a partition $\mathcal{S}$ of $\mathbb{R}^p$ where each $S \in \mathcal{S}$ is an intersection of predicates $\mathbf{1}[q \diamond 0]$ with $\diamond \in \{<,\geq\}$ and q has degree $\leq \alpha\beta^{l-1}$, such that $F_{\mathfrak{G}}(x_i,\cdot)$ restricted to each $S \in \mathcal{S}$ is a fixed polynomial of degree $\leq \beta^l$ for every example x_i, with $|\mathcal{S}| \leq \left(8enmt\alpha\beta^l\right)^{pl}$ and $\mathrm{Sh}(\mathcal{N}(\mathfrak{G});n) \leq \left(8enmt\alpha\beta^l\right)^{p(l+1)}$

The proof follows the same basic structure of the VC bound for networks with piecewise polynomial activation functions (Anthony and Bartlett, 1999, Theorem 8.8). The slightly modified proof here is also very similar to the proof of Lemma 3.2, performing an induction up through the layers of the network, arguing that each node computes a polynomial after restricting attention to some range of parameters. The proof of Lemma 4.2 manages to be multivariate (unlike Lemma 3.2), though this requires arguments due to Warren (1968) which are significantly more complicated than those of Lemma 3.2 (without leading to a strengthening of Theorem 1.1).

One minor departure from the VC dimension proof of piecewise polynomial networks (cf. (Anthony and Bartlett, 1999, Theorem 8.8)) is the following lemma, which is used to track the number of regions with the more complicated semi-algebraic networks. Despite this generalization, the VC dimension bound is basically the same as for piecewise polynomial networks.

Lemma 4.3 *Let a set of polynomials $\mathcal{Q}$ be given where each $\mathcal{Q} \ni q : \mathbb{R}^p \to \mathbb{R}$ has degree $\leq \alpha$. Define an initial family $\mathcal{S}_0$ of subsets of $\mathbb{R}^p$ as $\mathcal{S}_0 := \big\{\{a \in \mathbb{R}^p : q(a) \diamond 0\} \;:\; q \in \mathcal{Q}, \diamond \in \{<,\geq\}\big\}$. Then the collection $\mathcal{S}$ of all nonempty intersections of elements of $\mathcal{S}_0$ satisfies $|\mathcal{S}| \leq 2\left(\frac{4e|\mathcal{Q}|\alpha}{p}\right)^p$.*

5. Bibliographic notes and open problems

Arguably the first approximation theorem of a big class by a smaller one is the Weierstrass Approximation Theorem, which states that polynomials uniformly approximate continuous functions over compact sets (Weierstrass, 1885). Refining this, Kolmogorov (1936) gave a bound on how well subspaces of functions can approximate continuous functions, and Vitushkin (1955, 1959) showed

a similar bound for approximation by polynomials in terms of the polynomial degrees, dimension, and modulus of continuity of the target function. Warren (1968) then gave an alternate proof and generalization of this result, in the process effectively proving the VC dimension of polynomials (developing tools still used to prove the VC dimension of neural networks (Anthony and Bartlett, 1999, Chapters 7-8)), and producing an analog to Theorem 1.2 for polynomials.

The preceding results, however, focused on separating large classes (e.g., continuous functions of bounded modulus) from small classes (polynomials of bounded degree). Aiming to refine this, depth hierarchy theorems in circuit complexity separated circuits of a certain depth from circuits of a slightly smaller depth. As mentioned in Section 1, the seminal result here is due to Håstad (1986). For architectures closer to neural networks, *sum-product networks* (summation and product nodes) have been analyzed by Bengio and Delalleau (2011) and more recently Martens and Medabalimi (2015), and networks of linear threshold functions in 2 and 3 layers by Kane and Williams (2015); note that both polynomial gates (as in sum-product networks) and linear threshold gates are semi-algebraic gates. Most closely to the present work (excluding (Telgarsky, 2015), which is a vastly simplified account), Eldan and Shamir (2015) analyze 2- and 3-layer networks with general activation functions composed with affine mappings, showing separations which are exponential in the input dimension. Due to this result and also recent advances in circuit complexity (Rossman et al., 2015), it is natural to suppose Theorem 1.1 can be strengthened to separating k and $k+1$ layer networks when dimension d is large; however, none of the earlier works give a tight sense of the behavior as $d \downarrow 1$.

The triangle wave target functions considered here (e.g., cf. Lemma 3.10) have appeared in various forms throughout the literature. General properties of piecewise affine highly oscillating functions were investigated by Szymanski and McCane (2014) and Montúfar et al. (2014). Also, Schmitt (2000) investigated the map $z \mapsto 4z(1-z)$ (as in Lemma 3.10) to show that sigmoidal networks can not approximate high degree polynomials via an analysis similar to the one here, however looseness in the VC bounds for sigmoidal networks prevented exponential separations and depth hierarchies.

A tantalizing direction for future work is to characterize not just one difficult function (e.g., triangle functions as in Lemma 3.10), but many, or even all functions which are not well-approximated by smaller depths. Arguably, this direction could have value in machine learning, as discovery of such underlying structure could lead to algorithms to recover it. As a trivial example of the sort of structure which could arise, considering the following proposition, stating that any symmetric signal may be repeated by pre-composing it with the ReLU triangle function.

