JOURNAL OF COMPUTER AND SYSTEM SCIENCES **54**, 169–176 (1997)
ARTICLE NO. SS971477

# Polynomial Bounds for VC Dimension of Sigmoidal and General Pfaffian Neural Networks

Marek Karpinski*

*Department of Computer Science, University of Bonn, 53117 Bonn, Germany*

and

Angus Macintyre†

*Mathematical Institute, University of Oxford, Oxford OX1 3LB, United Kingdom*

Received May 10, 1995; revised October 25, 1995

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, that the VC dimension of analog neural networks with the sigmoidal activation function $\sigma(y) = 1/1 + e^{-y}$ is bounded by a quadratic polynomial $O((lm)^2)$ in both the number $l$ of programmable parameters, and the number $m$ of nodes. The proof method of this paper generalizes to much wider class of Pfaffian activation functions and formulas and gives also for the first time polynomial bounds on their VC dimension. We present also some other applications of our method. © 1997 Academic Press

## INTRODUCTION

This paper studies the VC dimension of general functional basis networks, and the resulting Boolean combinations of certain formulas. We develop a new method for proving explicit upper bounds for a wide class of analog neural networks with general Pfaffian activation functions.

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; M93a; 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 [MS93] the finiteness of VC dimension of sigmoidal neural networks has been established for the first time using a deep result in model theory. It is perhaps worth nothing that slightly more general analytic increasing activation functions do not always have a finite VC-dimension [S92].

In Maass's 1993 lecture notes [M93a] (see also [GJ93; MS93]), Open Problem 10 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 this paper we give an affirmative answer, with a polynomial bound in the number of programmable parameters. We believe that the bound can be improved to the one subquadratic in the number of programmable parameters and the number of nodes using a variant of our method. The result is a special case of a much more general result about the VC dimension of the classes defined by certain formulas. In contrast to [KM94], this paper does not use o-minimality and therefore can be applied to more general situations like the Pfaffian functions for which o-minimality is not yet even established(!).

In the case of boolean functions computed by sigmoidal neural networks (cf. [MSS91; M93b]), our result entails, also for the first time, by a simple counting argument, the fact that not every boolean function can be computed by a single polynomial size sigmoidal or general Pfaffian neural network with an appropriate weight assignment.

We refer to [AB92; GJ93; MS93] for all notions required for the VC dimension of neural networks, and to [H76] for all notions of differential geometry.

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.

The paper is organized as follows. In Section 1, we introduce the necessary formalism for the describing formulas, as well as all preparatory algebraic and topological facts. Section 2 contains the Main Result, and Sections 3 and 4 the applications.

* 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. Research partially done while visiting Dept. of Computer Science at Princeton University. E-mail: marek@cs.uni-bonn.de.

† Research supported in part by a Senior Research Fellowship of the EPSRC. E-mail: ajm@maths.ox.ac.uk.



0022-0000/97 $25.00
Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.

## 1. THE SETTING

1.1. We shall consider a standard model of a *feed-forward network architecture* $A$ with the *activation* function $\sigma$ (cf., e.g., [M93a, MS93]) with $k$ *inputs*, $m$ *computational nodes*, and $l$ *weights* (the number of *programmable parameters*). We assume (for simplicity) that the output gate of $A$ has range $\{0, 1\}$. We associate with $A$ an exponential formula $\Phi(\bar{v}, \tilde{y}) > 0$ for $\bar{v} \in \mathbb{R}^k$, and $\tilde{y} \in \mathbb{R}^l$, $\Phi$ 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 $\mathscr{C}_\Phi = \{\Phi_{\tilde{\beta}} : \tilde{\beta} \in \mathbb{R}^l\}$ for $\Phi_{\tilde{\beta}} = \{\bar{x} \in \mathbb{R}^k : \Phi(\bar{x}, \tilde{\beta}) > 0\}$ the partition of $\mathbb{R}^k$ by $A$ according to the weight assignment $\tilde{\beta}$. (The general reader is referred to [MS93; GJ93] for definitions and basic properties of the Vapnik–Chervonenkis (VC) dimension. We say a set $S \subseteq \mathbb{R}^k$ is *shattered* by $\mathscr{C}_\Phi$ if $\{S \cap C : C \in \mathscr{C}_\Phi\} = P(S)$. The VC dimension of $\mathscr{C}_\Phi$ is the maximal size of any set $S$ that can be shattered by $\mathscr{C}_\Phi$, or $\infty$ 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.

1.2. We start now with some definitions and notations. Fix integers $k$, $l$ and $C^\infty$ (infinitely differentiable) functions $\tau_1, ..., \tau_s$ from $\mathbb{R}^{k+l}$ to $\mathbb{R}$. Write $\tau_i$ as $\tau_i(v_i, ..., v_k, y_1, ..., y_l)$ (or $\tau_i(\bar{v}, \tilde{y})$).

Form a first-order language $L$ with primitives $<$ (for order) and function symbols $\bar{\tau}_1, ..., \bar{\tau}_s$, of arity $k + l$, corresponding to $\tau_1, ..., \tau_s$. (We drop $-$ for readability.)

Let $\Phi(\bar{v}, \tilde{y})$ be a quantifier-free $L$-formula, so $\Phi$ 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 $\tau$ is an $L$-term. For this paper, we assume each $\tau$ to be one of $\tau_i$'s.

For $\tilde{\beta} \in \mathbb{R}^l$, one defines

$$\Phi_{\tilde{\beta}} = \{\bar{v} \in \mathbb{R}^k : \mathbb{R} \models \Phi(\bar{v}, \tilde{\beta})\} \subseteq \mathbb{R}^k,$$

and the family

$$\mathscr{C}_\Phi = \{\Phi_{\tilde{\beta}} : \tilde{\beta} \in \mathbb{R}^l\}.$$

In this paper we give *good* explicit bounds on the VC-dimension of $\mathscr{C}_\Phi$, under certain assumptions about the $\tau_i$.

