Finiteness Results for Sigmoidal "Neural" Networks

Proc. 25th Annual Symp. Theory Computing, San Diego, May 1993

(Extended Abstract)

Angus Macintyre
Mathematical Inst., University of Oxford
Oxford OX1 3LB, England, UK
E-mail: ajm@maths.ox.ac.uk

Eduardo D. Sontag*
Dept. of Mathematics, Rutgers University
New Brunswick, NJ 08903
E-mail: sontag@hilbert.rutgers.edu

Abstract

This paper deals with analog circuits. It establishes the finiteness of VC dimension, teaching dimension, and several other measures of sample complexity which arise in learning theory. It also shows that the equivalence of behaviors, and the loading problem, are effectively decidable, modulo a widely believed conjecture in number theory. The results, the first ones that are independent of weight size, apply when the gate function is the "standard sigmoid" commonly used in neural networks research. The proofs rely on very recent developments in the elementary theory of real numbers with exponentiation. (Some weaker conclusions are also given for more general analytic gate functions.) Applications to learnability of sparse polynomials are also mentioned.

1 Introduction

"Multilayer neural networks" employing smooth threshold gates have been proposed during the past few years as models of parallel analog computation. In such networks, each processor computes a function of the type: $(u_1, \ldots, u_r) \mapsto \sigma\left(\sum_{i=1}^{r} w_i u_i - w_0\right)$ where σ is a differentiable function $\mathbb{R} \to \mathbb{R}$ (called the "gate" or "activation" function), and the w_i are real numbers (the "weights" or, for $i = 0$, the "threshold") associated to the particular processor. The u_i's include external inputs to the network as well as values calculated by other units. The output of one of the processors is singled out and designated as the output of the network. The graph describing the interconnections is assumed to be acyclic, so one thinks of such a device as computing, in the obvious manner, a function of the external inputs.

*Supported in part by US Air Force Grant AFOSR-91-0343. Partially completed while visiting Siemens Corporate Research.

The paper [14] compared the power of smooth circuits with the power of those based on step-function threshold gates (that is, $\sigma(x)$ would be 0 for $x < 0$ and 1 otherwise). Step-function threshold networks are common in circuit complexity studies, and this comparison was quantified in terms of the scaling of network size with respect to input size, as is common in that area. It was shown that, if weights do not grow more than polynomially, and under natural conventions on output separation, the usual classes of fixed-depth polynomial-sized circuits (namely, TC^0_d) are recovered. However, it was also shown that, perhaps surprisingly, a provable gain in efficiency exists for constant-sized circuits of depth 2. For computing functions of real inputs —as opposed to Boolean functions, but now leaving the input dimension constant— gains in effiency are also provable; see for instance the recent papers [18] and [4].

In applications work —see e.g. the textbook [9]— it is most common to use the activation function $\sigma_{\rm s}(x) = 1/(1+e^{-x})$, and we do so here for our main results. (Some results are developed in more generality, however.) Networks employing the function $\sigma_{\rm s}$ are nowadays routinely used in a wide variety of empirical learning problems. Differentiability of σ is needed since it is common to use gradient descent techniques for the numerical approximation of training data by a feedforward network. The particular choice $\sigma = \sigma_{\rm s}$ is due to ease of calculation of derivatives in terms of function values, since $\sigma'_{\rm s}(x) = (1-\sigma_{\rm s}(x))\sigma_{\rm s}(x)$, as well as for other qualitative considerations. (Another choice sometimes made is $\sigma(x) = \arctan(x)$.) These applications motivated to some extent the major work of Haussler in [8], which dealt with, among other topics, sample complexity results for smooth networks. The results in [8] left open, however, the question of whether, for distribution-independent learning, the sample complexity of the class of concepts defined by any fixed architecture (interconnection graph given, but weights to be chosen) is finite for any given error probabilities, *if no assumptions are made as to boundedness of weights*. One of the purposes of this work is to show that the Vapnik-Chervonenkis dimension of such concept classes is finite, which implies the desired sample complexity fact. The proof relies on some very recent developments in model theory which can be seen as an extension of both the classical Tarski-Seidenberg theory and the Whitney stratification of algebraic sets. The result will be proved in more generality, allowing more general polynomials, rather than just affine combinations, in the expressions that appear as arguments to σ; this has the advantage of allowing results for so-called "high-order" or "sigma-pi" neural nets, which have also appeared in the literature (cf. [5]).

In [20], it was remarked that, as an easy consequence of elementary properties of VC dimension and Tarski-Seidenberg theory, concept classes defined algebraically —in particular, neural nets with polynomial activations— do give rise to finite-VC dimension concept classes. (In that context, see [13, 6] for similar questions when using *piecewise polynomial* activations. Note also that for step-function threshold gates, sample complexity bounds are by now quite well understood, with fairly good bounds on VC dimension available; see for instance [1].) In [20], the authors also gave results that apply to a restricted class of (real-)analytic functions on bounded domains; for neural nets, their results would apply to the very special case of bounded inputs and a single adjustable (and bounded) weight. The recent model-theory work that we use allows dropping the single-weight assumption in this general result, but, far more importantly, it permits dropping all boundedness conditions when dealing with the standard sigmoid $\sigma_{\rm s}(x) = 1/(1+e^{-x})$. It is not hard to see that analyticity, for instance, is not sufficient for the validity of the VC dimension result; an easy counterexample is furnished by using networks based on $\sin(x)$. Even general qualitative properties such as monotonicity and existence of limits at infinity do not help, so more sophisticated tools must be brought in. (Under strong nondegeneracy conditions, one could also use techniques as in [11], but these assumptions are rarely satisfied in the networks application.)

We also apply these tools to the solution of several other open problems dealing with sigmoidal

nets, establishing finiteness of other "dimensions" which have appeared in this context. In particular, we show that the "teaching dimension" for real-valued functions determined by any given architecture is finite (and bounded by the number of weights plus one), and a similar result is given for a measure of interpolation capabilities. We also establish the finiteness of the Haussler/Pollard "pseudodimension." Finally, we show that, subject to the validity of a conjecture in number theory which is widely believed to be true, the question of equality of input/output behaviors (and the "loading" question in the sense of e.g. [3]) for sigmoidal neural nets is effectively decidable.

The techniques used here permit also answering in the positive the following open question, posed in [10], Section 4: is the VC dimension of the class of concepts defined by r-sparse polynomials in m variables finite?

We wish to emphasize that we do not provide explicit bounds for VC dimension, but merely establish finiteness. However, since the results being used are essentially constructive, it would be possible in principle to compute the VC dimension of a given architecture. In any case, we view our contribution as providing a motivation for the search for good bounds for particular architectures.

In closing this introduction, we remark that when the interconnection graph is permitted to have loops, it is natural to study neural networks as dynamical systems, in the role of language acceptors. For such "feedback" nets, the gains in capabilities when introducing continuous activations are even more drastic, allowing the passage from regular language recognition to Turing capabilities, and even to all of P/poly; see the recent work [16] and [17]. Moreover, in that context most reasonable concept classes turn out to have infinite VC dimension. We emphasize that in this paper we restrict attention to loop-free, feedforward, networks, and deal strictly with the computation of functions on Euclidean spaces.

2 Feedforward Networks

We will introduce the concept of an *architecture*, that is, a "neural network" for which all weights are thought of as variables. We retrict attention to architectures in which no feedback loops are allowed in signal transmission. An architecture serves to define a "wiring diagram" which indicates the flow of information among neurons, as well as, for each neuron, a rule that states what particular combination of the incoming signals will be used as its input. When these rules, specified by the polynomials P_l in the definition to be given below, happen to be given simply by affine functions, the architecture is said to be *first order*; this special case is the one most often considered in neural networks research. We treat more general (polynomial, or "high order") architectures, since we can obtain results for these with no extra effort.

