Polynomial Bounds for VC Dimension of Sigmoidal Neural Networks

Marek Karpinski *
Dept. of Computer Science
University of Bonn
53117 Bonn

Angus Macintyre †
Mathematical Institute
University of Oxford
Oxford OX1 3LB

Abstract. We introduce a new method for proving explicit upper bounds on the VC Dimension of general functional basis networks, and prove as an application, for the first time, the VC Dimension of analog neural networks with the sigmoid activation function $\sigma(y) = 1/1+e^{-y}$ to be bounded by a quadratic polynomial in the number of programmable parameters.

0 Introduction

The most commonly used activation function in various neural networks applications is the sigmoid $\sigma(y) = 1/1 + e^{-y}$ (cf. [HKP91]). We refer to [AB92], [M93], and [MS93] for all the necessary background on the computation by neural networks and the VC dimension (particularly, to the connection between their computational power, and the sample complexity). In Maass's 1993 lecture notes [M93], Open Problem 10 (see also [GJ93] and [MS93]) asks:

Is the VC-dimension of analog neural nets with the sigmoid activation function $\sigma(y) = 1/1 + e^{-y}$ bounded by a polynomial in the number of programmable parameters? (In [MS93] the *finiteness* of VC Dimension of sigmoidal neural networks has been established for the first time. The *explicit upper bounds* for VC dimension however, and consequently, the bounds on the sample sizes of single sigmoidal neural networks , remained an open problem.)

In this paper we give an affirmative answer, with a quadratic polynomial bound in a number of programmable parameters. We believe that the bound can be improved to the subquadratic one in the number of programmable parameters using a variant of our method. The details are given in Sections 2–3. The result is a special case of a much more general result about bounds for VC dimension in o-minimal theories.

The paper was inspired by the work of Goldberg and Jerrum [GJ93], who could deal with polynomial activation functions. A reference in [GJ93], to Warren's paper [W68], was of particular importance. Some other applications of our method will appear in the full version of this paper.

1 Model-theoretic Preliminaries

We shall consider a standard model of a *feedforward network architecture* A with the *activation* function σ (cf., e. g., [M93], [MS93]) with k *inputs*, m *computational nodes*, and ℓ *weights* (the number of *programmable parameters*). We assume that the output gate of A is thresholded to $\{0, 1\}$. We associate with A an exponentional formula $\Phi(\bar{v}, \tilde{y}) > 0$ for $\bar{v} \in \mathbf{R}^k$, and $\tilde{y} \in \mathbf{R}^\ell$ being a composition of polynomials, and activation functions over the computation nodes of A. $\Phi(\bar{v}, \tilde{y}) > 0$ represents the function computed by A. Alternatively, and this is crucial in our paper, we describe the computation of A as a Boolean combination of atomic formulas of two forms $\tau(\bar{v}, \tilde{y}) = 0$ or $\tau(\bar{v}, \tilde{y}) > 0$ describing local computations of A at its computational nodes (for appropriate $\bar{v}$'s, and $\tilde{y}$'s). The *VC dimension* of the *network* A is the *VC dimension* of the class $\mathcal{C}_\Phi = \{\Phi_{\tilde{\beta}} : \tilde{\beta} \in \mathbf{R}^\ell\}$ for $\Phi_{\tilde{\beta}} = \{\bar{x} \in \mathbf{R}^k : \Phi(\bar{x}, \tilde{\beta}) > 0\}$ the partition of $\mathbf{R}^k$ by A according to the weight assignment $\tilde{\beta}$. (The general reader is referred to [MS93] and [GJ93] for definitions and basic properties of Vapnik-Chervonenkis (VC) dimension. We call a set $S \subseteq \mathbf{R}^k$ to be *shattered* by $\mathcal{C}_\Phi$ if $\{S \bigcap C : C \in \mathcal{C}_\Phi\} = P(S)$. The

*Research partially supported by the International Computer Science Institute, Berkeley, by the DFG Grant KA 673/4-1, and by the ESPRIT BR Grants 7097 and ECUS 030. Email: marek@cs.uni-bonn.de

†Research supported in part by a Senior Research Fellowship of the SERC. Email: ajm@maths.ox.ac.uk

VC dimension of $\mathcal{C}_\Phi$ is the maximal size of any set S that can be shattered by $\mathcal{C}_\Phi$, or ∞ if arbitrary large subsets may be shattered.)

We turn our attention now to the analysis of general formulas resulting from the local computation descriptions of A. The method of our analysis is by no means restricted to the network architectures only, and can be applied to a much larger class of formulas, which could be of independent interest.

The principles behind this paper are of great generality. We do not seek here maximum generality, but restrict ourselves to work over the field of real numbers.

We work with structure M which are enrichments of the real field $\mathbf{R}$ by certain total C^∞ (infinitely differentiable) functions. Our underlying first-order language L has primitives $+, -, \cdot, <, 0, 1$ (with the usual interpretation on $\mathbf{R}$), together with various n-ary function symbols f. Each f has a fixed interpretation by a C^∞ function $\bar{f} : \mathbf{R}^n \to \mathbf{R}$, thereby determining an L-structure M. Obviously, if $\tau(v_1, \cdots, v_m)$ is an L-term with free variables $v_1, \cdots, v_m$, τ defines an m-ary C^∞ function (also denoted $\bar{\tau}$) from $\mathbf{R}^m$ to $\mathbf{R}$. L-formulas $\Phi(v_1, \cdots, v_k)$ define subsets of $\mathbf{R}^k$, and L-formulas $\Phi(v_1, \cdots v_k,\ y_1, \cdots, y_\ell)$ together with $\tilde{\beta} = (\beta_1, \cdots, \beta_\ell)$ in $\mathbf{R}^\ell$ define subsets of $\mathbf{R}^k$, namely

$$\Phi_{\tilde{\beta}} = \{\bar{x} \in \mathbf{R}^k : M \models \Phi(\bar{x}, \tilde{\beta})\}.$$

For $\Phi(\bar{v}, \tilde{y})$ as above, let

$$\{\Phi_{\tilde{\beta}} : \tilde{\beta} \in \mathbf{R}^\ell\} = \mathcal{C}_\Phi.$$

$\mathcal{C}_\Phi$ is a definable family of definable sets. In this paper we will give good bounds for the VC dimension of $\mathcal{C}_\Phi$, for many natural Φ.