Proposition 5.1 *Set* $f(z) := \sigma_{\mathrm{R}}(2\sigma_{\mathrm{R}}(z) - 4\sigma_{\mathrm{R}}(z-1/2))$ *(cf. Lemma 3.10), and let any* $g : [0,1] \to [0,1]$ *be given with* $g(z) = g(1-z)$*. Then* $h := g \circ f^k$ *satisfies* $h(x) = h(x + i2^k) = g(x2^k)$ *for every real* $x \in [0, 2^{-k}]$ *and integer* $i \in \{0, \ldots, 2^{-k} - 1\}$*; in other words,* h *is* 2^k *repetitions of* g *with graph scaled horizontally and uniformly to fit within* $[0,1]^2$.

Acknowledgments

The author is indebted to Joshua Zahl for help navigating semi-algebraic geometry and for a simplification of the multivariate case in Theorem 1.1, and to Rastislav Telgársky for an introduction to this general topic via Kolmogorov's Superposition Theorem (Kolmogorov, 1957). The author further thanks Jacob Abernethy, Peter Bartlett, Sébastien Bubeck, and Alex Kulesza for valuable discussions.

References

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

Yoshua Bengio and Olivier Delalleau. Shallow vs. deep sum-product networks. In *NIPS*, 2011.

Jacek Bochnak, Michal Coste, and Marie-Françoise Roy. *Real Algebraic Geometry*. Springer, 1998.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. pages 161–168, 2006.

John Duchi. Statistics 311/electrical engineering 377: Information theory and statistics. Stanford University, 2016.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. 2015. arXiv:1512.03965 [cs.LG].

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

Johan Håstad. *Computational Limitations of Small Depth Circuits*. PhD thesis, Massachusetts Institute of Technology, 1986.

Daniel Kane and Ryan Williams. Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. 2015. arXiv:1511.07860v1 [cs.CC].

Andrei Kolmogorov. Über die beste annäherung von funktionen einer gegebenen funktionenklasse. *Annals of Mathematics*, 37(1):107–110, 1936.

Andrey Nikolaevich Kolmogorov. On the representation of continuous functions of several variables by superpositions of continuous functions of one variable and addition. 114:953–956, 1957.

Alex Krizhevsky, Ilya Sutskever, and Geoffery Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

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.

James Martens and Venkatesh Medabalimi. On the expressive efficiency of sum product networks. 2015. arXiv:1411.7717v3 [cs.LG].

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *NIPS*, 2014.

Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. In *UAI 2011*, pages 337–346, 2011.

Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. An average-case depth hierarchy theorem for boolean circuits. In *FOCS*, 2015.

Michael Schmitt. Lower bounds on the complexity of approximating continuous functions by sigmoidal neural networks. In *NIPS*, 2000.

Lech Szymanski and Brendan McCane. Deep networks are effective encoders of periodicity. *IEEE Transactions on Neural Networks and Learning Systems*, 25(10):1816–1827, 2014.

Matus Telgarsky. Representation benefits of deep feedforward networks. 2015. `arXiv:1509.08101v2 [cs.LG]`.

Anatoli Vitushkin. On multidimensional variations. *GITTL*, 1955. In Russian.

Anatoli Vitushkin. Estimation of the complexity of the tabulation problem. *Fizmatgiz.*, 1959. In Russian.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Karl Weierstrass. Über die analytische darstellbarkeit sogenannter willkürlicher functionen einer reellen veränderlichen. *Sitzungsberichte der Akademie zu Berlin*, pages 633–639, 789–805, 1885.

Appendix A. Deferred proofs

This appendix collects various proofs omitted from the main text.

A.1. Deferred proofs from Section 2

The following mechanical proof shows that standard piecewise polynomial gates, maximization/minimization gates, and decision trees are all semi-algebraic gates.

Proof (of Lemma 2.3)

1. To start, since $\sigma : \mathbb{R} \to \mathbb{R}$ is piecewise polynomial, $\sigma \circ q$ can be written

$$\sigma(q(z)) := p_1(q(z))\mathbf{1}[q(z) \diamond_1 b_1] + \sum_{i=2}^{t-1} p_i(q(z))\mathbf{1}[-q(z) *_{i-1} -b_{i-1}]\mathbf{1}[q(z) \diamond_i b_i] + p_t(q(z))\mathbf{1}[-q(z) *_t -b_t]$$

   where for each $i \in [t]$, p_i has degree $\leq \beta$, $\diamond_i \in \{<, \leq\}$, $*_i = \text{“} < \text{”}$ when $\diamond_i = \text{“} \leq \text{”}$ and otherwise $*_i = \text{“} \leq \text{”}$, and $b_i \in \mathbb{R}$. As such, setting $q_i(z) := q(z) - b_1$ whenever $\diamond_i = \text{“} < \text{”}$ and $q_i(z) := b_i - q(z)$ otherwise, it follows that $\sigma \circ q$ is $(t, \alpha, \alpha\beta)$-sa.

2. Since $\min_{i\in[r]} x_i = -\max_{i\in[r]} -x_i$, it suffices to handle the maximum case, which has the form

$$\phi_{\max}(v) = \sum_{i=1}^{d} p_i(v) \left(\prod_{j<i} \mathbf{1}[p_i(v) > p_j(v)] \right) \left(\prod_{j>i} \mathbf{1}[p_i(v) \geq p_j(v)] \right).$$

   Constructing polynomials $q_{i,j} = p_j - p_i$ when $j < i$ and $q_{i,j} = p_i - p_j$ when $j > i$, it follows that $\phi_{\max}$ is $(r(r-1), \alpha, \alpha)$-sa.