We let Σ be any fixed set of functions $\mathbb{R} \to \mathbb{R}$. In the context of networks, we will refer to elements of Σ as *activations*.

Definition 2.1 A *feedforward network architecture* $\mathcal{A}$ (with activations in Σ) is a labeled directed acyclic graph as follows. We assume given an integer r; the space $\mathbb{R}^r$ will be the *weight space* for the architecture. There are a number m of input nodes (i.e., nodes of in-degree zero), and these are labeled by variables $x_1, \ldots, x_m$. The rest of the nodes are called *computation nodes*; exactly one of these is also an output node, i.e., a node of out-degree zero. The lth computation node N_l is labeled by a variable z_l as well as a polynomial P_l and an element σ_l of Σ, where, for some $\rho = \rho(l)$, $\mu = \mu(l)$, and $\nu = \nu(l)$, $P_l = P_l(w_{i_1}, \ldots, w_{i_\rho}, z_{j_1}, \ldots, z_{j_\mu}, x_{k_1}, \ldots, x_{k_\nu})$. Here $\{w_{i_1}, \ldots, w_{i_\rho}\}$ is a subset of the weight variables $\{w_1, \ldots, w_r\}$, while $\{z_{j_1}, \ldots, z_{j_\mu}\}$ are the computation variables and $\{x_{k_1}, \ldots, x_{j_\nu}\}$ the input variables corresponding to those nodes (computation and input, respectively) that are incident to N_l. In the case of the output node, we denote the corresponding variable z_l simply by y. □

The subsets of weights $\{w_{i_1}, \ldots, w_{i_\rho}\}$ appearing in each P_l are usually taken to be disjoint for

different l, but mathematically this will make no difference, so we do not make that assumption. In fact, we could equally well take all weights to appear in each P_l, thought of now as a function of all the variables.

Definition 2.2 Assume given a feedforward network architecture $\mathcal{A}$ with activations in the set Σ. We associate to $\mathcal{A}$ a *behavior*, which is a function $\beta_{\mathcal{A}} : \mathbb{R}^r \times \mathbb{R}^m \to \mathbb{R}$. The behavior is defined inductively on nodes, starting with inputs, as follows. If N is the ith input node, it computes the function $f(w,x) = x_i$. For the computation node N_l, its function is

$$f(w,x) := P_l(w_{i_1}, \ldots, w_{i_\rho}, \\ \sigma_{j_1}(f_{j_1}(w,x)), \ldots, \sigma_{j_\mu}(f_{j_\mu}(w,x)), x_{k_1}, \ldots, x_{k_\nu}) \ ,$$

where $f_{j_l}(w,x)$ denotes the function computed by the node j_l. Finally, the function $\beta_{\mathcal{A}}$ is defined as the function computed by the output node. □

The *function computed by the network corresponding to a given set of weights* $w_0 \in \mathbb{R}^r$ is by definition the function $\beta_{\mathcal{A}}(w_0, \cdot) : \mathbb{R}^m \to \mathbb{R}$. The *class of functions computed by* $\mathcal{A}$ is defined as the set of functions $\{\beta_{\mathcal{A}}(w_0, \cdot) : \mathbb{R}^m \to \mathbb{R}, w_0 \in \mathbb{R}^r\}$. When $\mathcal{A}$ is clear from the context, we write simply β instead of $\beta_{\mathcal{A}}$.

The main results will be for the special case $\Sigma = \{\sigma_{\rm s}\}$, where we use the notation $\sigma_{\rm s}$ to denote the *standard sigmoid*, given by: $\sigma_{\rm s}(x) := \frac{1}{1+e^{-x}}$ (alternatively, one may use the function $\tanh(x)$, which is obtained by rescaling and adding a constant). The choice $\sigma = \sigma_{\rm s}$ is standard in neural network practice, as discussed in the introduction. More generally, one may allow a larger set Σ, containing $\sigma_{\rm s}$, as follows.

Pick any positive integer l, and a cube $C = [-k,k]^l$ in $\mathbb{R}^l$. Assume that g is a real-valued function which is (real-)analytic in a neighborhood of C. By the *0-restriction* of g to C we will mean the function $f : \mathbb{R}^l \to \mathbb{R}$ which equals 0 outside C and equals g on C. A *restricted analytic* (RA) *function* is any function obtained in this manner. A function $f : \mathbb{R}^m \to \mathbb{R}$ will be said to be *exp*-RA *definable* if it can be defined in terms of a first-order logic sentence involving the standard propositional connectives, existential and universal quantification, algebraic operations, and symbols for the exponential function as well as all RA functions. Of course, $\sigma_{\rm s}$ is exp-RA definable, since $y = \sigma_{\rm s}(x)$ if and only if $y(1 + \exp(-x)) = 1$. Any RA function is in particular exp-RA definable. The function $\arctan(x)$ is also exp-RA definable, since $y = \arctan(x)$ if and only if $-\pi/2 < y < \pi/2$ and $\mathrm{SIN}\,(y) = x\,\mathrm{COS}\,(y)$, where SIN and COS denote the restrictions of sin and cos to $[-\pi/2, \pi/2]$. Compositions such as $\arctan(\exp(\exp(x)))$ are also allowed. However, the function $\sin(x)$ is *not* exp-RA definable (this will follow from the fact that the VC dimension result to be given below does not hold for this function).

When the set of activations Σ consists of exp-RA definable functions, the behavior β is again exp-RA definable. All results to be given for networks using such activations are in fact results that hold for *any* set of functions $\{\beta(w_0, \cdot) : \mathbb{R}^m \to \mathbb{R}, w_0 \in \mathbb{R}^r\}$, for any exp-RA definable function $\beta : \mathbb{R}^m \to \mathbb{R}$. We use the network terminology, however, since this is what motivated the paper, and because examples and counterexamples are given for networks.

3 Statements of Main Results

We next state our main results. Proofs are deferred to later sections.

3.1 Vapnik-Chervonenkis Dimension

Let $\mathcal{D}$ be any fixed set, to be thought of as the set of "inputs." For each subset $\mathcal{X}$ of $\mathcal{D}$, a *dichotomy* on $\mathcal{X}$ is a function $c : \mathcal{X} \to \{0,1\}$. We say that a function $f : \mathcal{D} \to \mathbb{R}$ *implements* the dichotomy c if and only if $c(s) > 0 \iff f(x) > 0$. Let $\mathcal{F}$ be any class of functions $\mathcal{D} \to \mathbb{R}$. The subset

$\mathcal{X}$ is *shattered* by $\mathcal{F}$ if each dichotomy on $\mathcal{X}$ can be implemented by some function $f \in \mathcal{F}$. The *Vapnik-Chervonenkis dimension* $\text{VC}(\mathcal{F})$ of the function class $\mathcal{F}$ is the supremum (possibly infinite) of the set of integers κ for which there is some subset $\mathcal{X} \subseteq \mathcal{D}$ of cardinality κ that can be shattered by $\mathcal{F}$.

In particular, consider, for any given architecture $\mathcal{A}$ with activation set Σ, the set of all functions of the type $\beta_{\mathcal{A}}(w_0, \cdot)$ for w_0 ranging over all possible $w_0 \in \mathbb{R}^r$. Here, $\mathcal{D} = \mathbb{R}^m$. In this case, we write simply $\text{VC}(\mathcal{A})$ and refer to the VC dimension of $\mathcal{A}$.