The following notion has in the last decade become central in the model theory of analysis [L92].

Definition. M is o-minimal if for every formula $\Phi(v_1, y_1, \cdots y_\ell)$ and every $\tilde{\beta} \in M^\ell$, $\Phi_{\tilde{\beta}}$ is a finite union of intervals with endpoints in $M \cup \{\pm\infty\}$.

Notes:

(a) "Interval" should be understood in all possible senses.

(b) It is not important that M lives on $\mathbf{R}$ or has C^∞-primitives.

(c) It follows, nontrivially, that the number of connected components of $\Phi_{\tilde{\beta}}$ above is bounded independent of $\tilde{\beta}$.

For our purposes we need two substantial results about o-minimality:

Theorem 1. If M is an o-minimal expansion of $\mathbf{R}$, then for any $\Phi(v_1, \cdots, v_k, y_1, \cdots, y_\ell)$, and any $\tilde{\beta}$, $\Phi_{\tilde{\beta}}$ has only finitely many connected components, and there is a bound $B(\Phi)$ independent of $\tilde{\beta}$.

Theorem 2. If M is an o-minimal structure then for every $\Phi(\bar{v}, \tilde{y})$, $\mathcal{C}_\Phi$ has finite VC-dimension.

For Theorem 1, see [D92], and [KPS86] for Theorem 2, see [L92]. Note that although the latter has constructive aspects, the use of Ramsey's theorem precludes realistic estimates for VC dimension of $\mathcal{C}_\Phi$.

2 The Main Result

2.1 Let M be a structure as above with enrichments of the real field $\mathbf{R}$, with C^∞ primitives, and o-minimal. $\Phi(v_1, \cdots, v_k,\ y_1, \cdots, y_\ell)$ is now assumed to be a *quantifier-free* formula. Thus Φ is a Boolean combination of atomic formulas, which can be of two forms:

$$\tau(\bar{v}, \tilde{y}) > 0,$$

or

$$\tau(\bar{v}, \tilde{y}) = 0$$

where τ is a term.

Now list as $\tau_1, \cdots, \tau_h$ the terms as above occurring in Φ. (One can delete repetitions). If we fix $\bar{\alpha} \in \mathbf{R}^k$, $\tilde{\beta} \in \mathbf{R}^\ell$ we get a sequence

$$\langle \tau_1(\bar{\alpha}, \tilde{\beta}), \cdots, \tau_h(\bar{\alpha}, \tilde{\beta}) \rangle$$

of reals, inducing a sequence of signs +, 0, -, via

$$\begin{aligned} \operatorname{sgn} \tau_i &= + \quad \text{if} \quad \tau_i(\bar{\alpha}, \tilde{\beta}) > 0 \\ &= 0 \quad \text{if} \quad \tau_i(\bar{\alpha}, \tilde{\beta}) = 0 \\ &= - \quad \text{if} \quad \tau_1(\bar{\alpha}, \tilde{\beta}) < 0. \end{aligned}$$

Call this sequence $\sigma(\bar{\alpha}, \tilde{\beta})$. Consider first, for fixed $\bar{\alpha}$, the $\tilde{\beta}$ such that $\sigma(\bar{\alpha}, \tilde{\beta})$ consists only of + and - 's. Then as $\tilde{\beta}$ varies one gets only finitely many $\sigma(\bar{\alpha}, \tilde{\beta})$, with a bound for the number being given by the number of connected components of

$$\mathbf{R}^\ell \setminus \cup_{i \le h} \{\tilde{y} : \tau_i(\bar{\alpha}, \tilde{y}) = 0\}.$$

[By Theorem 1, this number is finite, and has a bound independent of $\bar{\alpha}$].

To handle general $\sigma(\bar{\alpha}, \tilde{\beta})$ one uses a variational argument (Corollary 2.1 in [GJ93]) which is everywhere dense in what follows. Add a new variable ε, and replace $\langle \tau_i(\bar{x}, \tilde{y}) \rangle_{i \le h}$ by

$$\begin{aligned} &\langle \tau_1(\bar{x}, \tilde{y}) + \varepsilon, \tau_1(\bar{x}, \tilde{y}) - \varepsilon, \\ &\tau_2(\bar{x}, \tilde{y}) + \varepsilon, \cdots, -\varepsilon, \\ &\tau_h(\bar{x}, \tilde{y}) + \varepsilon, \tau_h(\bar{x}, \tilde{y}) - \varepsilon \rangle \end{aligned}$$

So we have replaced an h-tuple of terms in $k+\ell$ variables by a $(2h)$- tuple of terms in $k+\ell+1$ variables. In this way we get new sign sequences $\sigma^*(\bar\alpha, \tilde\beta, \varepsilon)$. The basic lemma is:

Lemma 3. For fixed $\bar\alpha$ the number of $\sigma(\bar\alpha, \tilde\beta)$ is bounded by the number of $\sigma^*(\bar\alpha, \tilde\beta, \varepsilon)$ consisting only of $+$ and $-$.