1.3. *Assumptions on the* $\tau_i$. Let $\bar{\alpha}_1, ..., \bar{\alpha}_V$ be elements of $\mathbb{R}^k$. Form the ($\leqslant sV$ *many*) $C^\infty$ funtions $\tau_i(\bar{\alpha}_j, \tilde{y})$ from $\mathbb{R}^l$ to $\mathbb{R}$. Choose $\Theta_1, ..., \Theta_r$ ($r \leqslant l$) from among these, and let

$$F : \mathbb{R}^l \to \mathbb{R}^r$$

be defined by

$$F(\tilde{y}) = \langle \Theta_1(\tilde{y}), ..., \Theta_r(\tilde{y}) \rangle.$$

By Sard's theorem [M65], the set of nonregular values $\langle \varepsilon_1, ..., \varepsilon_r \rangle$ of $F$ in $\mathbb{R}^r$ has Lebesgue measure 0. Recall that $\langle \varepsilon_1, ..., \varepsilon_r \rangle$ is a regular value of $F$ if either

(a) $F^{-1}(\langle \varepsilon_1, ..., \varepsilon_r \rangle) = \varnothing$, or

(b) $F^{-1}(\langle \varepsilon_1, ..., \varepsilon_r \rangle)$ is an $(l-r)$-dimensional $C^\infty$-submanifold of $\mathbb{R}^l$.

This motivates the assumption we now impose on the $\tau_i$.

*Assumption.* There is a bound $B$, independent on the $\bar{\alpha}_j$, $r$, and $\varepsilon_1, ..., \varepsilon_r$ such that if $F^{-1}(\langle \varepsilon_1, ..., \varepsilon_r \rangle)$ is an $(l-r)$-dimensional $C^\infty$-submanifold of $\mathbb{R}^l$ then $F^{-1}(\langle \varepsilon_1, ..., \varepsilon_r \rangle)$ has $\leqslant B$ connected components.

Fix such a $B$ henceforth.

1.4. Examples. (a) The $\tau_i$ are polynomials of degree $\leqslant d$ in $\tilde{y}$. Then $B$ can be taken as $2 \cdot (2d)^l$ by a result of Milnor [M64].

(b) Khovanski [K91, p. 91, Corollary 3] proved a basic result about exponential polynomials, namely:

Theorem 1. *Suppose* $l \geqslant m$. *Let* $Q_i$ ($i \leqslant m$) *be elements of* $\mathbb{R}[y_1, ..., y_l, e^{\Lambda_1}, ..., e^{\Lambda_q}]$, *where the* $\Lambda_i$ *are linear functions of* $y_1, ..., y_l$. *Suppose the* $Q_i$ *have degree* $d_i$, *and let* $k = l - m$, *and* $S = \sum_{i=1}^m d_i + k + 1$. *Suppose* $(0, ..., 0)$ *is a regular value of* $(Q_1, ..., Q_m)$, *with inverse image a manifold of dimension* $k$. *Then that manifold has no more than* $2^{q(q-1)/2} \cdot d_1 \cdots d_m \cdot S^k [(k+1) S - k]^q$ *connected components.*

This gives, for 1.3, if $\tau_i(\bar{v}, \tilde{y})$ is polynomial of degree $d$ in $\bar{v}$, $\tilde{y}$, and $q$ fixed subterms (independent of $i$) $\exp(g(\bar{v}, \tilde{y}))$, $g$ linear, $B \leqslant 2^{ql(ql-1)/2} \cdot d^l \cdot S^l [lS]^{ql}$, where $S = dl + l + 1$. The $q$ in Theorem 1 becomes $ql$ now, because of the substitutions of $\leqslant l$ many $\bar{\alpha}_j$ for $\bar{v}$. So

$$\log B \leqslant (ql)(ql-1)/2 + l \log d + l \log S + ql \log(lS)$$
$$\leqslant (ql)(ql-1)/2 + l \log d + l(q+1) \log S + ql \log l.$$

(c) If the $\tau_i$ are definable in an o-minimal expansion of the real field [KPS86], the existence of a $B$ is guaranteed, but *good* bounds are not.

(d) Examples for which o-minimality is unknown but where our method applies involved Pfaffian functions (cf. [K91, p. 91, Example 3]). We recall that a sequence of real functions $F_1, ..., F_q$ is a Pfaffian chain if all partial derivatives of every $F_i$, $1 \leqslant i \leqslant q$, can be expressed as polynomials in the first $i$ functions in the chain and the coordinate functions. Suppose the $\tau_i(\bar{v}, \tilde{y})$ are polynomials of degree $\leqslant d$ in the $\bar{v}, \tilde{y}$ and in functions $F_1, ..., F_q$ which form a Pfaffian chain of length $q$, where the polynomials are of degree $\leqslant D$. Let $r \leqslant l$ and $\Theta_1, ..., \Theta_r$ as in 1.3, defining a manifold of dimension $l-r$. Then if $S = r(d-1) + lD + 1$, we have

$$B \leqslant 2^{lq(lq-1)/2} \cdot d^r \cdot S^{l-r} [(l-r+1)\, S - (l-r)]^{lq}$$

giving, independent of $r$ a (crude) bound

$$B \leqslant 2^{lq(lq-1)/2} \cdot d^l \cdot (l(d+D))^l \, (l^2(d+D))^{lq}.$$

The bound in Theorem 1 corresponds to $D = 1$.

As for the exponential example Khovanski's $q$ becomes in our case $lq$ after the $\bar{\alpha}_j$ get substituted.