Theorem 1 *Assume that Σ consists of exp-*RA *definable functions. For every feedforward network architecture $\mathcal{A}$, $\text{VC}(\mathcal{A}) < \infty$.*

Thus nets using the standard activation σ_{s}, or even σ_{s} combined with certain other functions such as arctan, give rise to finite VC dimension. In general, however, it may be the case that $\text{VC}(\mathcal{A}) = \infty$, even for analytic σ, and even if one asks that σ be of "sigmoidal" type, as discussed later. But in the very special case of bounded weights, analytic activations, and bounded input space $\mathcal{D}$, there is a finiteness result which follows immediately from the above. More precisely, for each real $\gamma > 0$ and each architecture $\mathcal{A}$, consider the class of functions $\{\beta_{\mathcal{A}}(w_0, \cdot) : [-\gamma, \gamma]^m \to \mathbb{R},\ \|\omega_0\| \leq \gamma\}$, where we take any fixed norm in weight space $\mathbb{R}^r$. The VC dimension of this function class will be denoted by $\text{VC}(\mathcal{A}, \gamma)$.

Corollary 3.1 Assume that Σ consists of analytic functions. Then, for any feedforward network architecture $\mathcal{A}$ and real $\gamma > 0$, $\text{VC}(\mathcal{A}, \gamma) < \infty$. □

None of the assumptions can be relaxed. If the weights or the inputs are allowed to be unbounded, or if analyticity is replaced by infinite differentiability, the dimension may be infinite. The proof of the Corollary is trivial from Theorem 1. Indeed, if all weights are bounded and inputs are bounded as well, then there is an *a priori* bound on the arguments that appear in each activation. Thus one may replace each activation by a suitable restriction to a compact set, so that all functions become RA functions, and the Theorem applies.

3.2 Haussler's Pseudo-Dimension

We next define a few other notions of "dimension" of function classes. Again, $\mathcal{F}$ will be any class of functions $\mathcal{D} \to \mathbb{R}$, and when this class is that of all $\beta_{\mathcal{A}}(w_0, \cdot)$, $w_0 \in \mathbb{R}^r$, we will just write "$\mathcal{A}$" for the resulting function class and refer to the "dimension of the architecture."

For each subset $\mathcal{X} = \{x_1, \ldots, x_s\}$ of $\mathcal{D}$, we will say that $\mathcal{X}$ is *H-shattered* by $\mathcal{F}$ (the terminology is just "shattered" in Haussler's work [8]) if there exist real numbers $y_1, \ldots, y_s$ such that every dichotomy of $\widetilde{\mathcal{X}} := \{(x_1, y_1), \ldots, (x_s, y_s)\}$ can be implemented by some function of the form $(x, y) \mapsto \widetilde{f}(x, y) := f(x) - y$. The *Pseudo-Dimension* $\text{PD}(\mathcal{F})$ of the class $\mathcal{F}$ is the supremum (possibly infinite) of the set of integers κ for which there is some set $\mathcal{X}$ of cardinality κ that can be H-shattered by $\mathcal{F}$.

This notion of dimension turns out to be useful when studying learning-theoretic questions for real-valued (as opposed to binary-valued) functions, for instance, regression functions in statistics. According to [8], it originates in the work of Pollard. Note that $\text{VC}(\mathcal{F}) \leq \text{PD}(\mathcal{F})$ (case when all $y_i = 0$).

For any architecture $\mathcal{A}$, consider the architecture $\mathcal{A}'$ obtained from $\mathcal{A}$ by adding a new input "y" and a new computation node (the new output node) which computes $y_{\mathcal{A}} - y$, where $y_{\mathcal{A}}$ is the value of the function computed by the output node (now a non-output computation node) of $\mathcal{A}$. Obviously, $\text{PD}(\mathcal{A}) \leq \text{VC}(\mathcal{A}')$. (In fact, equality holds, because a set shattered by $\mathcal{A}'$ cannot contain two points with same x-component.) Thus, from Theorem 1 applied to $\mathcal{A}'$ we can conclude:

Theorem 2 *Assume that Σ consists of exp-*RA *definable functions. For every feedforward network architecture $\mathcal{A}$,* PD $(\mathcal{A}) < \infty$.

Similarly, Corollary 3.1 generalizes to this case.

3.3 Interpolation Dimension

For each finite subset $\mathcal{X}$ of $\mathcal{D}$, a *labeling* of $\mathcal{X}$ is a function $\lambda : \mathcal{X} \to \mathbb{R}$. The *error* of a given function $f : \mathcal{D} \to \mathbb{R}$ on a labeling λ is defined as: $E(f, \lambda) := \sum_{x \in \mathcal{X}} |f(x) - \lambda(x)|^2$. We say that λ *can be loaded* into the class of functions $\mathcal{F}$ (in the particular case of feedforward network architectures, "into the architecture $\mathcal{A}$") if

$$\inf_{f \in \mathcal{F}} E(f, \lambda) = 0 .$$

The set $\mathcal{X}$ will be said to be *I-shattered* by $\mathcal{F}$ if each labeling of $\mathcal{X}$ can be loaded into $\mathcal{F}$.

A weaker requirement is that, for some $\varepsilon > 0$, $\mathcal{X}$ be *ε-I-shattered* by $\mathcal{F}$, when one asks merely that each labeling $\lambda : \mathcal{X} \to (-\varepsilon, \varepsilon)$ of $\mathcal{X}$ can be loaded.

Note that for the case when $\mathcal{F} = \{\beta_{\mathcal{A}}(w_0, \cdot), w_0 \in \mathbb{R}^r\}$, requiring that $\mathcal{X} = \{x_1, \ldots, x_s\}$ be I-shattered amounts to asking that the mapping

$$\gamma_{\mathcal{X}} : \mathbb{R}^r \to \mathbb{R}^s : \; w \mapsto (\beta(w, x_1), \ldots, \beta(w, x_s)) , \tag{1}$$

from weights to outputs corresponding to inputs in $\mathcal{X}$, has a dense image, and ε-I-shattering is the same as the requirement that the image of this map intersect $(-\varepsilon, \varepsilon)^s$ densely.

The *interpolation dimension* ID $(\mathcal{F})$ of the function class $\mathcal{F}$ is the supremum (possibly infinite) of the set of integers κ for which there is an $\varepsilon > 0$ and some set $\mathcal{X}$ of cardinality κ that can be ε-I-shattered by $\mathcal{F}$. (Note that if one would define ID $(\mathcal{F})$ using I-shattering rather than ε-I-shatterings the dimension would be no greater; thus the upper bound to be given below holds in that case as well.)

This notion of capacity is natural in the context of neural network practice, where least-squares techniques are used in order to minimize the error $E(\beta(w, \cdot), \lambda)$ as a function of the parameters w, for experimental data given by λ. If weights in the output layer are restricted to be small, only small targets are reasonable, hence the interest in ε-shattering.

The next result is "as expected," but the use of σ_{s} is essential here. Similar results are false even for other analytic functions that qualitatively look very much like σ_{s} (strictly increasing, limits at $\pm\infty$, etc); see [18] for such counterexamples.

Theorem 3 *Assume that Σ consists of exp-*RA *definable functions. For every feedforward network architecture $\mathcal{A}$,* ID $(\mathcal{A}) \leq r$.