3. First consider a k-dt f, wherein the proof follows by induction on tree size. In the base case $k = 1$, f is constant. Otherwise, there exist functions f_l and f_r which are respectively l- and r-dt with $l + r < k$, and additionally an affine function q_f so that

$$\begin{aligned} f(x) &= f_l(x)\mathbf{1}[q_f(x) < 0] + f_r(x)\mathbf{1}[q_f(x) \geq 0] \\ &= \sum_{j=1}^{m_l} p_j^{(l)}(v)\mathbf{1}[q_f(x) < 0] \left(\prod_{i \in L_j^{(l)}} \mathbf{1}[q_i^{(l)}(v) < 0] \right) \left(\prod_{i \in U_j^{(l)}} \mathbf{1}[q_i^{(l)}(v) \geq 0] \right) \\ &\quad + \sum_{j=1}^{m_r} p_j^{(r)}(v)\mathbf{1}[q_f(x) \geq 0] \left(\prod_{i \in L_j^{(r)}} \mathbf{1}[q_i^{(r)}(v) < 0] \right) \left(\prod_{i \in U_j^{(r)}} \mathbf{1}[q_i^{(r)}(v) \geq 0] \right). \end{aligned}$$

   where the last step expanded the semi-algebraic forms of f_l and f_r. As such, by combining the sets of predicate polynomials for f_l and f_r together with $\{q_f\}$ (where the former two have cardinalities $\leq l$ and $\leq r$ by the inductive hypothesis), and unioning together the triples for f_l and f_r but extending the triples to include $\mathbf{1}[q_f < 0]$ for triples in f_l and $\mathbf{1}[q_f \geq 0]$ for triples in f_r, it follows by construction that f is $(k, 1, 0)$-semi-algebraic.

   Now consider a (t, k)-bdt g. By the preceding expansion, each individual tree f_i is $(k, 1, 0)$-sa, thus their sum is $(tk, 1, 0)$ by unioning together the sets of polynomials, triples, and adding together the expansions.

■

A.2. Deferred proofs from Section 3

The first proof shows that a collection of partitions may be refined into a single partition whose size is at most the total number of intervals across all partitions. As discussed in the text, while the proof has a simple idea (one need only consider boundaries of intervals across all partitions), it is somewhat painful since there is not consistent rule for whether specific endpoints endpoints of intervals are open or closed.

Proof (of Lemma 3.4) If $k = 1$, then the result follows with $B = A = A_1$ (since all intersections are empty), thus suppose $k \geq 2$. Let $\{a_1, \ldots, a_q\}$ denote the set of distinct boundaries of intervals of A, and iteratively construct the partition B as follows, where the construction will maintain that B_j is a partition whose boundary points are $\{a_1, \ldots a_j\}$. For the base case, set $B_0 := \{\mathbb{R}\}$. Thereafter, for every $i \in [q]$, consider boundary point a_i; since the boundary points are distinct, there must exist a single interval $U \in B_{i-1}$ with $a_i \in U$. B_i will be formed from B_{i-1} by refining U in one of the following two ways.

- Consider the case that each partition A_l which contains the boundary point a_i has exactly two intervals meeting at a_i and moreover the closedness properties are the same, meaning either a_i is contained in the interval which ends at a_i, or it is contained in the interval which starts at a_i. In this case, partition U into two intervals so that the treatment of the boundary is the same as those A_l's with a boundary at a_i.

- Otherwise, it is either the case that some A_l have a_i contained in the interval ending at a_i whereas others have it contained in the interval starting at a_i, or simply some A_l have three intervals meeting at a_i: namely, the singleton interval $[a_l, a_l]$ as well as two intervals not containing a_l. In this case, partition U into three intervals: one ending at a_i (but not containing it), the singleton interval $[a_i, a_i]$, and an interval starting at a_i (but not containing it).

(These cases may also be described in a unified way: consider all intervals of A which have a_i as an endpoint, extend such intervals of positive length to have infinite length while keeping endpoint a_i and the side it falls on, and then refine U by intersecting it with all of these intervals, which as above results in either 2 or 3 intervals.)

Note that the construction never introduces more intervals at a boundary point than exist in A, thus $|B| \leq |A| = kt$.

It remains to be shown that a union of intersections of elements of A is a union of elements of B. Note that it suffices to show that intersections of elements of A are unions of elements of B, since thereafter these encodings can be used to express unions of intersections of A as unions of B. As such, consider any intersection U of elements of A; there is nothing to show if U is empty, thus suppose it is nonempty. In this case, it must also be an interval (e.g., since intersections of convex sets are convex), and its endpoints must coincide with endpoints of A. Moreover, if the left endpoint of U is open, then U must be formed from an intersection which includes an interval with the same open left endpoint, thus there exists such an interval in A, and by the above construction of B, there also exists an interval with such an open left endpoint in B; the same argument similarly handles the case of closed left endpoints, as well as open and closed right endpoints, namely giving elements in B which match these traits. Let a_r and a_s denote these endpoints. By the above construction of B, intervals with endpoints $\{a_j, a_{j+1}\}$ for $j \in \{r, \ldots, s-1\}$ will be included in B, and since B is a partition, the union of these elements will be exactly U. Since U was an arbitrary intersection of elements of A, the proof is complete. ■

Next, the tools of Section 3.2 (culminating in the composition rule for semi-algebraic gates (Lemma 3.6)) are used to show crossing number bounds on semi-algebraic networks and boosted decision trees.