Let $\sigma(\bar\alpha, \tilde\beta_1), \cdots, \sigma(\bar\alpha, \tilde\beta_r)$ be the distinct $\sigma(\bar\alpha, \tilde\beta)$. Choose $\varepsilon > 0$ but $<$ all $\mid \tau_j(\bar\alpha, \tilde\beta_i) \mid (j \leq h, i \leq r)$ which are non zero. Then $\sigma^*(\bar\alpha, \tilde\beta_i, \varepsilon)$ has no zeros, and clearly $\sigma^*(\bar\alpha, \tilde\beta_i, \varepsilon) \neq \sigma^*(\bar\alpha, \tilde\beta_j, \varepsilon)$ if $\sigma(\bar\alpha, \tilde\beta_i) \neq \sigma(\bar\alpha, \tilde\beta_j)$. $\square$

Note: The essential point for future reference is that the number $\sigma(\bar\alpha, \tilde\beta)$ is bounded by the number of connected components of

$$\begin{aligned} \mathbf{R}^{\ell+1} \setminus \bigcup_{i\leq h} (\{(\tilde y, \varepsilon) : \tau_i(\bar\alpha, \tilde y) &= \varepsilon\} \\ \cup \{(\tilde y, \varepsilon) : \tau_i(\bar\alpha, \tilde y) &= -\varepsilon\}), \end{aligned}$$

and this has a bound independent of $\bar\alpha$.

2.2 Now we run through the argument of [GJ93]. Let $\Phi(v_1, \cdots, v_k,\ y_1, \cdots, y_\ell)$ be quantifier-free with terms $\tau_i(\bar v, \tilde y),\ i \leq s$.

Let $\{\bar\alpha_1, \cdots, \bar\alpha_v\}$ be distinct elements of $\mathbf{R}^k$ such that $\{\bar\alpha_1, \cdots \bar\alpha_v\}$ is shattered by $\mathcal{C}_\Phi$. Then exactly as in [GJ93] one sees:

$2^v \leq$ the number of sequences of signs $\{+, -, 0\}$ obtainable from

$$\langle \tau_1(\bar\alpha_1, \tilde y), \tau_1(\bar\alpha_2, \tilde y), \cdots, \tau_1(\bar\alpha_v, \tilde y), \tau_2(\bar\alpha_1, \tilde y), \cdots$$

$$\cdots \tau_1(\bar\alpha_v, \tilde y), \cdots, \tau_s(\bar\alpha_1, \tilde y), \cdots, \tau_s(\bar\alpha_v, \tilde y)\rangle.$$

Note that the latter sequence has length vs.

Then by the argument in 2.1, $2^v \leq$ number of connected components of

$$\mathbf{R}^{\ell+1} \setminus \bigcup_{\substack{i\leq s \\ j\leq v}} \{(\tilde y, \varepsilon) : \tau_i(\tilde\alpha_j, \tilde y) \pm \varepsilon\} \quad (*).$$

Our strategy is to use o-minimality to get a decent estimate for the right hand side.

2.3 We proceed axiomatically. We assume we have fixed a bound $\Gamma(\mu, m)$ for integers m and sequences $\mu = \langle \mu_i \rangle_{i\leq r}$ of terms

$$\mu_i(v_1, \cdots, v_k,\ y_1, \cdots, y_\ell) \qquad , \qquad i \leq r$$

for the number of connected components of

$$\bigcup_{j\leq m} \{\tilde y : \mu_{f(j)}(\bar\alpha_j, \tilde y) = 0\}$$

as $\bar\alpha_j$ varies through $(\mathbf{R}^k)^m$, and f is a function from $[O, m]$ to $[O, r]$.

Classical example. μ a polynomial of $\tilde y$ degree $\leq d$. Then $\Gamma(\mu, m)$ can be taken as $2.(2d)^\ell$.

This is due to Milnor [M64]. For $m = 1$ one has the bound $2.d^\ell$, and the general case reduces to this by replacing μ by $\sum_{j\leq m} \mu(\bar\alpha_j, \tilde y)^2$.

We shall see later the exponential analogue. In an o-minimal theory, Γ of course exists.

We show in 2.4 how the right hand side of (*) may be estimated in terms of Γ. The idea comes from Warren's 1968 paper [W68], and we now use the C^∞ property of M for the first time.

2.4 We assume given terms $\mu_1(\bar v, \tilde y), \cdots, \mu_n(\bar v, \tilde y)$ and $\bar\alpha_1, \cdots, \bar\alpha_n \in \mathbf{R}^k$.

We consider the definition

$$\{\bar y : \bigwedge_{i\leq n} \mu_i(\bar\alpha_i, \tilde y) = 0\} \qquad \text{in} \qquad \mathbf{R}^\ell,$$

and say it is *nonsingular* if either it defines ϕ, or at each point $(y_1, \cdots, y_\ell)$ in the above intersection the Jacobian matrix

$$\left| \frac{\partial}{\partial y_j} \mu_i(\bar\alpha_i, \tilde y) \right|_{\substack{i\leq n \\ j\leq \ell}}$$

has rank n.

Of course in the latter situation the implicit function theorem applies if $\ell > n$, and we have a C^∞ manifold of dimension $\ell - n$. If $\ell = n$, o-minimality gives a bound, depending only on $\mu_1, \cdots, \mu_n$, for the cardinality of the set defined.

When $\ell > n$, there are variables $y_{m_1}, \cdots, y_{m_{\ell-n}}$ such that locally all other y_i are given as $F_i(y_{m_1}, \cdots, y_{m_{\ell-n}})$ where F_i is a definable (from the $\bar\alpha_i$) C^∞ function.