Remark 3.2 Observe that it is possible for ID $(\mathcal{A})$ to be far smaller than r. For instance, consider the architecture with $\Sigma = \{\sigma_{\mathrm{s}}\}$ in which the nodes are totally ordered ordered and where at each computation node the polynomial P_l is affine on incoming node variables. Here $r = 2(k-1)$, where k is the number of nodes. As all the functions $\beta_{\mathcal{A}}(w, \cdot) : \mathbb{R} \to \mathbb{R}$ are necessarily monotone, ID $(\mathcal{A}) = 2$, independently of the number of nodes. (The same argument shows that VC $(\mathcal{A}) = 2$ as well.) □

Remark 3.3 As another example, take a *k-unit one hidden layer* architecture with a scalar input and $\Sigma = \{\sigma_{\mathrm{s}}\}$. Here k is a positive integer, there are $k+1$ computation nodes and 1 input node. The polynomials P_l are all affine on incident node variables, and the graph is such that the input node is incident to the k non-output computation nodes, and these in turn are all incident to the output node. Thus $r = 3k + 1$, and

$$\beta(w, x) = a_0 + \sum_{i=1}^{k} a_i \sigma(b_i x + c_i) , \tag{2}$$

where $w = (a_0, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k)$ and $\sigma = \sigma_s$. The above result says that $\text{ID}(\mathcal{A}) \leq 3k+1$. This fact had also been proved, for this very special architecture, by an ad-hoc argument in [18], where it was also shown that $\text{ID}(\mathcal{A}) \geq 2k-1$. The precise value of $\text{ID}(\mathcal{A})$ in this case seems to be open. □

3.4 Teaching Dimension

Again assume that a class $\mathcal{F}$ of functions $\mathcal{D} \to \mathbb{R}$ has been fixed. We first introduce some equivalence relations. For each $x \in \mathcal{D}$ and each two functions $f_1, f_2 \in \mathcal{F}$, we denote $f_1 \underset{x}{\sim} f_2$ if $f_1(x) = f_2(x)$. More generally, given a subset $\mathcal{X}$ of $\mathcal{D}$, $f_1 \underset{\mathcal{X}}{\sim} f_2$ means that $f_1 \underset{x}{\sim} f_2$ for all $x \in \mathcal{X}$. If $f_1 \underset{x}{\sim} f_2$ for all possible $x \in \mathcal{D}$, that is, if $f_1 = f_2$, we write also $f_1 \sim f_2$.

Following [7], we say that a *teaching subset* for a function $f_0 \in \mathcal{F}$ is a subset $\mathcal{X}$ of $\mathcal{D}$ such that, for every $f \in \mathcal{F}$, $f_0 \underset{\mathcal{X}}{\sim} f \;\Rightarrow\; f_0 \sim f$. The *teaching dimension* $\text{TD}(\mathcal{F})$ is the smallest integer κ (possibly infinite) with the property that for each $f_0 \in \mathcal{F}$ there is some teaching subset of size κ.

Thus, a teaching subset ("teaching sequence" in [7], but the order is immaterial) is a set of inputs that would allow a teacher to uniquely specify the particular function among all other functions of interest. The smallest bound on the size of such a set, over all f_0 to be taught, is the teaching dimension of the class.

For the case when an architecture $\mathcal{A}$ has been specified and we take $\mathcal{F} = \{\beta_{\mathcal{A}}(w_0, \cdot), w_0 \in \mathbb{R}^r\}$, we write $w_1 \underset{x}{\sim} w_2$ instead of $\beta(w_1, \cdot) \underset{x}{\sim} \beta(w_2, \cdot)$, and so forth, and we talk about teaching subsets for weights w_0, and the teaching dimension of $\mathcal{A}$.

A *universal identification set* $\mathcal{X}$ is one that is a teaching subset for all possible $f_0 \in \mathcal{F}$. In other words, "$\underset{\mathcal{X}}{\sim}$" is the same as simply "$\sim$" or equivalently for the case of architectures, the mapping in Equation (1) induces an embedding of $\mathbb{R}^r/\sim$ into $\mathbb{R}^s$. The *universal teaching dimension* $\text{UTD}(\mathcal{F})$ is the smallest integer κ (possibly infinite) with the property that there is some universal teaching subset of size κ. Clearly $\text{TD}(\mathcal{F}) \leq \text{UTD}(\mathcal{F})$.

The result to be given below provides a simple upper bound on the size needed for (universal) teaching subsets. Moreover, the result shows that in a precise sense, to be defined next, "almost every" subset of a given cardinality has the desired property.

Let M be any analytic manifold (in all results to follow, $M = \mathbb{R}^l$, for some positive integer l). A subset Z of M will be said to be *analytically thin* if it can be expressed as a finite or countable union of embedded analytic submanifolds of positive codimension. Such a set has zero measure and is of the first category (a countable union of nowhere dense sets), as discussed later. A subset Z of M will be said to be *finitely analytically thin* if it is a finite union of such submanifolds (it is hence nowhere dense). By abuse of terminology, we'll say that a family $\mathcal{Z}$ of k-element subsets of $\mathbb{R}^m$ is (finitely) analytically thin if the set of vectors $(x_1, \ldots, x_k) \in \mathbb{R}^{km}$ so that $\{x_1, \ldots, x_k\} \in \mathcal{Z}$ is (finitely) analytically thin.

Theorem 4 *Assume that Σ consists of analytic functions. Then, for every feedforward network architecture $\mathcal{A}$, $\text{TD}(\mathcal{A}) \leq r+1$ and $\text{UTD}(\mathcal{A}) \leq 2r+1$. Moreover, the set of universal teaching subsets of size $2r+1$, and for each w_0 the set of teaching subsets for w_0 of size $r+1$, have analytically thin complements. In the particular case in which Σ consists of exp-RA definable functions, one may replace "finitely analytically thin" for analytically thin in the above statement.*

Remark 3.4 The original definitions in [7] dealt with learning binary rather than real-valued functions. However, for infinite classes such as we consider in dealing with neural networks, the binary "teaching dimensions" obtained from just considering the sign of the outputs are, except for degenerate cases, always infinite (just from the signs of outputs for a finite number of inputs

one cannot predict the signs at all other points). Thus the binary case is not interesting in this context. Closely related to teaching subsets for classes of functions is the more general notion of "universal inputs for observability" that appears in the study of controlled dynamical systems; see [19], in particular Section 5.1 and the references for that section. □

It is also interesting to consider the case in which the space of inputs is restricted. Instead of Euclidean space, one might want to consider, for instance, only points $x \in \mathcal{D}$ with integer or rational coordinates. We can modify the definitions to deal with this more general situation as follows. We assume that a function class $\mathcal{F}$ has been fixed, and also that a subset $\mathcal{I}$ of the input space $\mathcal{D}$ has been chosen. Now we define a teaching subset, for a particular function f_0 and relative to $\mathcal{I}$, to be a subset $\mathcal{X}$ of $\mathcal{I}$ so that $f_0 \overset{\sim}{\mathcal{X}} f \;\Rightarrow\; f_0 \overset{\sim}{\mathcal{I}} f$. The teaching dimension $\text{TD}_{\mathcal{I}}(\mathcal{F})$ is the smallest integer integer κ (possibly infinite) with the property that for each $f_0 \in \mathcal{F}$ there is some such teaching subset, relative to $\mathcal{I}$, of size κ. In the case of architectures, we again write just $\text{TD}_{\mathcal{I}}(\mathcal{A})$. Analogously, we can also define a universal dimension relative to $\mathcal{I}$.

Theorem 5 *Let $\Sigma = \{\sigma_s\}$. For every feedforward network architecture $\mathcal{A}$, and each $\mathcal{I} \subseteq \mathcal{D}$, there is a finite universal identification set; in particular, $\text{TD}_{\mathcal{I}}(\mathcal{A}) \leq \text{UTD}_{\mathcal{I}}(\mathcal{A}) \leq \infty$.*

3.5 Decidability Issues

Assume that $\Sigma = \{\sigma_s\}$. Given an architecture $\mathcal{A}$, and two weight vectors $w_1, w_2 \in \mathbb{R}^r$, it is not entirely trivial to determine if the resulting functions are the same, that is, if $w_1 \sim w_2$. (One notable exception is the "single hidden layer architecture," where, as noted by Pascal Koiran – personal communication,– the results in [21] make the question trivial.) Our next result asserts –modulo the validity of a conjecture in number theory– that the equivalence relation $\sim$ is indeed computable. *Schanuel's conjecture* is the following statement: **(SC)** For any set $\{z_1, \ldots, z_l\}$ of $\mathbb{Q}$-linearly independent complex numbers, $\text{trdeg}_{\mathbb{Q}}\, \mathbb{Q}[z_1, \ldots, z_l, e^{z_1}, \ldots, e^{z_l}] \geq l$. Property (SC) is widely believed to be true; note that it encompasses many classical open problems; for instance, for $z_1 = 1$ and $z_2 = e$ it would imply that e and e^e are algebraically independent.