## 2. THE MAIN RESULT

2.1. We shall prove:

THEOREM 2 (Assumption as above). *VC-Dimension* $(\mathscr{C}_\Phi) \leqslant 2 \log B + (16 + 2 \log s)\, l$.

(Note. In this paper log is logarithm to base 2.)

2.2. Let $\bar{a}_1, ..., \bar{a}_V$ be elements of $\mathbb{R}^k$ such that $\{\bar{a}_1, ..., \bar{a}_V\}$ is shattered by $\mathscr{C}_\Phi$. For each subset $E$ of $\{\bar{a}_1, ..., \bar{a}_V\}$, pick $\tilde{y}_E$ in $\mathbb{R}^l$ such that $E = \{\bar{a}_j : \mathbb{R} \models \Phi(\bar{a}_j, \tilde{y}_E)\}$. Choose $\varepsilon > 0$ such that if any $\tau_i(\bar{a}_j, \tilde{y}_E)$ $(1 \leqslant i \leqslant s, 1 \leqslant j \leqslant V, E \subseteq \{\bar{a}_1, ..., \bar{a}_V\})$ is $\neq 0$, then $|\tau_i(\bar{a}_j, \tilde{y}_E)| > \varepsilon$.

Note that for $\tilde{y} \in \mathbb{R}^l$, the set $\{\bar{a}_j : \mathbb{R} \models \Phi(\bar{a}_j, \tilde{y})\}$ depends only on the *signs* $(+, -, \text{or } 0)$ taken at $\tilde{y}$ by the functions $\tau_i(\bar{a}_j, \tilde{y})$ $(1 \leqslant i \leqslant s, 1 \leqslant j \leqslant V)$. (The sign of $\lambda$ is $+$ if $\lambda > 0$, $-$ if $\lambda < 0$, and 0 if $\lambda = 0$.)

Because of the $\tilde{y}_E$ one has $\geqslant 2^V$ such sign series as $\tilde{y}$ varies. The $\tilde{y}_E$ now show the following.

LEMMA 3. *If* $0 < \varepsilon_{ij} < \varepsilon$ $(1 \leqslant i \leqslant s, 1 \leqslant j \leqslant V)$ *the complement in* $\mathbb{R}^l$ *of the union of the sets* $\{\tilde{y} : \tau_i(\bar{a}_j, \tilde{y}) = \varepsilon_{ij}\} \cup \{\bar{y} : \tau_i(\bar{a}_j, \tilde{y}) = -\varepsilon_{ij}\}$ $(1 \leqslant i \leqslant s, 1 \leqslant j \leqslant V)$ *has at least* $2^V$ *connected components* ($V, \bar{a}_j, \varepsilon$ *are fixed as above*).

2.3. This can now be combined with Sard [S42, M65], and a combinatorial idea of Warren [W68], to give Theorem 2.

We use the following cases of Sard's theorem. We have a $C^\infty$ map $F: \mathbb{R}^m \to \mathbb{R}^n$. A point $p$ of $\mathbb{R}^n$ is called a *regular value* of $F$ if either $m \leqslant n$ and $F^{-1}(p)$ is a submanifold of $\mathbb{R}^m$ of dimension $m-n$, or empty, or $m > n$ and $F^{-1}(p)$ is empty. Then the basic result is that the set of $q$ in $\mathbb{R}^n$ which are not regular values of $F$ has Lebesgue measure 0.

(It is easily seen that the normal definition of regular value, in terms of $F^{-1}(p)$ containing no critical points, is equivalent to that given above.)

Now we apply Sard [S42]. Let $P = \{\langle i, j\rangle : 1 \leqslant i \leqslant s, 1 \leqslant j \leqslant V\}$. For $\langle i, j\rangle \in P$, let $\tau_{i,j}(\tilde{y}) = \tau_i(\bar{a}_j, \tilde{y})$. For $A \subseteq P$, and $f \in \{1, -1\}^A$, let $F_{A,f}(\tilde{y}) = \langle ..., f(\langle i, j\rangle) \cdot \tau_{i,j}, (\tilde{y}), ...\rangle_{\langle i,j\rangle \in A}$. So $F_{A,f}$ is a $C^\infty$ map from $\mathbb{R}^l$ to $\mathbb{R}^A$. For $\boldsymbol{\varepsilon}$ in $\mathbb{R}^P$, let $Z(A, f)(\boldsymbol{\varepsilon}) = \{\tilde{y} : \text{for all } \langle i, j\rangle \in A,\ \tau_{i,j}(\tilde{y}) = f(\langle i, j\rangle)\, \tau_{i,j}(\tilde{y})\}$. Finally, let $I = [-1, 1]$, which has measure 2.

LEMMA 4. *Let* $\Gamma$ *be the set of all* $\boldsymbol{\varepsilon}$ *in* $I^P$ *such that for all* $A \subseteq P$ *with* $\operatorname{card}(A) = j \leqslant l$, *and all* $f \in \{-1, 1\}^A$, $Z(A, f)(\boldsymbol{\varepsilon})$ *is either empty, or a manifold of dimension* $l - j$. *Then* $\Gamma$ *has measure* $2^{\operatorname{card}(P)}$.

*Proof.* Look at the $\boldsymbol{\varepsilon}$ for which the condition fails for some $A, f$. Let $\Pi_A$ be the projection of $I^P$ onto $I^A$. Then $\Pi_A(\boldsymbol{\varepsilon})$ is not a regular value of $F_{A,f}$, so it belongs to a subset of $\mathbb{R}_A$ of measure 0. So the $\boldsymbol{\varepsilon}$ in $I^P$ for which the condition fails for $A, f$ have measure 0. Since there are only finitely many $A, f$ the result follows. ∎

Now a slight refinement.