Proof (of Lemma 3.2)

1. This proof first shows $f \circ h$ is $(2^i t_i \alpha_i \prod_{j \leq i-1} t_j \alpha_j \beta_j^{i-j+1} k_j, \prod_{j \leq i} \beta_j)$-poly, and then relaxes this expression and applies Lemma 3.3 to obtain the desired bound.

   First consider the case $d = 1$ and h is the identity map, thus $f \circ h = f$. For convenience, set

   $$A_i := \prod_{j \leq i} \alpha_j, \quad B_i := \prod_{j \leq i} \beta_j, \quad C_i := \prod_{j \leq i} \beta_j^{i-j+1} = \prod_{j \leq i} B_j, \quad M_i := \prod_{j \leq i} m_j, \quad T_i := \prod_{j \leq i} t_j.$$

   The proof proceeds by induction on the layers of f, showing that each node in layer i is $(2^i T_i A_i C_{i-1} M_{i-1}, B_i)$-poly.

   For convenience, first consider layer $i = 0$ of the inputs themselves: here, node i outputs the i^{th} coordinate of the input, and is thus affine and $(1, 1)$-poly. Next consider layer $i > 0$, where the inductive hypothesis grants that each node in layer $i-1$ is $(2^{i-1} T_{i-1} A_{i-1} C_{i-2} M_{i-2}, B_{i-1})$-poly. Consequently, since any node in layer i is (t_i, α_i, β_i)-sa, Lemma 3.6 grants it is also

$(2^{i-1}t_iT_{i-1}A_{i-1}C_{i-2}M_{i-2}m_{i-1}(1+\alpha_iB_{i-1}), \beta_iB_{i-1})$-poly as desired (since $1+\alpha_iB_{i-1} \leq 2\alpha_iB_{i-1}$).

Next, consider the general case $d \geq 1$ and $h : \mathbb{R} \to \mathbb{R}^d$ is an affine map. Since every coordinate of h is affine (and thus $(1,1)$-poly), composing h with every polynomial in the semi-algebraic gates of layer 1 gives a function $g \in \mathcal{N}_1((m_i, t_i, \alpha_i, \beta_i)_{i=1}^l)$ which is equal to $f \circ h$ everywhere and whose gates are of the same semi-algebraic complexity. As such, the result follows by applying the preceding analysis to g.

Lastly, the simplified terms give $f \circ h$ is $((2t\alpha)^l \beta^{l(l-1)/2} \prod_{j \leq l-1} m_j, \beta^{l(l+1)/2})$-poly. Since $\ln(\cdot)$ is strictly increasing and concave and $m_l = 1$,

$$\ln\left(\prod_{j\leq l-1} m_j\right) = \ln\left(\prod_{j\leq l} m_j\right) = \sum_{j\leq l} \ln(m_j) \leq l\ln(m/l) = \ln((m/l)^l).$$

It follows that $f \circ h$ is $((2tm\alpha/l)^l \beta^{l(l-1)/2}, \beta^{l(l+1)/2})$-poly, whereby the crossing number bound follows by Lemma 3.3.

2. Given any k-dt f, the affine function evaluated at each predicate may be composed with h to yield another affine function, thus $f \circ h : \mathbb{R} \to \mathbb{R}$ is still a k-dt, and thus $(k, 1, 0)$-sa by Lemma 2.3. As such, by Lemma 3.6 (with $g_1(z) = z$ as the identity map), $f \circ h$ is $(k, 0)$-poly. (Invoking Lemma 3.6 without massaging in h introduces a factor d.) Similarly, for a (t, k)-bdt g, $g \circ h : \mathbb{R} \to \mathbb{R}$ is another (t, k)-bdt after pushing h into the predicates of the constituent trees, thus Lemma 2.3 grants $g \circ h$ is $(tk, 1, 0)$-sa, and Lemma 3.6 grants it is $(tk(1+1), 0)$-poly. The desired crossing number bounds follow by applying Lemma 3.3.

■

Next, elementary computations verify that the three functions listed in Lemma 3.10 are indeed $(1, [0, 1])$-triangle.

Proof (of Lemma 3.10)

1-2. By inspection, $f(0) = f(1) = 0$ and $f(1/2) = 1$. Moreover, for $x \in [0, 1/2]$, $f(x) = 2x$ meaning f is increasing, and $x \in [1/2, 1]$ means $f(x) = 2(1-x)$, meaning f is decreasing. Lastly, the properties of g follow since $f = g$.

3. By inspection, $h(0) = h(1) = 0$ and $h(1/2) = 1$. Moreover h is a quadratic, thus can cross 0 at most twice, and moreover $1/2$ is the unique critical point (since g' has degree 1), thus g is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$.