We want to take a naive, presentation-sensitive, notion of submanifold, or, better, locally flat submanifold. We assume a presentation as above, with $y_{m1}, \cdots, y_{m_{\ell-n}}$ specified, and suppose that we add some new equations $\mu_{n+1} = 0, \cdots, \mu_{n+u} = 0$ giving a manifold of dimension $\ell - (n+u)$ with a subset of $\{y_{m_1}, \cdots y_{m_{\ell-n}}\}$ as its *basis*.
This gives notion of locally flat. Thus $\mathcal{M}_1 \to \mathcal{M}_2$ is locally like the canonical $\mathbf{R}^{\ell-(n+u)} \to \mathbf{R}^{\ell-n}$ which puts 0's on all but first $\ell - (n+u)$ entries.

We will be able to use Warren's Theorem 1 exactly as he does. For convenience we repeat it:

Theorem 4. Let $\mathcal{M}$ be a connected topological n-manifold, and let $M_1, \cdots, M_n$ be connected $(n-1)$-manifolds embedded in M so that:

(1) The M_i are topologically closed and locally flat in M;

(2) The intersection of any given j of the $M_i, 1 \leq j \leq n$, is either empty or is an $(n-j)$ manifold locally flat in the intersection of any $(j-1)$ of the M_i, and

(3) any intersection of more than n of the M_i is empty.

Let b_j be the number of connected components among all intersections of any j of the M_i with M.

Then $M - \cup_{i=1}^{n} M_i$ has $\leq \Sigma b_j$ connected components.

Regular Configurations. If we are given a sequence $\mu_i(\bar{\alpha}_i, \tilde{y}) 1 \leq i \leq n$, of terms in parameters, we say they form a *regular* sequence if every formal intersection of a subset is nonsingular in the sense explained earlier. In that case, if one takes as $M_1, \cdots, M_b$ the connected components of the $\{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = 0\}$, the hypotheses of Theorem 2 are satisfied, with $M = \mathbf{R}^\ell$. Connectedness is clear, and local flatness is direct. That the intersection of more than ℓ of the M_i is empty follows from regularity and a dimension count. The finiteness of components follows from o-minimality, since each original component is definable ([D92], [KPS86]).

And now we come to the crunch, which reduces all calculations, via small perturbations, to ones covered by Theorem 4. This corresponds to Warren's Lemmas 2.2, 2.3 and 2.4. We have to change his argument for 2.2, which appeals to complex projective geometry. We use instead Sard's Theorem [M65].

Lemma 5. Let $\mu_1(\bar{\alpha}_1, \tilde{y}), \cdots, \mu_m(\bar{\alpha}_m, \tilde{y})$ be as usual, and $\tau(\bar{\alpha}, \tilde{y})$ arbitrary. If the definition

$$V = \cap \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = 0\}$$

is regular, or defines the set $\mathbf{R}^\ell$, then for all but finitely many real numbers λ the definition

$$V \cap \{\tilde{y} : \tau(\bar{\alpha}, \tilde{y}) - \lambda = 0\}$$

is regular.

Proof. If $m > \ell, V = \phi$, and there is nothing to prove.

If $m = \ell, V$ is finite, and so choose λ outside the range of $\tau(\bar{\alpha}, \tilde{y})$ on V.

If $m < \ell$, or $V = \mathbf{R}^\ell$, we have the C^∞ map $\tau(\bar{\alpha}, \tilde{y})$ from V to $\mathbf{R}$, and by Sard's Theorem and o-minimality the set of regular values of τ is cofinite. Any regular value λ works. □

The rest is almost formal.

Lemma 6. (Same notation as Lemma 3).

There exists $\delta > 0$ such that if $0 < \varepsilon < \delta$ then every connected component of

$$E = \mathbf{R}^\ell - (V \cup \{\tilde{y} : \tau(\bar{\alpha}, \tilde{y}) = 0\})$$

contains a connected component of one of the sets

$$E_\varepsilon = \mathbf{R}^\ell - (V \quad \cup \quad \{\tilde{y} : \tau(\bar{\alpha}, \tilde{y}) = \varepsilon\} \quad \cup \quad \{\tilde{y} : \tau(\bar{\alpha}, \tilde{y}) = -\varepsilon\})$$

Proof: E has finitely many connected components, by o-minimality. Let them be $C_\gamma (\gamma < \gamma_o)$, and pick

$$\tilde{c}_\gamma \in C_\gamma$$

Let

$$\delta = \min_{\gamma < \gamma_o} \mid \tau(\bar{\alpha}, \tilde{c}_\gamma) \mid > 0.$$

C_γ is contained in some component J of $\mathbf{R}^\ell \setminus V$, and is the maximal connected set of points $\tilde{y}$ in J containing $\tilde{c}_\gamma$ and such that

$$\operatorname{sgn} \tau(\bar{\alpha}, \tilde{y}) = \operatorname{sgn} \tau(\bar{\alpha}, \tilde{c}_\gamma).$$

If $\varepsilon < \delta$, some component K of $J \cup \{\tilde{y} : |\tau(\bar{\alpha}, \tilde{y})| > \varepsilon\}$ contains $\tilde{c}_\gamma$. This is a component of E_ε and is contained in C_γ. □

Lemma 7. Let $\mu_1, \cdots, \mu_m$ be as usual. Then there are real numbers $\varepsilon_1, \cdots, \varepsilon_m$ such that the collection

$$\mu_1 - \varepsilon_1, \mu_1 + \varepsilon_1, \mu_2 - \varepsilon_2, \mu_2 + \varepsilon_2, ...$$

form a regular configuration, and such that every connected component of

$$E = \mathbf{R}^\ell \setminus \bigcup \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = 0\}$$

contains one of

$$F = \mathbf{R}^\ell \setminus \bigcup \tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = \pm\varepsilon_i\}.$$

Proof: By recursion on m.