LEMMA 5. *Let* $\Gamma'$ *be the subset of* $\Gamma$ *consisting of all* $\boldsymbol{\varepsilon}$ *such that if* $\operatorname{card}(A) > l$ *and* $f \in \{1, -1\}^A$, *then* $Z(A, f)(\boldsymbol{\varepsilon})$ *is empty. Then* $\Gamma'$ *has measure* $2^{\operatorname{card}(P)}$.

*Proof.* Again, consider the $\boldsymbol{\varepsilon}$ for which condition fails for a fixed $A, f$. As before, this set has measure 0. Since there are only finitely many $A, f$, the result follows. ∎

We now take up the notations of Lemma 3. The $\boldsymbol{\varepsilon}_{ij}$ in $I^P$ with $0 < \varepsilon_{ij} < \varepsilon$ form a set of measure $\varepsilon^{\operatorname{card}(P)}$. (Of course, $\operatorname{card}(P) = sV$). Combining this with Lemma 5, we get that $\Gamma'$ intersected with the above has measure $\varepsilon^{\operatorname{card}(P)}$, and so, in particular, it is nonempty.

Note finally, before we approach Theorem 2 via a theorem of Warren, that for $\boldsymbol{\varepsilon}$ in $\Gamma'$, if $A_1 \subseteq A_2$ and $f_1 \subseteq f_2$ then $Z(A_2, f_2)(\boldsymbol{\varepsilon})$ is a *submanifold* of $Z(A_1, f_1)(\boldsymbol{\varepsilon})$.

Warren [W68] proved:

THEOREM 6. *Let* $\mathscr{M}$ *be a connected topological* $n$*-manifold, and let* $\mathscr{M}_1, ..., \mathscr{M}_n$ *be connected* $(n-1)$*-manifolds which are submanifolds of* $\mathscr{M}$ *so that*

(1) *The* $\mathscr{M}_i$ *are closed in* $\mathscr{M}$;

(2) *The intersection of any given* $j$ *of the* $\mathscr{M}_i$, $1 \leqslant j \leqslant n$, *is either empty, or is an* $(n-j)$*-submanifold of the intersection of any* $(j-1)$ *of the* $\mathscr{M}_i$;

(3) *Any intersection of more than n of the* $\mathscr{M}_i$ *is nonempty.*

*Let* $b_j$ $(0 \leqslant j \leqslant n)$ *be the number of connected components among all intersections of any* $j_n$ *of the* $\mathscr{M}_i$. *Then* $\mathscr{M} - \bigcup_{i=1}^{n} \mathscr{M}_i$ *has no more than* $\sum_{j=0}^{n} b_j$ *connected components.*

*Proof.* See [W68, Theorem 1]. We want to apply this by fixing $\boldsymbol{\varepsilon}$ in $\Gamma'$, taking $\mathscr{M} = \mathbb{R}^l$ and the $\mathscr{M}_i$ as the zerosets of the $\tau_{i,j}(\tilde{y}) \pm \varepsilon_{ij}$. All that is missing is that we did not guarantee that these zerosets are connected. But if we rather take the $\mathscr{M}_i$ as the connected components of the zerosets, the hypotheses of Theorem 6 are satisfied. Indeed, Warren's result clearly remains true if the condition on connectedness of the $\mathscr{M}_i$ is dropped.

Going back to Lemma 3, we see to bound $2^V$ by the number of connected components of the complement in $\mathbb{R}^l$ of the union of the sets $\{\tilde{y}: \tau_{ij}(\tilde{y}) = \varepsilon_{ij}\} \cup \{\tilde{y}: \tau_{ij}(\tilde{y}) = -\varepsilon_{ij}\}$, where each $\varepsilon_{ij}$ is between 0 and $\varepsilon$, and $\boldsymbol{\varepsilon}$ is in $\Gamma'$. To apply Warren, we have to bound the $b_j$ for $0 \leqslant j \leqslant l$. Of course, $b_0 = 1$.

Now $n = 2sV$. Let $1 \leqslant j \leqslant l$. There are $n$ many zero sets, but of course any intersection $\{\tilde{y}: \tau_{ij}(\tilde{y}) = \varepsilon_{ij}\} \cap \{\tilde{y}: \tau_{ij}(\tilde{y}) = -\varepsilon_{ij}\} = \varnothing$.

Any intersection of no more than $j$ of the zerosets has $\leqslant B$ connected components (original assumption). So by these two remarks

$$b_j \leqslant 2^j \cdot \binom{sV}{j} \cdot B,$$

giving

$$\sum_{j=0}^{l} b_j \leqslant B \cdot \sum_{j=0}^{l} 2^j \cdot \binom{sV}{j} \leqslant B \cdot \left(\frac{2sVe}{l}\right)^l$$

by [W68].

So now we have

$$2^V \leqslant B \cdot \left(\frac{2sVe}{l}\right)^l.$$

*Conclusion of Proof of Theorem* 2. *Case* 1. $V \leqslant 4sel$. Then

$$2^V \leqslant B(8s^2e^2)^l \leqslant B(4se)^{2l},$$

so

$$V \leqslant \log B + 2l \log(4se) \leqslant \log B + 10l + 2l \log s.$$

*Case* 2. $V > 4sel$. Then

$$2^V \leqslant B \binom{V}{l}^{2l},$$

so

$$2^{V/l} \leqslant B^{1/l} \binom{V}{l}^2.$$

Now $2^{V/2l} > (V/l)^2$ if $V > 16l$, so either $2^{V/2l} < B^{1/l}$, or $V \leqslant 16l$. So either $2^V < B^2$, or $V \leqslant 16l$. So either $V < 2 \log B$, or $V \leqslant 16l$. ∎

## 3. APPLICATIONS

3.1. If we now work with polynomials and Milnor's bound for $B$, we get the results from [GJ93].

3.2. *An example involving exponentiation.* Fix $q$ and linear functions $\Lambda_1, ..., \Lambda_q$ of $\bar{v}$, $\tilde{y}$. Let $\tau_i(\bar{v}, \tilde{y})$, $1 \leqslant i \leqslant s$, be polynomials, of total degree $d_i$, in $\bar{v}$, $\tilde{y}$ and the $e^{\Lambda_i}$'s.