In order for the following result to make sense, we assume that the coefficients of the polynomials P_l are all rational. The same result will hold under more general conditions involving various types of computable real numbers.

Theorem 6 *Assume that $\Sigma = \{\sigma_s\}$. If (SC) is true, then there is a decision procedure for determining, for any given $w_1, w_2 \in \mathbb{R}^r$, if $w_1 \sim w_2$.*

3.6 A Remark on Sparse Polynomials

Consider, for any two fixed positive integers m, r, the class $\mathcal{P}_{m,r}$ of those functions $f : \mathbb{R}^m \to \mathbb{R}$ that can be expressed as a linear combination of no more than r monomials in the m variables $x_1, \ldots, x_m$. Elements of $\mathcal{P}_{m,r}$ are sometimes called *r-sparse polynomials* (or "fewnomials"). See [10] for background on sparse polynomials, including applications to the harmonic analysis of Boolean circuits and learnability of Boolean functions; this reference states as an open problem the finiteness of VC dimension for $\mathcal{P}_{m,r}$, when $m > 1$.

Theorem 7 *For $\mathcal{F} = \mathcal{P}_{m,r}$, $\text{VC}(\mathcal{F}) < \infty$.*

4 Some Facts on Analytic Functions

We collect here some technical results that will be needed in the proofs.

4.1 Analytically Thin Sets

Let M be any (second countable) analytic manifold of dimension l. Recall that an embedded submanifold Z of M, of dimension q, is a connected subset which, locally around each of its points and up to diffeomorphisms, looks like a "slice" $\{(x_1, \ldots, x_l) \,|\, x_{q+1} = \ldots = x_l = 0\}$. When $q = l$, this is just an open set. Assume now that Z has positive codimension, that is, $q \leq l - 1$. From the definition, it follows that for some open subset $M_0 \subseteq M$, Z is a closed subset of M_0 (in the relative topology). Observe that Z is nowhere dense, that is, its closure has empty interior. (That is, if U is any open subset of M, then $Z \bigcap U$ cannot be dense in U: if U does not intersect M_0 then this is clear; otherwise $U \bigcap M_0$ is a nonempty open set and we may assume without loss of generality that $U \subseteq M_0$; then $U = (U \setminus Z) \bigcup (U \bigcap Z)$, so either $U \setminus Z$ is a nonempty open set, and we are again done, or $U \bigcap Z = U$, but in this latter case Z would contain an open set and hence could not have positive codimension.) Also, such a Z has measure zero.

If Z is a finite or countable union of embedded analytic submanifolds of M of dimension $\leq q$, and q is the smallest such integer, we call q the *dimension* of Z. (It is not hard to verify that the dimension is well-defined, in the sense that it doesn't depend on the particular union of countably many submanifolds being used.) From now on, if we write "$\dim Z$" for a subset Z of a manifold M, we mean implicitly that Z is such a finite or countable union. Note that being analytically thin is equivalent to $\dim Z < \dim M$. We collect next some facts that will be needed later. Essentially, these facts amount to saying that naive parameter counts are well-justified when dealing with analytic mappings.

Proposition 4.1 Assume that M, N, and M_i, $i = 1, \ldots, k$, are analytic manifolds. Let $f : M \to N$ be an analytic mapping. Then:

1. For all $Z \subseteq M$, $\dim f(Z) \leq \dim Z$.
2. For all $Z \subseteq M$,
$$\dim Z \ \leq \ \dim N \ + \max_{y \in N} [\dim f^{-1}(y) \bigcap Z].$$
3. If $Z_i \subseteq M_i$, for $i = 1, \ldots, k$, then $Z = Z_1 \times \ldots \times Z_k \subseteq M_1 \times \ldots \times M_k$ satisfies $\dim Z = \sum_i \dim Z_i$.

Proof. We need first to recall the notion of a semianalytic subset T of M. This is a set T so that, for each $z \in M$, there is some neighborhood U of z so that $T \bigcap U$ is in the Boolean algebra generated by a finite family of subsets of the form $\{f_j(x) > 0\}$, for some analytic functions $f_j : U \to \mathbb{R}$, $j = 1, \ldots, J_z$. Semianalytic subsets are countable unions of embedded submanifolds, as follows from the stratification theorems cited later. Conversely, it is easy to see from the definition of embedded submanifold that if Z is such a submanifold, then Z is a countable union of semianalytic subsets. Moreover, as M can be written as a countable union of compacts, after intersecting with suitable compact sets one knows that Z is a countable union of semianalytic subsets with compact closure. Thus saying "countable union of embedded submanifolds" is the same as "countable union of semianalytic subsets with compact closure." Note that $\dim Z = q$ if and only if Z can be written as such a union in such a way that the maximum of the dimensions of the subsets is q.

Now let $f : M \to N$ be analytic and let Z be as above. In order to calculate the dimension of $f(Z)$, it is enough, by the above considerations, to do this when Z is a relatively compact semianalytic subset. But in that case, $f(Z)$ is a subanalytic (= proper analytic image of a semianalytic) subset of N, so it is indeed a countable union of embedded submanifolds of N, by the stratification theorem for subanalytic sets (for which see, for instance, [22] or [2]), and the dimension inequality follows by the stratification theorem applied to f.

To prove the statement about fibres $f^{-1}(y)$, we proceed as follows. Note that each such fibre is semianalytic, so its dimension is well-defined. Write Z as a countable union of compact semianalytic subsets Z_i; then $\max_y[\dim f^{-1}(y) \bigcap Z] = \max_{y,i} \dim[f^{-1}(y) \bigcap Z_i]$. Fix any i. By the stratification theorem applied to f and relative to Z_i, as in for instance Theorem 9.2 in [22], we can partition N into a countable union of connected analytic embedded submanifolds T_j in such a manner that Z_i is a countable union of embedded submanifolds diffeomorphic to $\mathbb{R}^{n_j} \times T_j$ for various integers n_j and various j's, and on each such set the mapping f is (up to the same diffeomorphism) the projection $\mathbb{R}^{n_j} \times T_j \to T_j$. Thus $q = \max_y \dim[f^{-1}(y) \bigcap Z_i]$ is the largest of these n_j's, while $\dim Z_i$ is at most $q + t$, t= largest dimension of the T_j's, and N has dimension t. This shows that $\dim Z_i \leq \dim N + \max_{y \in N}[\dim f^{-1}(y) \bigcap Z_i] \leq \dim N + \max_{y \in N}[\dim f^{-1}(y) \bigcap Z]$ from which, since $\dim Z = \max_i \dim Z_i$, the conclusion follows.