First choose β as in Lemma 6, with $\mu_2, \cdots, \mu_m$ as the μ's, and μ_1 as the τ. By Lemma 5, for all but finitely many λ, $\{\tilde{y} : \mu_1 - \lambda = 0\}$ is regular. So choose $0 < \varepsilon_1 < \beta$ so both

$$\{\tilde{y} : \mu_1 = \varepsilon_1\} \qquad \text{and} \qquad \{\tilde{y} : \mu_1 = -\varepsilon_1\}$$

are regular. (Of course their intersection is empty).

Assume $\varepsilon_1, \cdots, \varepsilon_r$ have been determined. Now choose β as in Lemma 6, with μ_{r+1} as τ and $\mu_i \pm \varepsilon_i$, $i = 1, \cdots, r$ and $\mu_{r+2}, \cdots, \mu_n$ as the μ's.
Choose $\varepsilon_{r+1} \in (0, \beta)$ so that the

$$\{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) - \varepsilon_{r+1} = 0\}$$

and

$$\{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) + \varepsilon_{r+1} = 0\}, \qquad i \leq r+1$$

form a regular configuration. This is possible by Lemma 5, because the number of intersections to consider is finite, and all preceding intersections are regular. □

2.4 Recall the bound Γ from 2.3.

Theorem 6. Let $\mu_1(\bar{\alpha}_1, \tilde{y}), \cdots, \mu_m(\bar{\alpha}_m, \tilde{y})$ be as usual. Then the number of connected components of

$$\mathbf{R}^\ell \setminus \bigcup \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = 0\}$$

is bounded by

$$\sum_{j+0}^{\ell} C_{m,j} \cdot 2^j \cdot \Gamma(\mu_1^+, \mu_1^-, \cdots, \mu_m^+, \mu_m^-, j)$$

where $C_{m,j}$ is $\binom{m}{j}$ if $j \leq m$, and $= 0$ otherwise, and

$$\mu_i^+(v_1, \cdots, v_k,\ w_1, \cdots, w_m, \tilde{y})$$

$$= \mu_i(\bar{v}, \tilde{y}) - w_i$$

and

$$\mu_i^-(v_1, \cdots v_k,\ w_1, \cdots, w_m, \tilde{y}) = \mu_1(\bar{v}, \tilde{y}) + w_i.$$

Proof: By Lemma 5 it suffices to bound the number of connected components of

$$\mathbf{R}^\ell \setminus \cup \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = \pm\varepsilon_i\}$$

for small ε, in the case of a regular configuration. So Warren's result, Theorem 4, applies. To calculate b_j, observe that

$$\begin{aligned} \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) &= \varepsilon_i\} \\ \cup\{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) &= -\varepsilon_i\} = \phi, \end{aligned}$$

and there are $\leq \Gamma(\mu_1^+, \cdots, \mu_m^+, j)$ components for each j-intersection $\neq \phi$, giving

$$\left(\binom{m}{j}\right) \Gamma(\mu_1^+, \cdots, \mu_m^+, j) 2^j$$

possibilities. [We have $\pm\varepsilon_1, \cdots, \pm\varepsilon_m$ to choose from, but never pick both $\pm$, giving $\left(\binom{m}{j}\right) \cdot 2^j$ choices].

If $j > \ell$, regularity forces $b_j = 0$. □

Corollary: If

$$\Gamma^*(\mu_1, \cdots, \mu_m) = \sup_{j \leq \ell} \Gamma(\mu_1^+, \cdots, \mu_m^+, j),$$

then the number of connected components of $\mathbf{R}^\ell \setminus \bigcup \{\tilde{y} : \mu_i(\bar{\alpha}_i, \tilde{y}) = 0\}$

$$\leq \left(\frac{2me}{\ell}\right)^\ell \cdot \Gamma^*(\mu_1, \cdots \mu_m).$$

Proof. Warren essentially showed

$$\Sigma_{r=0}^{\ell} 2^r \left(\binom{m}{r}\right) \leq \left(\frac{2me}{\ell}\right)^\ell$$

□.

2.5 Estimating $VC - \dim$ of $\mathcal{C}_\Phi$

We have Φ as before, with associated

$$\tau_1, \cdots, \tau_s.$$

Now by (*) and the Corollary to Theorem 6, we get

$$2^v \leq \left(\frac{2sve}{\ell}\right)^\ell \sup_{j \leq \ell} \Gamma(\tau_1^+, \cdots, \tau_v^+, j) \quad (**).$$

Note that we already started in * with $\tau_i \pm \varepsilon$, and translating by extra ε_i changes nothing.

The problem with (**) is that v occurs in the sup term. But this is quite illusory, since the τ_i+ for $i \leq sv$ are substitution instances of $\tau_i^+(\bar{v}, \tilde{y}),\ i \leq h$.
So

$$2^v \leq \left(\frac{2sve}{\ell}\right)^\ell \cdot \sup_{j \leq \ell} \Gamma(\tau_1^+, \cdots \tau_h^+, j)$$

and $B(\Phi)(= B) = \sup_{j \leq \ell} \Gamma^1(\tau_1^+, \cdots \tau_h^+, j)$ is independent of v.
If $v/\ell \leq 4se$ we get

$$2^v \leq B \cdot (4se)^{2\ell},$$

so

$$v \leq \log B + 2\ell \log(4se).$$

If $v/\ell > 4se$, we get

$$2^v < B \cdot (\frac{v}{\ell})^{2\ell}$$

so

$$2^{v/\ell} < B^{1/\ell}(v/\ell)^2.$$

Now $2^{v/2\ell} > (v/\ell)^2$ if $v/\ell > 16$, so either

$$v \leq 17\ell,$$

or

$$2^{v/2\ell} < B^{1/\ell},$$

i.e.

$$2^v < B^2 \qquad \text{, i.e} \qquad v < 2\log B.$$

So in all cases,

$$v < 2\log B + (17\log s)\ell$$

where log is to base 2.