■

In the case of the ReLU $(1, [0, 1])$-triangle function f given in Lemma 3.10, the exact form of f^k may be established as follows. (Recall that this refined form allows for the use of Lebesgue measure in Theorem 1.1, and also the repetition statement in Proposition 5.1.)

Proof (of Lemma 3.11) The proof proceeds by induction on the number of compositions l. For the base case $l = 1$,

$$f^1(z) = f(z) = \begin{cases} 2z & \text{when } z \in [0, 1/2], \\ 2(1-z) & \text{when } z \in (1/2, 1], \\ 0 & \text{otherwise.} \end{cases}$$

For the inductive step, first note for any $z \in [0, 1/2]$, by symmetry of f^l around 1/2 (i.e., $f^l(z) = f^l(1-z)$ by the inductive hypothesis), and by the above explicit form of f^1,

$$f^{l+1}(z) = f^l(f(z)) = f^l(2z) = f^l(1-2z) = f^l(f(1/2-z)) = f^l(f(z+1/2)) = f^{l+1}(z+1/2),$$

meaning the case $z \in (1/2, 1]$ is implied by the case $z \in [0, 1/2]$. Since the unique nonnegative integer i_{l+1} and real $z_{l+1} \in [0, 1)$ satisfy $2z = 2(i_{l+1} + z_{l+1})2^{-l-1} = (i_{l+1} + z_{l+1})2^{-l}$, the inductive hypothesis grants

$$(f^l \circ f)(z) = f^l(2z) = \begin{cases} 2z_{l+1} & \text{when } 0 \le z_{l+1} \le 1/2, \\ 2(1-z_{l+1}) & \text{when } 1/2 < z_{l+1} < 1, \end{cases}$$

which completes the proof. ■

The proof of the slightly more general form of Theorem 1.1 is as follows; it does not quite imply Theorem 1.1, since the constructed measure is not the Lebesgue measure even for the ReLU-based $(1, [0, 1])$-triangle function from Lemma 3.10.

Proof (of Theorem 3.12) First note some general properties of f^k. By Corollary 3.9, f^k is $(2^{k-1}, [0, 1])$-triangle, which means there exist $s := 2^k + 1$ points $(z_i)_{i=1}^s$ so that $f^k(z_i) = \mathbf{1}[i \text{ is odd}]$, and moreover f^k is continuous and equal to $1/2$ at exactly 2^k points (by the strict increasing/decreasing part of the triangle wave definition), which is a finite set of points and thus has Lebesgue measure zero. Taking $p_y : \mathbb{R} \to \mathbb{R}^d$ to be the map $p_y(z) = (z, y)$ where $y \in \mathbb{R}^{d-1}$, then $(h \circ p_y)(z) = \widetilde{h((z, y))} = f^k(z)$, thus letting $\mathcal{I}$ denote the 2^k pieces within which $\widetilde{f^k}$ is constant, it follows that $\widetilde{h \circ p_y}$ is constant within the same set of pieces and thus $\mathrm{Cr}(h \circ p_y) = s$.

Now consider the discrete case, where ν denotes the uniform measure over the s points $(x_i)_{i=1}^s$ defined as $x_i := p_0(z_i) \in \mathbb{R}^d$. Further consider the two types of distance.

- Since $z_i < z_{i+1}$ and $\widetilde{f^k}(z_i) \neq \widetilde{f^k}(z_{i+1})$, then taking $(U_i)_{i=1}^s$ to denote the intervals of $\mathcal{I}$ sorted by their left endpoint, $z_i \in U_i$ for $i \in [s]$. By Lemma 3.1,

$$\begin{aligned} \int |\tilde{h} - \tilde{g}| d\nu &= \frac{1}{s} \sum_{i=1}^s |\tilde{h}(x_i) - \tilde{g}(x_i)| = \frac{1}{s} \sum_{i=1}^s |\widetilde{f^k}(z_i) - \widetilde{g \circ p_0}(z_i)| \\ &\ge \frac{1}{s} \sum_{i=1}^s \mathbf{1}[\forall z \in U_i \centerdot \widetilde{f^k}(z) \neq \widetilde{g \circ p_0}(z)] \\ &\ge \frac{1}{2} \left(1 - 2\left(\frac{2^{k-2}}{s}\right)\right) \ge \frac{1}{4}. \end{aligned}$$

- Since $f^k(z_i) \in \{0, 1\}$, then $\widetilde{f^k}(z_i) \neq \widetilde{g}(x_i)$ implies $|f^k(z_i) - g(x_i)| \ge 1/2$, thus $\int_{[0,1]^d} |h - g| d\nu \ge \int_{[0,1]^d} |\tilde{h} - \tilde{g}| d\nu / 2 \ge 1/8$.

Construct the continuous measure μ as follows, starting with the construction of a univariate measure μ_0. Since f^k is continuous, there exists a $\delta \in (0, \min_{i\in[s-1]} |z_i - z_{i+1}|/2)$ so that $|f^k(z) - f^k(z_i)| \leq 1/4$ for any $i \in [s]$ and z with $|z - z_i| \leq \delta$. As such, let μ_0 denote the probability measure which places half of its mass uniformly on these s balls of radius δ (which must be disjoint since f^k alternates between 0 and 1 along $(z_i)_{i=1}^s$), and half of its mass uniformly on the remaining subset of $[0, 1]$. Finally, extend this to a probability measure μ on $[0, 1]^d$ uniformly, meaning μ is the product of μ_0 and the measure μ_1 which is uniform over $[0, 1]^{d-1}$. Now consider the two types of distances.