Finally, to prove that $\dim Z_1 \times Z_2 = \dim Z_1 + \dim Z_2$, simply note that $Z_1 \times Z_2$ equals a union of the type $Z_1^j \times Z_2^k$, for countable coverings by submanifolds for each of Z_1 and Z_2 respectively, and dimensions add as they should for submanifolds. ∎

4.2 Order-Minimality

We will need to use certain recent techniques from model theory. For this purpose, we consider the structure $L = (\mathbb{R}, +, \cdot, <, 0, 1, \exp, \{f, f \in \text{RA}\})$, and the corresponding language for the real numbers with addition, multiplication, and order, as well as one function symbol for real exponentiation and one for each restricted analytic function. The set of (first order) formulas over L is the set of all well-formed logical expressions obtained by using propositional connectives, real numbers as constants, the operations of addition and multiplication, the relations $<$ and $=$, and exp and restricted analytic functions as functions; quantification is allowed over variables. By abuse of notation, when giving such a formula, we will also allow other symbols, such as "$-$" or "$\geq$" which could be in turn defined on the basis of the above primitives, or even symbols for any function already shown to be exp-RA definable. The following is an example of a formula $\Phi(x, y)$ over L: $\forall z \left[e^{7z^2e^y} - \pi xz \geq \arctan(e^x)\right]$. We write $\Phi(x, y)$ to indicate the fact that the only free –i.e., non-quantified– variables in the formula are x and y. Each such formula will be interpreted over the real numbers, that is, all variables are assumed to take real values. Thus all quantifiers are implicitely assumed to be over $\mathbb{R}$. Given a formula Φ with free variables $x_1, \ldots, x_l$, we write $\mathcal{S}(\Phi)$ for the subset of $\mathbb{R}^l$ that it defines. For instance, the above $\Phi(x, y)$ gives rise to: $\mathcal{S}(\Phi)=$

$$\left\{(x, y) \in \mathbb{R}^2 \,|\, (\forall z \in \mathbb{R}) \left[e^{7z^2e^y} - \pi xz \geq \arctan(e^x)\right]\right\} .$$

Similarly, the truth of a formula Φ with no free variables is defined as the truth of the statement obtained when quantifying over the reals. A *definable* set is a set of the form $\mathcal{S}(\Phi)$, for some first order formula Φ over the language L. An exp-RA definable function is the same as one whose graph is definable in this sense.

When the exponential is left out, the definable sets are precisely those called "finitely subanalytic" in [24]. Restricted analytic functions were introduced in [26]. (The definition in that reference is slightly different from the one we gave in the previous section: it assumes that the functions g have a convergent power series representation valid on all of the cube $C = [-k, k]^l$, but a standard compactness argument shows that the two definitions are equivalent.) Van den Dries had shown in [24] that the theory of real numbers with restricted analytic functions is model-complete, which means roughly that every formula is equivalent to one that involves only existential quantification. (We do not give the precise definition here, as it is not needed for explaining the further material.) In a recent major development, Wilkie showed in [28, 29] that using exponentiation (but now leaving out the RA functions), model-completeness obtains as well. Finally, in [26] and [27], it was shown that the full theory (RA as well as exponentials) is model-complete, and hence order-minimal:

Fact. ([26], Theorem 6.9, and [27]) The theory of L is *order-minimal*, that is, for each formula Φ having just one free variable, $\mathcal{S}(\Phi)$ is a subset of $\mathbb{R}$ consisting of a finite union of intervals (possibly unbounded or just points).

The terminology arises from the fact that such finite unions are the smallest Boolean algebra of subsets that can be defined using order. The forthcoming book [25] by van den Dries deals in detail with order-minimal theories. Sets definable (in any dimension) for order-minimal theories admit finite cell decompositions into topological submanifolds, and are in every sense very small (for instance, unless of full dimension, there are directions along which every line intersects the set in at most finitely many points). When dealing with the language L, where the primitives denote analytic functions, one has a stronger result as well:

Proposition 4.2 Let S be a definable subset of $\mathbb{R}^q$. Then, either S contains an open subset or it is finitely analytically thin. □

This is a consequence of [26], Theorem 8.8, which shows that each definable subset is a finite union of "analytic cells" each of which is definable and definably-isomorphic to an Euclidean space. The definition of analytic cell in that paper implies that each such cell is an embedded analytic submanifold.

5 Proofs

We now prove the results.

5.1 VC Result

A critical ingredient that we need is provided by a recent paper by Laskowski. In order to make the application of the results in [12] easier to understand, we next re-express some of the definitions given there in the terminology used in this work. We will say that an *L-architecture with r parameters and m inputs* is a formula $\Phi(w_1, \ldots, w_r, x_1, \ldots, x_m)$ (we write often just $\Phi(w, x)$, using variables $w \in \mathbb{R}^r$ and $x \in \mathbb{R}^m$). For each fixed evaluation of the "weight" vector w, we may consider the binary function $\Phi_w : \mathbb{R}^m \to \{0, 1\}$ given by $\Phi_w(x) = 1 \iff \Phi(w, x)$ true . This defines a class of functions $\mathcal{F} = \{\Phi_w, w \in \mathbb{R}^m\}$. If $\mathcal{F}$ arises in this manner from a formula Φ, we write $\text{VC}(\Phi)$ for the VC dimension $\text{VC}(\mathcal{F})$.

One of the main results in [12] (page 383, first paragraph) shows that order-minimality of a theory, which essentially amounts to providing a finite VC dimension conclusion for subsets of $\mathbb{R}$, implies the same conclusion for *any* number of variables: If the theory of L is order-minimal, then $\text{VC}(\Phi) < \infty$ for every formula Φ. At the time when [12] was written, order-minimality was open for the theory of the language L of interest in this paper. This in turn has now been established, as remarked above. We can then conclude as follows:

Theorem 8 $\text{VC}(\Phi) < \infty$ *for every formula* Φ.

The proof of Theorem 1 is a trivial consequence of Theorem 8, as one can characterize the behavior of a feedforward architecture by an obvious formula.

Remark 5.1 The constructions in the model theory literature would result in tremendously large upper bounds for VC dimension. For the "single hidden layer" one-input architecture mentioned in Remark 3.3, one can easily obtain a bound that is exponential in the number k of units (just use the fact that a linear combination of exponentials in 1 variable cannot have more zeroes than number of terms, which is essentially the Descartes Rule of Signs). □

Remark 5.2 The result in Theorem 8 is a generalization of that proved in [20] for formulas involving only algebraic operations. More precisely, consider a formula Φ in the language of the real numbers with addition, multiplication, and order (no exponentiation). The proof in [20] that $\text{VC}(\Phi) < \infty$ for every formula Φ of this type is far easier than in the general case, and is sketched next. By the Tarski-Seidenberg theorem on quantifier elimination, the set defined by $\Phi(w, x)$, for each w, can also be defined by a propositional formula involving just terms of the type $\{x \mid P_i(w, x) > 0\}$, for some finite set of polynomials P_i (the formula and the P_i's depend only on Φ). As a formula, each "$P_i > 0$" defines a class with finite VC dimension, since the class of functions determined in this manner is a subset of a finite dimensional space of functions (the space of all polynomials of degree at most equal to the degree of P_i). Now the sets defined by the formulas $\Phi(w, x)$ can be obtained by a finite number of Boolean operations from the above sets, and this can be easily shown to preserve finite VC dimension. It was recently observed in [13, 6] that a far better result can be given in the algebraic case, resulting in an estimate of VC dimension which is polynomial on the size of the architecture (the second reference uses the Milnor bounds on the number of connected components of a semi-algebraic set.) □

5.2 The Interpolation Result

Note that a *finitely* analytically thin subset is nowhere dense (as it is a finite union of nowhere dense subsets). So this follows from Proposition 4.2:

Corollary 5.3 If S is a definable subset of $\mathbb{R}^q$, then either it has nonempty interior or it is nowhere dense.

Now Theorem 3 is easy to establish. Indeed, if $\mathcal{X}$ is a set that can be ε-I-shattered, then the image of the map in Equation (1) intersects $(-\varepsilon, \varepsilon)^s$ densely, for some $\varepsilon > 0$. By the above Corollary this image, being a definable set, must have nonempty interior. But the map is analytic, so then Sard's Theorem implies that its differential must have full rank s at some point. In particular, it must then be the case that $s \leq r$, establishing the result.

Note that the inequality $\text{ID}(\mathcal{A}) \leq r$ is trivial in the case of bounded weights, assuming only that Σ consists of smooth activations. That is, if one takes any class of funtions of the type $\{\beta_{\mathcal{A}}(w_0, \cdot), \|\omega_0\| \leq \gamma\}$, then the image of the map (1) (with domain $\|\omega\| \leq \gamma$) is compact, hence closed. Thus the image cannot intersect $(-\varepsilon, \varepsilon)^s$ densely unless it contains all of $(-\varepsilon, \varepsilon)^s$. Now Sard's theorem again provides the conclusion.