We showed after Theorem 1(1.4) that

$$\log B \leqslant (ql)(ql-1)/2 + l \log d + l(q+1) \log S + ql \log l,$$

where

$$S = dl + l + 1 \leqslant (d+1)(l+1).$$

So

$$\begin{aligned} VC-\mathrm{Dim}(\mathscr{C}_\Phi) &\leqslant (ql)(ql-1) + 2l \log d + 2l(q+1) \log S \\ &\quad + 2ql \log l + (16 + 2 \log s)\, l \\ &\leqslant (ql)(ql-1) + 2l \log d + 2l(q+1) \log(d+1) \\ &\quad + 2l(q+1) \log(l+1) + 2ql \log l + (16 + 2 \log s)\, l. \end{aligned}$$

So

$$\begin{aligned} VC-\mathrm{Dim}(\mathscr{C}_\Phi) &\leqslant (ql)(ql-1) + 4l(q+1) \log(l+1) \\ &\quad + 2l(q+2) \log(d+1) + (16 + 2 \log s)\, l. \end{aligned}$$

3.3. *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 none on the degrees involved. So this is all we now assume about the $\tau_i(\bar{v}, \tilde{y})$.

The strategy is to break the $\tilde{y}$-space $\mathbb{R}^l$ into $3^l$ pieces according to $y_j<0$, $y_j=0$, $y_j>0$.

Having chosen for each $j$ one such sign, one changes to variables $y'_j$ with $y'_j=\log(-y_j)$ if $y_j<0$, $y'_j=y_j$ if $y_j=0$, and $y'_j=\log(y_j)$ if $y_j>0$. Then $\tau_i(\bar{v}, \tilde{y})$ transforms to a function *linear* in no more than $q_i$ exponentials of linear functions of the $\tilde{y}'$, where $q_i$ is the number of nonzero coefficients of $\tau_i$. In particular any $\tau_i(\bar{a}_j, \tilde{y})$ will satisfy the hypotheses of Khovanski's Theorem 1, with $d_i=1$.

So we can apply 3.2 $3^l$ times. After taking $\log s$ we get for $VC-\mathrm{Dim}(\mathscr{C}_\Phi)$ the bound

$$(ql)(ql-1)+2l(q+1)+2l(q+1)\log(l+1)$$
$$+2ql\log l+(16+2\log s)\,l+l\log 3.$$

## 4. APPLICATION TO SIGMOIDAL NEURAL NETWORKS

4.1. Let us recall again [MS92] the definition of a sigmoidal network architecture $A$. The data involves:

(a) A directed acyclic graph $G$, labeled by variables and polynomials as explained below;

(b) an integer $l$, the dimension of the space of *weights* (the number of *programmable* parameters), and the weight variables $y_1, ..., y_l$;

(c) if there are $k$ input nodes (i.e., nodes of in-degree 0) these are labeled by variables $v_1, ..., v_k$;

(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 labeled by a variable $z_m$, and a polynomial

$$P_{N_m}(v_{t_1}, ..., v_{t_\rho}, z_{u_1}, ..., z_{u_\gamma}, y_{\lambda_1}, ..., y_{\lambda_\delta}),$$

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

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

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

Then $A$ computes a function $\beta_A\colon \mathbb{R}^{k+l}\to\mathbb{R}$, as described recursively below:

(a) If $N$ is an input node, with associated variable $v_i$, $f_N(\bar{v}, \tilde{y})=v_i$,

(b) If $N$ is a computation node with variable $z_m$,

$$f_N(\bar{v}, \tilde{y})=P_N(v_{t_1}, ..., v_{t_\rho}, \sigma(f_{N_1}(\bar{v}, \tilde{y})), ..., \sigma(f_{N_\gamma}(\bar{v}, \tilde{y})), y_{\lambda_1}, ..., y_{\lambda_\delta}) \qquad (\#)$$

where $N_i$ corresponds to $z_{u_i}$, $1\leqslant i\leqslant\gamma$.

Then $\beta_A$ is $f_{N_\omega}$, where $N_\omega$ is the output node.

Now, if we work in a language with $+$, $-$, $\cdot$, 0, 1, $<$, and a symbol $\sigma$ for the activation function, then $f_{N_w}(\bar{v}, \tilde{y})$ 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 $\mathscr{C}_\Phi$. By [L92] (which appeals to [W94]) this dimension is finite, since $\sigma$ is definable in $+$, $-$, $\cdot$, 0, 1, $<$, $e^x$.

We now apply our method to get a good polynomial bound for $VC-\dim(A)$. So we need to know a bound on the number of connected components of a manifold of dimension $l-j$ defined by the conditions

$$\tau(\bar{\alpha}_i, \tilde{y})=\varepsilon_i, \quad 1\leqslant i\leqslant j\ (\leqslant l).$$

We are aware of several approaches to this computation and may in the future look more closely at the relative merits of various methods. For now we appeal directly to the Khovanski estimates previously used, but now they are applied in a high-dimensional space.