- By Lemma 3.1,

$$\begin{aligned}
\int |\tilde{h} - \tilde{g}| d\mu(x) &= \iint |\widetilde{f^k}(p_y(z)) - \tilde{g}(p_y(z))| d\mu_0(z) d\mu_1(y) \\
&= \int \sum_{U\in\mathcal{I}} \int \mathbf{1}[z \in U \wedge \widetilde{f^k}(z)) \neq \tilde{g}(p_y(z))] d\mu_0(z) d\mu_1(y) \\
&\geq \int \frac{1}{2s} \sum_{U\in\mathcal{I}} \mathbf{1}[\forall z \in U \,.\, \widetilde{f^k}(z)) \neq \widetilde{g \circ p_y}(z)] d\mu_1(y) \\
&\geq \frac{1}{4}\left(1 - 2\left(\frac{2^{k-2}}{s}\right)\right) \geq \frac{1}{8}.
\end{aligned}$$

- For any $y \in \mathbb{R}^{d-1}$ and $U_i \in \mathcal{I}$ (with corresponding $z_i \in U_i$), if $\widetilde{f^k}(z) \neq \widetilde{g \circ p_y}(z)$ for every $z \in U_i$, then

$$\int_{U_i} |f^k(z) - g(p_y(z))| d\mu_0(z) \geq \int_{|z-z_i|\leq\delta} |f^k(z) - 1/2| d\mu_0(z) \geq \frac{1}{4}\mu_0(\{z \in U_i : |z - z_i| \leq \delta\}) \geq \frac{1}{8s}.$$

  By Lemma 3.1,

$$\begin{aligned}
\int |h - g| d\mu(x) &= \iint |h(p_y(z)) - g(p_y(z))| d\mu_0(z) d\mu_1(y) \\
&\geq \int \sum_{U\in\mathcal{I}} \mathbf{1}[\forall z \in U \,.\, \widetilde{f^k}(z) \neq \tilde{g}(p_y(z))] \int_U |f^k(z) - g(p_y(z))| d\mu_0(z) d\mu_1(y) \\
&\geq \int \frac{1}{8s} \sum_{U\in\mathcal{I}} \mathbf{1}[\forall z \in U \,.\, \widetilde{f^k}(z) \neq \widetilde{g \circ p_y}(z)] d\mu_1(y) \\
&\geq \frac{1}{16}\left(1 - 2\left(\frac{2^{k-2}}{s}\right)\right) \geq \frac{1}{32}.
\end{aligned}$$

■

As a closing curiosity, Theorem 3.12 implies the following statement regarding polynomials.

Corollary A.1 *For any integer $k \geq 1$, there exists a polynomial $h : \mathbb{R}^d \to \mathbb{R}$ with degree 2^k and a corresponding continuous measure μ which is positive everywhere over $[0, 1]^d$ so that every polynomial $g : \mathbb{R}^d \to \mathbb{R}$ of degree $\leq 2^{k-3}$ satisfies $\int |h - g| d\mu \geq 1/32$.*

Proof Set $f(z) = 4z(1-z)$, which by Lemma 3.10 is $(1,[0,1])$-triangle, thus f^k is $(2^{k-1},[0,1])$-triangle with $\mathrm{Cr}(f^k) = 2^k + 1$ by Corollary 3.9, and f^k has degree 2^k directly; thus set $h(x) = f^k(x_1)$. Next, for any polynomial $g : \mathbb{R}^d \to \mathbb{R}$ of degree $\leq 2^{k-3}$, $g \circ p_y : \mathbb{R} \to \mathbb{R}$ is still a polynomial of degree $\leq 2^{k-3}$ for every $y \in \mathbb{R}^{d-1}$ (where $p_y(z) = (z,y)$ as in Theorem 3.12), and so Lemma 3.3 grants $\mathrm{Cr}(g \circ p_y) \leq 1 + 2^{k-3} \leq 2^{k-2}$. The result follows by Theorem 3.12. ■

A.3. Deferred proofs from Section 4

First, the proof of a certain VC lower bound which mimics the Gilbert-Varshamov bound; the proof is little more than a consequence of Hoeffding's inequality.

Proof (of Lemma 4.1) For convenience, set $m := \mathrm{Sh}(\mathcal{F}; n)$, and let $(a_1, \ldots, a_m)$ denote these dichotomies (meaning $a_j \in \{0,1\}^n$), and with foresight set $\epsilon := \sqrt{\ln(m/\delta)/(2n)}$. Let $(Y_i)_{i=1}^n$ denote fair Bernoulli random labellings for each point, and note by symmetry of the fair coin that for any fixed dichotomy a_j,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n |(a_j)_i - Y_i| < 1/2 - \epsilon\right] = \Pr\left[\frac{1}{n}\sum_{i=1}^n Y_i < 1/2 - \epsilon\right].$$

Consequently, by a union bound over all dichotomies and lastly by Hoeffding's inequality,

$$\begin{aligned}\Pr\left[\exists f \in \mathcal{F} \centerdot \frac{1}{n}\sum_{i=1}^n |\tilde{f}(x_i) - Y_i| < 1/2 - \epsilon\right] &\leq \sum_{j=1}^m \Pr\left[\frac{1}{n}\sum_{i=1}^n |(v_j)_i - Y_i| < 1/2 - \epsilon\right] \\ &= m\Pr\left[\frac{1}{n}\sum_{i=1}^n Y_i < 1/2 - \epsilon\right] \\ &\leq m\exp(-2n\epsilon^2) \leq \delta,\end{aligned}$$

where the last step used the choice of ϵ. ■

The remaining deferred proofs do not exactly follow the order of Section 4, but instead the order of dependencies in the proofs. In particular, to control the VC dimension, first it is useful to prove Lemma 4.3, which is used to control the growth of numbers of regions as semi-algebraic gates are combined.