Theorem 7. VC-dim$(\Phi) \leq [2\log B + (17\log s)\ell]$.

Proof. Done. □

2.6. An example involving exponentiation.

We work with +, -, ., 0,1, <, e^x, and appeal to Wilkie's work [W94], or [DMM94], for a proof of *o*-minimality. Let us suppose about Φ that its terms $\tau_i(\bar{v}, \tilde{y})(i \leq s)$ are polynomials of degree $\leq d$ in $\bar{v}, \tilde{y}$ and no more than q subterms $\exp(g(\bar{v}, \tilde{y}))$, where g is linear.

Khovanski [K91] has proved a basic result relating to this situation, namely:

Theorem 8. Let $Q_i\,(i \leq m)$ be elements of $\mathbf{R}[y_1, \cdots y_\ell, e^{\Lambda_1}, \cdots e^{\Lambda_q}]$, where the Λ_i are linear functions of $y_1, \cdots, y_\ell$. Suppose that the system

$$Q_1 = \cdots = Q_m = 0$$

is regular, so defining a manifold $\mathcal{M}$ of dimension $\ell - m$. Then if Q_i has degree d_i (in $y_1, \cdots y_\ell, e^{\Lambda_1}, \cdots, e^{\Lambda_q}$), $k = \ell - m$ and

$$S = \Sigma_{i=1}^m d_i + k + 1,$$

$\mathcal{M}$ has no more than

$$2^{q(q-1)/2} d_1, \cdots, d_m S^k [(k+1)S - k]^q$$

connected components.

Now note that in the proof of our main estimate we needed estimates only on number of connected components for regular configurations. So applying the above, in the notation of 2.5

$$B \leq 2^{q(q-1)/2} d^\ell [(\ell+1)(d+1)]^{\ell+q}.$$

Then $\log B$

$$\begin{aligned} \leq q(q-1)/2 \quad &+ \quad \ell \log d \\ &+ \quad (\ell+q)\log(\ell+1)(d+1). \end{aligned}$$

So in this case VC-dimension of $\mathcal{C}_\Phi$

$$\begin{aligned} \leq q(q-1)/2 \quad &+ \quad q\log(\ell+1)(d+1) \\ &+ \quad \ell(\log d + 17\log s). \end{aligned}$$

2.7 Application to sparse formulas. Since Khovanski's [K91] one has known how to use Finiteness Theorems about exponentiation to give uniform estimates in problems involving families of polynomials where there is an absolute bound to the number of nonzero coefficients occurring, but no bound on the degree of the polynomials. Using 2.6 we can readily get uniform bounds in for $VC - \dim \mathcal{C}_\Phi$ where Φ is a quantifier-free formula of the language of ordered fields. $\Phi(\bar{v}, \tilde{y})$ is, as usual, built from terms $\tau_i(\bar{v}, \tilde{y})\,(i \leq s)$, and in this case the τ_i are polynomial. Let us assume about the τ_i only that they involve at most q many $\tilde{y}$-monomials, as i varies.

The strategy is to break the $\tilde{y}$-space $\mathcal{R}^\ell$ into 3^ℓ pieces according to $y_i < 0, y_i = 0, y_i > 0$. Having made a choice for each i, one changes to variables y_i^1, with $y_i^1 = \log y_i$ if $y_i > 0, y_i^1 = \log(-y_i)$ if $y_i < 0$, and $y_i^1 = y_i$ if $y_i = 0$. Then $\Phi(\bar{v}, \tilde{y})$ transforms to $\Phi^1(\bar{v}, \tilde{y}^1)$, where Φ^1 involves terms polynomial in $\bar{v}$ and *linear* in no more than q exponentials of linear functions of the $\tilde{y}^1$.

To proceed, we have to inspect the main proof (2.4). The VC-dimension of $\mathcal{C}_\Phi$ is bounded by the number of connected components of a set in $\tilde{y}$-space. So clearly it is bounded by $3^\ell \cdot b$, where b is a uniform bounded covering all the subcases when we have made a fixed change of variable. But to the latter 2.6 applies, giving

$$b \leq e^\ell \cdot [2^{q(q-1)/2} \cdot [2(\ell+1)]^{\ell+q},$$

so in usual notation

$$\begin{aligned} \log B \leq \ell \log 3 \quad &+ \quad q(q-1)/2 \\ &+ \quad (\ell+q)\log 2(\ell+1), \end{aligned}$$

whence

$$\begin{aligned} VC - dim \quad \mathcal{C} \quad _\Phi \\ \leq \quad & q(q-1)/2 + q\log 2(\ell+1) \\ + \quad & \ell(\log 3 + \log 2(\ell+1) + 17\log s). \end{aligned}$$

3 Application to sigmoidal neural networks

We define (cf. [MS93]) a *sigmoidal* network architecture A. The data involves:
a) A directed acyclic graph G, labelled by variables and polynomials as explained below:
b) an integer ℓ, the dimension of the space of *weights*, and the weight variables $y_1, \cdots, y_\ell$ (ℓ is the number of *programmable* parameters);
c) if there are k input nodes (i.e. nodes of in-degree 0) these are labelled by variables $v_1, \cdots, v_k$;
d) there is exactly one output node (i.e. a node of out-degree zero);
e) those nodes which are not input nodes are called computation nodes, and the m^{th} such N_m is labelled by a variable z_m, and a polynomial

$$P_m(v_{t_1}, \cdots, v_{t_\rho}, z_{u1}, \cdots, z_{u\gamma}, y_{\lambda 1}, \cdots y_{\lambda\delta})$$

where the $y's$ are a subset of the weight variables, the $v's$ correspond to the input nodes immediately below m (i.e. connected *to* m) and the $z's$ correspond to the computation nodes immediately below m.