For each $i$ with $1\leqslant i\leqslant j$, and each computation node $N$ we add variables $Z_{N,i}$ and $\hat{Z}_{N,i}$. Among these are the output variables $Z_{w,i}$ for each $i$. Finally, we add input variables

$$v_{c,i} \qquad \text{for} \quad c\leqslant k, \quad i\leqslant j.$$

Now consider the system of equations

$$Z_{N,i}=P_N(v_{t_1,i}, ..., v_{t_\rho,i}, \hat{Z}_{N_1,i}, ..., \hat{Z}_{N_\gamma,i}, y_{\lambda_1}, ..., y_{\lambda_\delta})$$
$$1=\hat{Z}_{N,i}(1+e^{-Z_{N,i}})$$

as $N$ ranges over computation nodes, and $1\leqslant i\leqslant j$. To see the meaning, refer to (#).

Write the system as

$$S(\bar{v}_1, ..., \bar{v}_j, z_{w,1}, ..., z_{w,j}, \tilde{y}, \tilde{\tilde{w}}),$$

where $\bar{v}_c=(v_{c,1}, ..., v_{c,j})$ and $\tilde{\tilde{w}}$ denotes all the remaining variables.

The essential points are:

(1) $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})\Rightarrow\tau(\bar{\alpha}_i, \tilde{y})=\varepsilon_i$, $1\leqslant i\leqslant j$;

(2) If $\tau(\bar{\alpha}_i, \tilde{y})=\varepsilon_i$ for all $1\leqslant i\leqslant j$, then there are unique $\tilde{\tilde{w}}$ such that $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})$;

(3) The set in $\mathbb{R}^l$ defined by the conditions $\tau(\bar{\alpha}_i, \tilde{y}) = \varepsilon_i$, $1 \leqslant i \leqslant j$, is homeomorphic to that in $(\tilde{y}, \tilde{\tilde{w}})$ space defined by $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})$, so either both or neither are manifolds.

So now we can use the Khovanski estimates on $S$, assuming $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})$ defines a manifold of dimension $l-j$. Note that there are $l+(2m+1)\cdot j$ variables among $(\tilde{y}, \tilde{\tilde{w}})$, if $m$ is the number of nonoutput computation nodes of $A$. $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})$ is defined by $2mj$ equations and, of course, $l+(2m+1)\, j-2mj = l-j$.

Let $d$ be a bound for the degree of all $P_N$. Then, by Khovanski, $S(\bar{\alpha}_1, ..., \bar{\alpha}_j, \varepsilon_1, ..., \varepsilon_j, \tilde{y}, \tilde{\tilde{w}})$ defines a set with no more than $2^{((m+1)j)((m+1)j-1)/2}\, d^{2mj}(2mjd+(l-j)+1)^{l-j}$ $((l-j+1)\, 2mjd-(l-j))^{(m+1)j}$ connected components.

So this gives us a bound $B$ for the $\tau$-problem, namely,

$$B \leqslant 2^{nl(nl-1)/2} \cdot d^{2nl} \cdot (l\cdot(2nd+1))^l \cdot (2nl^2 d)^{nl},$$

where $n = m+1 =$ number of computation nodes of $A$. So $\log B \leqslant nl(nl-1)/2 + 2nl \log d + l \log l + l \log(2nd+1) + 2nl \log l + nl \log(2nd) = \beta(A)$, say.

Now, applying Theorem 2, we get:

THEOREM 7. *The VC-dimension of $A$ is bounded above by*

$$2\beta(A) + 16l.$$

The term $(nl)(nl-1)/2$ is obviously the dominant term, if $d$ is small. Since in the general case $l$ could majorize $n$, one can argue that our bound is of degree 4 as a function of $l$ only.

4.2. *Generalizations*. The estimation above with a dominant term $(ml)^2$ does not depend essentially on the type of the activation function used. An alternative approach to the above result works directly with the function $f_{N_w}(\bar{v}, \tilde{y})$, and uses the fact that $f_{N_w}$ is a *Pfaffian* function. For the fundamental work on Pfaffian functions one should consult [K91].

$\sigma(x)$ is Pfaffian, since $\sigma'(x) = \sigma(x) - (\sigma(x))^2$. Clearly $f_N(\bar{v}, \tilde{y})$ is Pfaffian, for $N$ an input node, for $f_N(\bar{v}, \tilde{y}) = v_i$, where $v_i$ is the input variable corresponding to $N$. Using (#) we have, for a computation node $N$,

$$\begin{aligned}\frac{\partial}{\partial y_j} f_N(\bar{v}, \tilde{y}) = {} & \frac{\partial P_N}{\partial Z_{u_1}} \cdot \frac{\partial f_{N_1}}{\partial y_j} \cdot (\sigma(f_{N_1}) - \sigma(f_{N_1})^2) \\ & + \frac{\partial P_N}{\partial Z_{u_2}} \cdot \frac{\partial f_{N_2}}{\partial y_j} \cdot (\sigma(f_{N_2}) - \sigma(f_{N_2})^2) \\ & + \cdots + \frac{\partial P_N}{\partial Z_{u_r}} \cdot \frac{\partial f_{N_r}}{\partial y_j} \cdot (\sigma(f_{N_r}) - \sigma(f_{N_r})^2) \\ & + \sum \frac{\partial P_N}{\partial y_{\lambda_r}} \cdot \frac{\partial y_{\lambda_r}}{\partial y_j}\end{aligned}$$

and

$$\frac{\partial}{\partial y_j} \sigma(f_N(\bar{v}, \tilde{y})) = \frac{\partial}{\partial y_j} f_N(\bar{v}, \tilde{y}) \cdot (\sigma(f_N) - \sigma(f_N)^2).$$

From this one sees that if $\alpha_1, ..., \alpha_r$ $(r \leqslant l)$ are arbitrary values of $\bar{v}$ then the collection of all $f_N(\bar{\alpha}_i, \tilde{y})$ and $\sigma(f_N(\bar{\alpha}_i, \tilde{y}))$ for $r \leqslant l$ and $N$ an input or computation node, from a Pfaffian chain of length $2ml$, in which all polynomials have degree $\leqslant d+2$.