Proof (of Lemma 4.3) Fix some ordering $(q_1, q_2, \ldots, q_{|\mathcal{Q}|})$ of the elements of $\mathcal{Q}$, and for each $i \in [|\mathcal{Q}|]$ define two functions $l_i(a) := \mathbf{1}[q_i(a) < 0]$ and $u_i(a) := \mathbf{1}[q_i(a) \geq 0]$, as well as two sets $L_i := \{a \in \mathbb{R}^p : l_i(a) = 1\}$ and $U_i := \{a \in \mathbb{R}^p : u_i(a) = 1\}$. Note that

$$\mathcal{S} := \left\{(\cap_{i\in A} L_i) \cap (\cap_{i\in B}) : A \subseteq [|\mathcal{Q}|], B \subseteq [|\mathcal{Q}|]\right\} \setminus \{\emptyset\}.$$

Additionally consider the set of sign patterns

$$V := \left\{\left(l_1(a), u_i(a), \ldots, l_{|\mathcal{Q}|}(a), u_{|\mathcal{Q}|}(a)\right) : a \in \mathbb{R}^p\right\}.$$

Distinct elements of $\mathcal{S}$ correspond to distinct sign patterns in V: namely, for any $C \in \mathcal{S}$, using the ordering of $\mathcal{Q}$ to encode A and B as binary vectors of length $|\mathcal{Q}|$, the corresponding interleaved

binary vector of length $2|\mathcal{Q}|$ is distinct for distinct choices of (A, B). (For each i that appears in neither A nor B, there two possible encodings in V: having both coordinates corresponding to i set to 1, and having them set to 0. On the other hand, a more succinct encoding based just on $(l_i)_{i=1}^{|\mathcal{Q}|}$ fails to capture those sets arising from intersections of proper subsets of $\mathcal{Q}$.) As such, making use of growth function bounds for sets of polynomials (Anthony and Bartlett, 1999, Theorem 8.3),

$$|\mathcal{S}| \leq |V| \leq 2\left(\frac{4e\alpha|\mathcal{Q}|}{p}\right)^p.$$

■

Thanks to Lemma 4.3, the proof of the VC dimension bound Lemma 4.2 follows by induction over layers, effectively keeping track of a piecewise (regionwise?) polynomial function as with the proof of Lemma 3.2 (but now in the multivariate case).

Proof (of Lemma 4.2) First note that this proof follows the scheme of a VC dimension proof for networks with piecewise polynomial activation functions (Anthony and Bartlett, 1999, Theorem 8.8), but with Lemma 4.3 allowing for the more complicated semi-algebraic gates, and some additional bookkeeping for the (semi-algebraic) shapes of the regions of the partition $\mathcal{S}$.

Let examples $(x_j)_{j=1}^n$ be given with $n \geq p$, let m_i denote the number of nodes in layer i (whereby $m_1 + \cdots + m_l = m$), and let $f := F_{\mathfrak{G}} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ denote the function evaluating the neural network (as in Section 2.1), where the two arguments are the parameters $w \in \mathbb{R}^p$ and the input example $x \in \mathbb{R}^d$. The goal is to upper bound the number of dichotomies

$$K := \mathrm{Sh}(\mathcal{N}(\mathfrak{G}); n) = |\{(\mathrm{sgn}(f(w, x_1)), \ldots, \mathrm{sgn}(f(w, x_n))) : w \in \mathbb{R}^p\}|.$$

The proof will proceed by producing a sequence of partitions $(\mathcal{S}_i)_{0=1}^l$ of $\mathbb{R}^p$ and two corresponding sequences of sets of polynomials $(\mathcal{P}_i)_{i=0}^l$ and $(\mathcal{Q}_i)_{i=0}^l$ so that for each i, $\mathcal{P}_i$ has polynomials of degree at most β^i, $\mathcal{Q}_i$ has polynomials of degree at most $\alpha\beta^{i-1}$, and over any parameters $S \in \mathcal{S}_i$, there is an assignment of elements of $\mathcal{P}_i$ to nodes of layer i so that for each example x_j, every node in layer i evaluates the corresponding fixed polynomial in $\mathcal{P}_i$; lastly, the elements of $\mathcal{S}_i$ are intersections of sets of the form $\{w \in \mathbb{R}^p : q(w) \diamond 0\}$ where $q \in \mathcal{Q}_i$ and $\diamond \in \{<, \geq\}$, and the partition $\mathcal{S}_{i+1}$ refines $\mathcal{S}_i$ for each i (meaning for each $U \in \mathcal{S}_{i+1}$ there exists $S \supseteq U$ with $S \in \mathcal{S}_i$). Setting the final partition $\mathcal{S} := \mathcal{S}_l$, this in turn will give an upper bound on K, since the final output within each element of $\mathcal{S}$ is a fixed polynomial of degree at most β^l, whereby the VC dimension of polynomials (Anthony and Bartlett, 1999, Theorem 8.3) grants

$$K \leq \sum_{S \in \mathcal{S}} |\{(\mathrm{sgn}(f(w, x_1)), \ldots, \mathrm{sgn}(f(w, x_n))) : w \in S\}| \leq 2|\mathcal{S}|\left(\frac{2en\beta^l}{p}\right)^p. \tag{A.1}$$