5.3 Teaching Dimension Bounds

Consider a feedforward network architecture $\mathcal{A}$, and the various relations "$\sim$" on weights. Fix an $w_0 \in \mathbb{R}^r$; we will characterize the teaching subsets of size $r+1$ for the weight w_0. Let: $\mathcal{W}_0 := \{w \in \mathbb{R}^r \mid w \not\sim w_0\}$. For each $w \in \mathcal{W}_0$ let: $\mathcal{B}(w) := \{x \mid x \in \mathbb{R}^m \text{ and } w \underset{x}{\sim} w_0\}$. If $w \in \mathcal{W}_0$, this set is a semianalytic subset of $\mathbb{R}^m$ of dimension at most $m-1$, as it is the set of zeroes of a nonzero analytic function, namely $\beta(w, x) - \beta(w_0, x)$.

Therefore the following subset of $\mathbb{R}^{m(r+1)}$: $\mathcal{T}(w) = \{(x_1, \ldots, x_{r+1}) \mid x_i \in \mathcal{B}(w)\ \forall i = 1, \ldots, r+1\} = \prod_{i=1}^{r+1} \mathcal{B}(w)$ has dimension at most $(m-1)(r+1)$. (Apply Proposition 4.1, Part 3.)

With $k := m(r+1)$, take the following subset of $\mathcal{W}_0 \times \mathbb{R}^k$: $\mathcal{G} := \{(w, x_1, \ldots, x_{r+1}) \mid w \in \mathcal{W}_0, x_i \in \mathcal{B}(w)\ \forall i = 1, \ldots, r+1\}$. Consider the projection $\pi_1 : \mathcal{W}_0 \times \mathbb{R}^k \to \mathcal{W}_0$ on the first r coordinates. For each $w \in \mathcal{W}_0$, $\pi^{-1}(w) \bigcap \mathcal{G} = \mathcal{T}(w)$ has dimension at most $(m-1)(r+1)$. Applying Proposition 4.1, Part 2, it follows that the subset $\mathcal{G}$ has dimension at most $r + (m-1)(r+1) = m(r+1) - 1$.

Now consider the projection π_2 of $\mathcal{G}$ on the last $m(r+1)$ coordinates. The image is exactly the set B consisting of those vectors $(x_1, \ldots, x_{r+1})$ which give rise to *non* teaching sets $\mathcal{X} = \{x_1, \ldots, x_{r+1}\}$

for w_0. But projections cannot increase dimension. (Apply Proposition 4.1, Part 1, with $f = \pi_2$.) Therefore the set B has dimension $\leq m(r+1)-1$, as desired for the first part of Theorem 4. Observe that when Σ consists of exp-RA definable functions, the set B is definable, so from Proposition 4.2 and the above dimension count it follows that B is finitely analytically thin.

Proving the existence of universal teaching sets of cardinality $2r+1$, and in fact that almost all sets of that cardinality are universal teaching sets, is now easy. Indeed, consider a new architecture with weight space $\mathbb{R}^{2r}$ and such that the new behavior satisfies $\beta'((w_1, w_2), x) := \beta(w_1, x) - \beta(w_2, x)$. Fix any w_0. Then, any teaching set for (w_0, w_0), that is, a teaching set for the identically zero function, is a universal teaching set. Since the parameter space is now of dimension $2r$, the result follows.

Remark 5.4 It can be shown by examples that the bounds are best possible. Also, for smooth, rather than analytic, activations, a local result is possible: there is a dense open subset of $\mathbb{R}^r$, and an open covering of this set, so that on each subset V of this cover, some set of r inputs serves as a universal teaching set with respect to weights on V. □

We now turn to the proof of Theorem 5. Now there is no algebraic structure on the input space, as we are restricted to working with the subset $\mathcal{I}$ of $\mathbb{R}^m$. Thus we can only exploit the dependence on weights. For each $x \in \mathcal{I}$, consider: $\mathcal{V}(x) := \{w \in \mathbb{R}^r \mid w \underset{x}{\sim} w_0\}$. The problem is simply to show that there is some finite subset $\mathcal{X} = \{x_1, \ldots, x_l\} \subset \mathcal{I}$ so that $\bigcap_{i=1}^{l} \mathcal{V}(x_i) = \bigcap_{x \in \mathcal{I}} \mathcal{V}(x)$. All we need for this is a descending chain condition (DCC) on sets obtained as finite intersections of sets of the type $\mathcal{V}(x)$, which are definable.

To get this, we need to use some results from Tougeron's [23], which gives various sufficient conditions for DCC to hold. In particular, the first sentence of the proof of 3.5 gives a condition phrased in terms of finiteness of connected components for regular points of varieties, which together with dimensionality counts is sufficient for our purposes. Details are omitted.

5.4 Decidability

Theorem 6 is immediate from the fact that, subject to a positive answer to conjecture (SC), there is a decision procedure for determining the truth of any formula Φ in the language introduced earlier, if exponentials are used (but not RA functions). See [27] for this recent result,

We note also that the *loading problem*, that is, the problem of determining, for a given set of data and architecture, if there is a network interpolating at the given data, is also decidable, modulo (SC), for the same reasons.

Observe that the use of other analytic functions may lead to undecidability. For instance, if $\sin(x)$ is used instead of σ_s, one may easily encode integers into the loading problem (by asking that certail values be zero), and hence the solution of diophantine equations can be reduced to this question.

5.5 The Sparse Polynomials Result

The proof of Theorem 7 is a simple consequence of the previous material. Consider the class of all functions given by exponential polynomials $P(e^{y_1}, \ldots, e^{y_m})$, where P is a polynomial having at most r terms. As one can write a formula for these functions, over the language L —use as parameters the coefficients of the monomials and the exponents— this has finite VC dimension, let's say κ. We claim that the VC dimension of $\mathcal{P}_{m,r}$ is at most $3^m\kappa$.

This is proved as follows. For each real number x, let $\operatorname{sign} x$ be zero if $x = 0$ and $x/|x|$ otherwise. For a vector $x \in \mathbb{R}^m$, let $\operatorname{sign} x$ be the vector of signs of its coordinates. Assume that there would be some $X \subset \mathbb{R}^m$ of cardinality $3^m\kappa + 1$ which can be shattered by $\mathcal{P}_{m,r}$. Then there is some subset X' of X of cardinality $\kappa + 1$, which consists of vectors all having the same sign, $(\varepsilon_1, \ldots, \varepsilon_m)$, and of course this set can still be shattered by $\mathcal{P}_{m,r}$. Now consider a subset $Y \subset \mathbb{R}^m$ of cardinality $\kappa + 1$

so that, for each $x = (x_1, \ldots, x_m)$ in X', there is some $y = (y_1, \ldots, y_m)$ in Y with $x_i = \varepsilon_i e^{y_i}$ for each i. This substitution maps polynomials into exponential polynomials, so Y can be shattered by exponential polynomials, contradicting the definition of κ.

6 Counterexamples

We now show by means of counterexamples that none of the hypotheses in Corollary 3.1 can be dropped. If the activation is merely infinitely differentiable, even with bounded inputs and weights, or analytic with either unbounded inputs or unbounded weights, the VC dimension is infinite. Moreover, the counterexamples are in terms of the single-hidden layer architectures standard in neural nets research (see Equation 2) and σ is a squashing function (strictly increasing and bounded), a qualitative property which is usually imposed on activations.