Finally, let $\Theta_i(\tilde{y})$ be $f_{N_w}(\bar{\alpha}_i, \tilde{y})$, a polynomial of degree $\leqslant d$ in the variables and the elements of the chain. Our task was to bound the number of connected components of

$$\{\tilde{y}\colon \Theta(\tilde{y}) = \varepsilon_1, ..., \Phi_r(\tilde{y}) = \varepsilon_r\}$$

under the assumption this is an $(l-r)$-submanifold of $\mathbb{R}^l$. We can apply [K91, p. 91, Example 3], described in 1.3.

So we get, in the present case,

$$B \leqslant 2^{2ml(2ml-1)/2} \cdot (d+2)^l \cdot S^{l-1} (lS)^{2ml},$$

where

$$S \leqslant (d+3)(l+1).$$

This is slightly inferior to the bound given in Theorem 7. However, the method used here clearly generalizes to give a huge variety of examples in which, as in 3.2 or Theorem 7, we get a dominant term quadratic in $ql$. (In the above $q = 2ml$).

In particular, the analogue of Theorem 3.2 is: Let $\tau_i(\bar{v}, \tilde{y})$, $1 \leqslant i \leqslant s$, be polynomials of degree $\leqslant d$ in the $\bar{v}$, $\tilde{y}$ and functions $f_1, ..., f_q$ in a Pfaffian chain of length $q$ and degree $\leqslant D$. Then

$$\begin{aligned} & VC-\mathrm{Dim}(\mathscr{C}_\Phi) \\ & \quad \leqslant 2(ql)(ql-1) + 2l \log d + 2l \log(ld + lD + 1) \\ & \qquad \times 2ql \log l + 2ql \log(ld + lD + 1) + l(16 + 2 \log S).\end{aligned}$$

As for Theorem 7, it generalizes to architectures with Pfaffian activation functions. The only difference is that a $q$ and a $D$ appear. Suppose that the activation functions of $A$ are all members of a Pfaffian chain of length $q$ and degree $D$. Then the argument outlined earlier for the sigmoid case gives

$$\begin{aligned} B & \leqslant 2^{lmq(lmq-1)/2} \cdot d^l \cdot l^l (d+D)^l \cdot (l^2(d+D))^l\, mq \\ & \leqslant 2^{lmq(lmq--1)/2} \cdot d^l \cdot l^{l+2lmq} \cdot (d+D)^{l+2lmq}\end{aligned}$$

so

$$\log B \leqslant lmq(lmq-1)/2 + l \log d + (l+2lmq) \log l + (l+2lmq) \log(d+D),$$

given the VC bound $\leqslant lmq(lmq-1) + (2 \log d + 16 + 2 \log s)\, l + 2(l+2lmq) \log l + 2(l+2lmq) \log(d+D)$. Thus there is a quadratic effect from $q$, but only a logarithmic one from $D$.

4.3. *Arctangent.* A special case is worth recording. Take the arctangent as the activation function of a network architecture. The Pfaffian chain is $1/(1+x^2)$, $\arctan x$, so $q=2$, and one readily verifies $D=2$. So one has for arctangent activation the dominant term $4lm$, rather than $lm$ for the sigmoid.

4.4. *Sparse networks.* We maintain the notations of 4.1, but now we consider families of $A$'s, based on some graph and $\sigma$, but where the $P_N$ can vary, subject to the restriction that none of them have more than $\Delta$ many nonzero coefficients. Combining the ideas of 3.3 and 4.1, we easily get for $\log B$ a bound with dominant term quadratic in $\ln \Delta$, and this is, of course, dominant in the VC-dimension bound for the $A$'s in the family.

4.5. *Haussler's pseudodimension.* 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 [M93a].

4.6. *Boolean functions.* We are interested now in computation of boolean functions $f\colon \{0,1\}^k \to \{0,1\}$ by neural networks (cf. [MSS91, M93b]). It is known that applying some single nonboolean activation functions enhances, sometimes dramatically, the computational power of a neural network (cf. [MSS91]) even if restricted to the boolean functions. However, it has been open for sometime now how much this increase in computational power of a neural network could be. The fundamental inability to answer this problem was caused by the lack of a method bounding the amount of information that can be encoded in the weights of a neural network. Particularly, no known methods were sufficient even to show that there always exists a boolean function $f\colon \{0,1\}^k \to \{0,1\}$ which *cannot* be computed by single constant depth, polynomial size (number of nodes and programmable parameters) neural network with sigmoidal activation function with an appropriate weight assignment. Main results of this paper entail a solution to this problem. In fact the polynomial bounds on the VC dimension entail that no subexponential size $2^{o(k)}$ sigmoidal or general Pfaffian neural network can compute all boolean function $f\colon \{0,1\}^k \to \{0,1\}$ under appropriate weight assignments. Let $A$ be a sigmoidal or general Pfaffian neural network with $m$ nodes and $l$ programmable parameters. Denote by $\mathbb{B}_A$ the set of *all boolean* functions computed by $A$ under an appropriate weight assignment, and by $d$ the VC dimension of $A$.

Observe that also the VC dimension of $A$ restricted to the boolean functions is bounded by $d$. We have $\ln(|\mathbb{B}_A|) \leqslant O(kd)$ (cf., e.g., [AB92]). Our $O((lm)^2)$ upper bounds on the VC dimension $d$ of $A$ entail now the following formula for the number $|\mathbb{B}_A|$ of different boolean functions computed by $A$: $|\mathbb{B}_A| \leqslant 2^{O(kl^2m^2)}$.