To start, consider layer 0 of the input coordinates themselves, a collection of d affine maps. Consequently, it suffices to set $\mathcal{S}_0 := \{\mathbb{R}^p\}$, $\mathcal{Q}_0 := \emptyset$, and $\mathcal{P}_0$ to be the nd possible coordinate maps corresponding to all d coordinates of all n examples.

For the inductive step, consider some layer $i + 1$. Restricted to any $S \in \mathcal{S}_i$, the nodes of the previous layer i compute fixed polynomials of degree β^i. Each node in layer $i + 1$ is (t, α, β)-sa, meaning there are t predicates, defined by polynomials of degree $\leq \alpha$, which define regions wherein this node is a fixed polynomial. Let Q_S denote this set of predicates, where $|Q_S| \leq tnm_{i+1}$ by

considering the n possible input examples and the t possible predicates encountered in each of the m_{i+1} nodes in layer $i+1$, and set $Q_{i+1} := Q_i \bigcup (\cup_{S\in\mathcal{S}_i} Q_S)$. By the definition of semi-algebraic gate, each node in layer $i+1$ computes a fixed polynomial when restricted to a region defined by an intersection of predicates which moreover are defined by Q_{i+1}. As such, defining $\mathcal{S}_{i+1}$ as the refinement of $\mathcal{S}_{i+1}$ which partitions each $S \in \mathcal{S}_i$ according to the intersections of predicates encountered in each node, then Lemma 4.3 on each Q_S grants

$$|\mathcal{S}_{i+1}| \leq \sum_{S\in\mathcal{S}_i} |\{\text{all nonempty intersections of } Q_S\}| \leq 2|\mathcal{S}_i| \left(\frac{4enm_{i+1}t\alpha\beta^i}{p}\right)^p, \tag{A.2}$$

which completes the inductive construction.

The upper bound on K may now be estimated. First, $|\mathcal{S}|$ may be upper bounded by applying eq. (A.2) recursively:

$$|\mathcal{S}| \leq |\mathcal{S}_0| \prod_{i=1}^{l} \left(\frac{8enm_i t\alpha\beta^{i-1}}{p}\right)^p \leq \left(8enmt\alpha\beta^{l-1}\right)^{pl}.$$

Continuing from Equation (A.1),

$$K \leq 2|\mathcal{S}| \left(\frac{2em\beta^l}{p}\right)^p \leq \left(8enmt\alpha\beta^l\right)^{p(l+1)}.$$

To compute $\mathrm{VC}(\mathcal{N}(\mathfrak{G}))$, it suffices to find N such that $\mathrm{Sh}(\mathcal{N}(\mathfrak{G}); N) < 2^N$, which in turn is implied by $p(l+1)\ln(N) + p(l+1)\ln(8emt\alpha\beta^l) < N\ln(2)$. Since $\ln(N) = \ln(N/(2p(l+1)) + \ln(2p(l+1)) \leq N/(2p(l+1)) - 1 + \ln(2p(l+1))$ and $\ln(2) - 1/2 > 1/6$, it suffices to show

$$6p(l+1)\left(\ln(2p(l+1)) + \ln(8emt\alpha\beta^l)\right) \leq N.$$

As such, the left hand side of this expression is an upper bound on $\mathrm{VC}(\mathcal{N}(\mathfrak{G}))$. ■

The proofs of Lemma 1.3 and Theorem 1.2 from Section 1 are now direct from Lemma 4.2 and Lemma 4.1.

Proof (of Lemma 1.3) This statement is the same as Lemma 4.2 with some details removed. ■

Proof (of Theorem 1.2) By the bound on $\mathrm{Sh}(\mathcal{N}(\mathfrak{G}); n)$ from Lemma 4.2,

$$\begin{aligned}
n = \frac{n}{2} + \frac{n}{2} &\geq 2\ln(1/\delta) + 4pl^2\ln(8emt\alpha\beta p(l+1)) + \frac{n}{2} \\
&\geq 2\ln(1/\delta) + 2p(l+1)\ln(8emt\alpha\beta^l) + 2p(l+1)\left(\ln(p(l+1))) + \frac{n}{2p(l+1)} - 1\right) \\
&\geq 2\ln(1/\delta) + 2p(l+1)\ln(8emt\alpha\beta^l) + 2p(l+1)\ln(n) \\
&\geq 2\ln(1/\delta) + 2\ln(\mathrm{Sh}(\mathcal{N}(\mathfrak{G}); n)).
\end{aligned}$$

The result follows by plugging this into Lemma 4.1. ■

A.4. Deferred proofs from Section 5

Proof (of Proposition 5.1) Immediate from Lemma 3.11. ■