Assume that $\alpha : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and satisfies the following properties: $\alpha \in L^1(\mathbb{R})$, $\alpha(x) > 0$ $\forall x$, α is even, and $|\alpha'(x)| \leq c\alpha(x)$ $\forall x$. Assume that $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and satisfies the following properties: f and f' are bounded and f is even. Define $\beta(x) := \int_0^x \alpha(t)dt$. Observe that this is an odd function. Now let $g := K\beta + \alpha f$, where K is a constant so that $c|f(x)| + |f'(x)| < K$ for all x. By construction, g is bounded. We claim that g is strictly increasing as well. Indeed, $|\alpha'(x)f(x) + \alpha(x)f'(x)| < K\alpha(x)$ for all x, and hence $g > 0$ everywhere. So g is a squashing function. Now consider the 2-unit one hidden layer architecture with activation $\sigma = g$, $r = 1$, and $a_0 = 0$, $a_1 = a_2 = 1$, $c_1 = c_2 = 1$, and $b_1 = b_2 = w$. (There is only one programmable weight, but of course the result to be proved, infinite VC dimension, will remain true if all coefficients in Equation 2 are taken as weights.) The behavior is $\beta(w, x) = g(wx) + g(-wx) = 2\alpha(wx)f(wx)$. Since α is everywhere positive, the dichotomies implemented by this architecture are precisely the same as those implemented by the set of functions $\{f(w \cdot (\cdot)), w \in \mathbb{R}\}$. Thus if there exists for each integer s a set of s real numbers $x_i, i = 1, \ldots, s$ and weight choices w_j, $j = 1, \ldots, 2^s$ so that the matrix $\text{sign}(f(w_j x_i))$ has all its 2^s columns of distinct signs, the same is true of $\beta(w_j, x_i)$. An example is furnished by $f(x) = \cos(x)$, and $\alpha = \frac{1}{1+x^2}$. This shows that arbitrary (not exp-RA definable) analytic functions may result in architectures with infinite VC dimension. (Moreover, the architecture used is the simplest one that appears in neural nets practice.)

Note that if we wish the x_i's to be bounded, for instance to be restricted to the interval $[-1, 1]$, one may replace the above x_i's and w_j's by $\frac{x_i}{c}$ and cw_j, where $c = \sum |x_i|$. Similarly, if one wants to restrict the weights w_j to be bounded, one can use cx_i and $\frac{w_j}{c}$, with $c = \sum |w_j|$. Thus bounded weights or inputs (but not simultaneously), even with analytic activations, do not suffice.

Finally, consider a function f as above, and let $\rho : \mathbb{R} \to \mathbb{R}$ be 0 at 0 and equal to e^{-1/x^2} elsewhere. Write $h(x) := \rho(x)f(\frac{1}{x})$. Note that h is even, bounded, has bounded derivative, and is smooth. Thus h can be used in the above constructions instead of f; consider the architecture that results. Now, if $x_i > 1$, $i = 1, \ldots, s$ and $w_j > 1, j = 1, \ldots, 2^s$ are so that the original matrix $f(w_j x_i)$ had all columns of distinct signs, the same is true of $\beta(\hat{w}_j, \hat{x}_i)$ with the new "f", where $\hat{w}_j = 1/w_j$ and $\hat{x}_i = 1/x_i$. The inputs and weights are now all in the interval $(0, 1)$. Starting with $f = \cos$, this illustates that even with bounded weights and inputs, infinite differentiability is not sufficient to guarantee finite VC dimension.

References

[1] Baum, E.B., and D. Haussler, "What size net gives valid generalization?," *Neural Computation* **1** (1989): 151-160.

[2] Bierstone, E., and Pierre D. Milman, "Semianalytic and subanalytic sets," *Inst. Hautes Études Sci. Publ. Math.* **67**(1988): 5-42.

[3] Blum, A., and R.L. Rivest, "Training a 3-node neural network is NP-complete," in *Advances in Neural Information Processing Systems 2* (D.S. Touretzky, ed), Morgan Kaufmann, San Mateo, CA, 1990, pp. 9-18.

[4] DasGupta, D., and G. Schnitger, "The power of approximating: a comparison of activation functions," in *Advances in Neural Information Processing Systems 5* (Giles, C.L., Hanson, S.J., and Cowan, J.D., eds), Morgan Kaufmann, San Mateo, CA, 1993, to appear.

[5] Durbin, R., and D.E. Rumelhart, "Product units: a computationally powerful and biologically plausible extension to backpropagation networks", *Neural Computation* **1** (1989): 133-142.

[6] Goldberg, P., and M. Jerrum, "Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers," submitted.

[7] Goldman, S.A., and M.J. Kearns, "On the complexity of teaching," *Proc. Forth ACM Workshop on Computational Learning Theory*, July 1991, pp. 303-314.

[8] Haussler, D., "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. and Computation* **100**(1992): 78-150.

[9] Hertz, J., A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, 1991.

[10] Karpinski, M., and T. Werther, "VC dimension and uniform learnability of sparse polynomials and rational functions," *SIAM J. Computing*, to appear. (Preprint 8537-CS, Bonn, 1989.)

[11] Khovanskii, A.G., *Fewnomials*, American Mathematical Society, Providence, R.I., 1991.

[12] Laskowski, M.C., "Vapnik-Chervonenkis classes of definable sets," *J. London Math. Soc.* **2 45**(1992): 377-384.

[13] Maass, W.G., "Bounds for the computational power and learning complexity of analog neural nets," *Proc. of the 25th ACM Symp. Theory of Computing*, 1993.

[14] Maass W., G. Schnitger, and E.D. Sontag, "On the computational power of sigmoid versus Boolean threshold circuits," in *Proc. 32nd IEEE Symp. Foundations of Comp. Sci.*, 1991: 767-776.

[15] Macintyre, A.J., and A.J. Wilkie, "Schanuel's conjecture implies the decidability of real exponentiation," handwritten, Oxford University, 1992.

[16] Siegelmann, H.T., and E.D. Sontag, "On the computational power of neural nets," in *Proc. Fifth ACM Workshop on Computational Learning Theory*, Pittsburgh, July 1992, 440-449.

[17] Siegelmann, H.T., and E.D. Sontag, "Analog computation, neural networks, and circuits," submitted.

[18] Sontag, E.D., "Feedforward nets for interpolation and classification," *J. Comp. Syst. Sci.* **45**(1992): 20-48.

[19] Sontag, E.D., *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer, New York, 1990.

[20] Stengle, G. and Yukich, J.E., "Some new Vapnik-Chervonenkis classes," *The Annals of Statistics* **14** (1989): 1441-1446.

[21] Sussmann, H.J., "Uniqueness of the weights for minimal feedforward nets with a given input-output map," *Neural Networks* **5**(1992): 589-593.

[22] Sussmann, H.J., "Real analytic desingularization and subanalytic sets: An elementary approach," *Trans. Amer. Math. Soc.* **317**(1990): 417-461.

[23] Tougeron, J.C., "Sur certaines algebres de fonctions analytiques", *Seminaire sur la geometrie algebrique reelle*, Tome I, II Publ. Math. Univ. Paris VII **24** (1986): 35-121.

[24] van den Dries, L., "A generalization of the Tarski-Seidenberg theorem, and some nondefinability results," *Bull. AMS* **15**(1986): 189-193.

[25] van den Dries, L., " Tame topology and 0-minimal structures", preprint, University of Illinois, Urbana, 1991-2.

[26] van den Dries, L., and C. Miller, "On the real exponential field with restricted analytic functions," *Israel J. Math.*, to appear.

[27] van den Dries, L., A. Macintyre, and D. Marker, "The elementary theory of restricted analytic fields with exponentiation," *Annals of Math.*, to appear.

[28] Wilkie, A.J., "Some model completeness results for expansions of the ordered field of reals by Pfaffian functions," preprint, Oxford, 1991, submitted.

[29] Wilkie, A.J., "Smooth 0-minimal theories and the model completeness of the real exponential field," preprint, Oxford, 1991, submitted.