One now fixes an activation function $\sigma : \mathbf{R} \to \mathbf{R}$, in our case the function

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

Then A computes a function $\beta_A : \mathbf{R}^{k+\ell} \to \mathbf{R}$:
a) If N is a computation node, as above, labelled by z_m

$$\begin{aligned} f_N(\bar{v}, \tilde{y}) &= P_m(v_{t_1}, v_{t_\rho}, \sigma(f_{N_1}(\bar{v}, \tilde{y}), .. \\ &\quad \sigma(f_{N_\gamma}(\bar{v}, \tilde{y})), y_{\lambda_1}, \cdots y_{\lambda_\delta}) \end{aligned}$$

where N_i corresponds to u_i, $1 \leq i \leq \gamma$.

Then β_A is f_{N_w}, where N_w is the output node.

Now, if we work in a language with $+, -, \cdot, 0, 1$ and a symbol σ for the activation function, then $f_A(\bar{v}, \tilde{y})$ is given by a term $\tau(\bar{v}, \tilde{y})$, by transcribing naively the above recursion. Let $\Phi(\bar{v}, \tilde{y})$ be

$$\tau(\bar{v}, \tilde{y}) > 0.$$

Then (by definition) the VC-dimension of A is the VC-dimension of $\mathcal{C}_\Phi$. By [L92] (which appeals to Wilkie's [W94]) this dimension is finite, since σ is definable in $+, -, \cdot, 0, 1, e^x$.

Given a sigmoidal network architecture A, we now apply our results to get a very good estimate for $VC - \dim(A)$. We have simply to bound $\Gamma^*(\tau, j)$ for $j \leq \ell$ in order to get B. That is, we need to know a bound on the number of connected components of an intersection of no more than j sets of the form

$$\{\tilde{y} : \tau(\bar{\alpha}_i, \tilde{y}) = \varepsilon_i\} \qquad 1 \leq i \leq j.$$

This estimate is given by working in a higher-dimensional space and using the Khovanski estimate used earlier.

We use the computation variables Z_m, and others $\hat{Z}_m$ in correspondence with those. Write Z_w for the output variable. Now consider

$$\begin{aligned} \sum_m [(Z_m &- P_m(v_{t_1,\cdots,t_p}, \hat{Z}_{N1}, \cdots \hat{Z}_{N\gamma}, y_{\lambda_1 \cdots y_{\lambda_\delta}}))^2 \\ &+ (1 - \hat{Z}_m(1+e^{-Z_m}))^2] \\ &= \mu(\bar{v}, \bar{\bar{z}}, \tilde{y}). \end{aligned}$$

Notice that

$$\begin{aligned} &\mu(\bar{v}, \bar{\bar{z}}, \tilde{y}) = 0 \\ &\quad \to \quad Z_w = \tau(\bar{v}, \tilde{y}) \quad , \quad \text{and} \\ Z_w &= \tau(\bar{v}, \tilde{y}) \Leftrightarrow (\exists \bar{\bar{z}}_.)\mu(\bar{v}, \bar{\bar{z}}_., \tilde{y}) = 0). \end{aligned}$$

Let $m =$ the number of computation nodes. Then for fixed $\bar{\alpha}$, the number of connected components in $\mathbf{R}^{\ell+2m}$ of $\mu(\bar{\alpha}, \bar{\bar{z}}, \tilde{y}) = 0$ is, by [K91],

$$\begin{aligned} \leq \quad & 2^{m(m-1)/2}(2d)^{\ell+2m} \\ & [(\ell+2m+1)\cdot(2d+1)]^{\ell+3m} \end{aligned}$$

and this clearly gives a bound for the number for

$$\tau(\bar{\alpha}, \tilde{y}) = \varepsilon.$$

But we need to handle $\leq j$, $\tau(\bar{\alpha}_i, \tilde{y}) = \varepsilon_i$ together. So we need now variables

$$v_{r,i}, Z_{N,i}, \hat{Z}_{N,i}, \qquad , i \leq j,$$

thereby having us work in $\mathbf{R}^{\ell+2mj}$ space, and obtaining an estimate

$$\begin{aligned} & 2^{(mj)(mj-1)/2} \cdot (2d)^{\ell+2mj} \\ & [(\ell+2mj+1)(2d+1)]^{\ell+3mj} \end{aligned}$$

and since B can be chosen no larger than the supremum of these $j \leq \ell$, we get

$$\begin{aligned} \log B &\leq (m\ell)(m\ell-1)/2 + (\ell+2mj)(\log 2d) \\ &+ (\ell+3m\ell)\log(\ell+2m\ell+1) \\ &+ (\ell+3m\ell)\log(2d+1), \end{aligned}$$

so $VC - \dim(A) \leq (m\ell)(m\ell-1)/2 + \ell(2m+1)\log 2d + \ell(3m+1)\log((2m+1)\ell+1) + \ell(3m+1)\log(2d+1)$. $\square$

Remark: The estimation above with a dominant term $(m\ell)^2$ does not depend essentially on the type of the activation function used, and can be straightforwardly generalized to the large class of (multivariate) Pfaffian ([K91]) activation functions. To see this we generalize the condition of Theorem 8 defined by

$$p(\tilde{y}, e^{\Lambda_1(\tilde{y})}, \ldots, e^{\Lambda_q(\tilde{y})}) = 0$$

for p a polynomial, and Λ_i's linear, by appealing to Khovanski ([K91], p. 91). We replace the $e^{\Lambda_i(\tilde{y})}$ by q many functions occurring in a Paffian chain of length $\leq q$. We can extend above estimation to Pfaffian activation functions, taking into account the bound D on the degree of the polynomials occurring in the Pfaffian chain (resulting only in additional $\log D$ factors).