4.7. *Multivariate activation.* There is also more remarkable further generalization. There is an obvious way to consider network architectures with multivariate activation functions. If these are Pfaffian, we still get a quadratic dominant term. We will elaborate this in a future publication.

## 5. OPTIMALITY OF KHOVANSKI'S $2^{q(q-1)/2}$ BOUND?

We strongly suspect that this bound can be lowered to the order $q^q$ ($\sim 2^{q \log q}$). Obviously this would improve our upper bounds on the VC dimension. The best lower bound on the VC dimension of neural networks is $\Omega(l \log l)$ (cf. [M93a; M94]) for the threshold, and $\Omega(l^2)$ (cf. [KS95]) for piecewise polynomial and sigmoidal activation functions. There is still a large gap between $\Omega(l^2)$ lower bound and our $O(l^4)$ upper bound for sigmoidal and Pfaffian activation functions. The current bound on $B$ in our paper comes because of Khovanski's technique of removing one variable at a time (cf. [K91, p. 13]). We are looking closely at a method for getting to a kind of *Bezout's estimate* in one step, removing all variables simultaneously.

## ACKNOWLEDGMENTS

We thank Gregory Cherlin, Mark Jerrum, and Eduardo Sontag for stimulating remarks and discussions. In particular, the current organization of the paper was suggested by Eduardo Sontag.

## REFERENCES

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

[AS93] M. Anthony and J. Shawa-Taylor, A result of Vapnik with applications, *Discrete Appl. Math.* **47** (1993), 207–217.

[BT90] A. Borodin and P. Tiwari, On the decidability of sparse univariate polynomial interpolation, *in* "Proceedings, 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 exponentiation, *Ann. of Math.* **140** (1994), 183–205.

[GJ93] P. Goldberg and M. Jerrum, Bounding the Vapnik Chervonenkis dimension of concept classes parameterized by real numbers, *Mach. Learning* (to appear); a preliminary version *in* "Proc. 6th ACM Workshop on Computational Learning Theory, 1993," pp. 361–369.

[H12] G. H. Hardy, Properties of logarithmic-exponential functions, *Proc. London Math. Soc.* **10** (1912), 54–90.

[H92] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. Comput.* **100** (1992), 78–150.

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

[H76] M. W. Hirsch, "Differential Topology," Springer-Verlag, New York/Berlin, 1976.

[KM94] M. Karpinski and A. Macintyre, Polynomial bounds for VC dimension of sigmoidal neural networks, *in* "Proc. 27th ACM STOC, 1995," pp. 200–208.

[KM95] M. Karpinski and A. Macintyre, Bounding VC dimension for neural networks: Progress and prospects (invited lecture), *in* "Proc. EuroCOLT'95," Lecture Notes in Artif. Intell., Vol. 904, pp. 337–341, Springer-Verlag, New York/Berlin, 1995.

[KW93] M. Karpinski and T. Werther, VC dimension and uniform learnability of sparse polynomials and rational functions, *SIAM J. Comput.* **22** (1993), 1276–1285.

[K91] A. G. Khovanski, "Fewnomials," Amer. Math. Soc., Providence, RI, 1991.

[KPS86] J. Knight, A. Pillay, and C. Steinhorn, Definable sets and ordered structures II, *Trans. Amer. Math. Soc.* **295** (1986), 593–605.

[KS95] P. Koiran and E. D. Sontag, Neural networks with Quadratic VC dimension, *in* "Advances in Neural Information Processing Systems (NIPS '95), 1995," to appear.

[L92] M. C. Laskowski, Vapnik–Chervonenkis classes of definable sets, *J. London Math. Soc.* **45** (1992), 377–384.

[M93a] W. Maass, Perspectives of current research about the complexity of learning on neural nets, *in* "Theoretical Advances in Neural Computation and Learning" (V. P. Roychowdhury, K. Y. Siu, A. Orlitsky, Eds.), pp. 295–336, Kluwer Academic, Amsterdam, 1994.

[M93b] W. Maass, Bounds for the computational power and learning complexity of analog neural nets, *in* "Proc. 25th ACM STOC, 1993," pp. 335–344.

[M94] W. Maass, Neural nets with superlinear VC-dimension, *in* "Proc. of the International Conference on Artificial Neural Networks 1994, ICANN '94, Berlin, 1994," pp. 581–584; *Neural Comput.* **6** (1994), 875–882.

[MSS91] W. Maass, G. Schnitger, and E. D. Sontag, On the computational power of sigmoidal versus boolean threshold circuits, *in* "Proc. 32nd IEEE FOCS, 1991," pp. 767–776.

[MS93] A. J. Macintyre and E. D. Sontag, Finiteness results for sigmoidal neural networks, *in* "Proc. 25th ACM STOC, 1993," pp. 325–334.

[M64] J. Milnor, On the Betti numbers of real varieties, *Proc. Amer. Math. Soc.* **15** (1964), 275–280.

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

[S42] A. Sard, The measure of the critical points of differentiable maps, *Bull. Amer. Math. Soc.* **48** (1942), 883–890.

[S-T94] J. Shawe-Taylor, Sample sizes for sigmoidal neural networks, preprint, University of London, 1994; *in* "Proc. ACM COLT, 1995."

[S92] E. D. Sontag, Feedforward nets for interpolation and classification, *J. Comput. System Sci.* **45** (1992), 20–48.

[TV94] G. Turan and F. Vatan, On the computation of boolean functions by analog circuits of bounded fan-in, *in* "Proc. 35th IEEE FOCS, 1994," pp. 553–564.

[W68] H. E. Warren, Lower bounds for approximation by non-linear manifolds, *Trans. Amer. Math. Soc.* **133** (1968), 167–178.

[W94] A. J. Wilkie, Model completeness results of restricted Pfaffian functions and the exponential function, *J. Amer. Math. Soc.*, to appear.