3.1 Sparse Networks.

We maintain the notations of Section 3, but now we consider families of $\mathcal{A}$, based on the same graph, but where the P_N can vary, subject to the restriction that none of them have more than Δ many nonzero coefficients. Let us go directly to the point where we work in $\mathbf{R}^{\ell+2mj}$ with the $v_{r,i}$ etc. By squaring and summing we will have changed Δ to $(2mj)\Delta^2$. Taking account of the need to consider the various quadrants in $\mathbf{R}^{\ell+2mj}$ we now derive easily for B an estimate

$$\begin{aligned} &\leq 3^{\ell+2mj} \cdot 2^{(2m\ell)\Delta^2(2m\ell\delta^2-1)/2} \\ &\cdot \quad 2(2+\ell+2m\ell)^{\ell+2m\ell-1}, \end{aligned}$$

thereby getting a dominant term $m^2\ell^2 \cdot \Delta^4$ for the VC-dimension.

3.2 Haussler's Pseudo Dimension.

We refer to [MS93] for the definition of the pseudo-dimension of an architecture. Since the pseudo-dimension of an architecture A is bounded by the VC-Dimension of a new architecture A' (see [MS93]) got directly from A, we get polynomial bounds for the pseudo-dimension. This answers affirmatively the second part of problem 10 in [M93]. □

Acknowledgement: We thank Gregory Cherlin, Mark Jerrum and Eduardo Sontag for a number of stimulating remarks and discussions.

References

[AB92] M. Anthony, N. Biggs, Computational Learning Theory: An Introduction, Cambridge University Press, 1992.

[AS93] M. Anthony, J. Shawe-Taylor, A Result of Vapnik with Applications, Discrete Applied Math. **47** (1993), pp. 207–217.

[BT90] A. Borodin, P. Tiwari, On the Decidability of Sparse Univariate Polynomial Interpolation, Proc. 22nd ACM STOC (1990), pp. 535–545.

[D92] L. van den Dries, Tame Topology and 0-minimal Structures, preprint, University of Illinois, Urbana, 1992; to appear as a book.

[DMM94] L. van den Dries, A.Macintyre and D.Marker, The Elementary Theory of Restricted Analytic Fields with Exponentation, Annals of Mathematics **140** (1994), pp 183-205.

[GJ93] P.Goldberg and M.Jerrum, Bounding the Vapnik Chervonenkis Dimension of Concept Classes Parametrized by Real Numbers. Machine Learning, 1994 (to appear). A preliminary version appeared in Proc. 6th ACM Workshop on Computational Learning Theory, pp. 361–369, 1993.

[H92] D. Haussler, Decision Theoretic Generalizations of the PAC Model for Neural Net and other Learning Applications, Information an Computation **100**, (1992), pp. 78–150.

[HKP91] J. Hertz, A. Krogh and R. G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, 1991.

[H76] M. W. Hirsch, Differential Topology, Springer-Verlag, 1976.

[KW93] M.Karpinski and T.Werther, VC Dimension and Uniform Learnability of Sparse Polynomials and Rational Functions, SIAM J. Computing **22** (1993), pp 1276–1285.

[K91] A.G.Khovanski, Fewnomials, American Mathematical Society, Providence, R.I., 1991.

[KPS86] J.Knight, A.Pillay and C.Steinhorn, Definable Sets and Ordered Structures II, Trans. American Mathematical Society **295** (1986), pp.593-605.

[L92] M.C.Laskowsky, Vapnik-Chervonenkis Classes od Definable Sets, J.London Math. Society **45** (1992), pp 377–384.

[M93] W.Maass, On the Complexity of Learning on Feedforward neural Nets, in Proc. EATCS Advanced School on Computational Learning and Cryptography, Vietri sul Mare, 1993.

[MSS91] W. Maass, G. Schnitger and E. D. Sontag, On the Computational Power of Sigmoidal versus Boolean Threshold Circuits, Proc. 32nd IEEE FOCS (1991), pp. 767–776.

[MS93] A.J.Macintyre and E.D.Sontag, Finiteness results for Sigmoidal Neural Networks, Proc. 25th ACM STOC (1993), pp.325–334.

[M64] J.Milnor, On the Betti Numbers of Real Varieties, Proc. of the American Mathematical Society **15** (1964), pp 275–280.

[M65] J.Milnor, Topology from the Differentiable Viewpoint, Univ.Press, Virginia, 1965.

[TV94] G. Turan and F. Vatan, On the Computation of Boolean Functions by Analog Circuits of Bounded Fan-in, Proc. 35th IEEE FOCS (1994), pp. 553–564.

[W68] H.E.Warren, Lower Bounds for Approximation by Non-linear Manifolds, Trans. of the AMS **133** (1968), pp. 167–178.

[W94] A.J.Wilkie, Model Completeness Results of Restricted Pfaffian Functions and the Exponential Function, to appear in Journal of the AMS, 1994.