Polynomial Bounds for the VC-Dimension of Sigmoidal, Radial Basis Function, and Sigma-pi Networks

Akito SAKURAI
Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-03, JAPAN
E-mail: sakurai@harℓ.hitachi.co.jp

Abstract. W^2h^2 is an asymptotic upper bound for the VC-dimension of a large class of neural networks including sigmoidal, radial basis functions, and sigma-pi networks, where h is the number of hidden units and W is the number of adjustable parameters, which extends Karpinski and Macintyre's resent results.* The class is characterized by polynomial input functions and activation functions that are solutions of first order differential equations with rational function coefficients and that can be represented in an implicit function form of a composition of the natural logarithm and polynomials. $O(W \log h)$ is a lower bound for the VC-dimension of sigmoidal, radial basis function, and sigma-pi networks.

1. Introduction

When Vapnik and Chervonenkis developed in [13] the concept that later became called Vapnik-Chervonenkis dimension, or VC-dimension, they dared not dream that it would become as common as is now. After Blumer et al. showed in [2] that the number of samples needed to accomplish certain learning task and that the expected error for unseen data are bounded from above by using the VC-dimension of a hypothesis space, many researchers, including those in neural network fields, have recognized the importance of the VC-dimension.

The VC-dimension is an index to measure the expressive capability of hypothesis space or, in neural network field, that of functional space whose member is a specific neural network with fixed weights and connections. The VC-dimension is roughly the maximum number of input data, most favorably arranged for the networks in the input data space, such that any of the 0–1 valued functions defined on these data (considered as points in the input data space) can be realized by a network with appropriately chosen weights. For networks that output continuous or analog value, the value is thresholded to 1 or 0 before considering the VC-dimension. That is, the network is assumed to be used for classification tasks. Still, the VC-dimension is known to relate to interpolation capability of the original networks.

We had been unable to obtain VC-dimension values for practical types of networks, until fairly tight upper and lower bounds were obtained ([10], [11], and [12]) for linear threshold element networks in which all elements perform a threshold function on weighted sum of inputs. Roughly, the lower bound for the networks is $(1/2)W \log h$ and the upper bound is $W \log h$ where h is the number of hidden elements and W is the number of connecting weights (for one-hidden-layer case $W \approx nh$ where n is the input dimension of the network).

In many applications, though, sigmoidal functions, specifically a typical sigmoid function $1/(1+\exp(-x))$, are used instead of the threshold function. This is mainly because the differentiability of the functions is needed to perform backpropagation or other learning algorithms. Unfortunately, we were yet to learn explicit bounds on the VC-dimension for this type of networks, although many related results were getting available in the literature. W. Maass obtained a result for piecewise linear functions networks, for which the VC-dimension is $O(W^2 \log p)$, where p is the number of linear pieces of the function ([8]). Macintyre and Sontag proved that the VC-dimension of neural networks with the sigmoid functions (and other broad types of functions) is finite ([9]). Bartlett and Williamson proved that $8W \log(11WD)$ is an upper bound on the VC-dimension of the two-layer (*i.e.*, one hidden layer) sigmoid networks with inputs restricted in $\{-D, -(D-1), \ldots, D-1, D\}$, where D is an integer ([1]).

Recently Karpinski and Macintyre proved a polynomial bound for the VC-dimension of the sigmoidal networks ([5]). They combined ideas of Warren ([14]), Khovanskii ([7]), and Goldberg and Jerrum ([3]); and obtained an upper bound $O(W^2h^2)$.

---

* M.Karpinski and A.J.Macintyre informed us that they also extended their previous results and will present it at EuroCOLT'95 ([6]).

We independently but later found the same method (the following section 3 is basically the same as [5]) but applied Khovanskii's results on Pfaffian eqaution cases rather than exponentiation function cases and proved that the similar upper bound holds for a larger class of networks such as radial basis functions and sigma-pi networks. M.Karpinski and A.J.Macintyre informed us that they also extended their previous results to Pfaffian cases and will present it at EuroCOLT'95 ([6]).

More precisely, our asymptotic upperbound for the VC-dimension is W^2h^2. The upper bound applies to a large class of networks, namely those with polynomial input functions and the activation functions that solve first order differential equations with rational function coefficients, which can be represented in a implicit function form of composing of the natural logarithm ln and polynomials.

2. Notations

Let F be a class of functions from the n-dimensional Euclidean space $\mathcal{R}^n$ to $\{0,1\}$, a set of two elements, and let S be a set of N points in $\mathcal{R}^n$. A *dichotomy* of S is a set of assignments of 0 or 1 to each point in S. Here, a dichotomy is *realized* by $f \in F$ if the value assigned by the dichotomy to each point $x \in S$ is equal to $f(x)$. If any of the possible dichotomies of S can be realized by some $f \in F$, S is *shattered* by F. The largest cardinality of S shattered by F is the *VC-dimension* of F.

A neural network is defined as having the following constituents:

1. a directed acyclic graph G with W edges and $h+n+1$ nodes, where
   - 1.1 there is exactly one output node (*i.e.*, a node with out-degree zero) labeled as N_{h+1},
   - 1.2 there are n input nodes (*i.e.*, a node with in-degree zero) labeled as $x_1, x_2, \ldots, x_n$,
   - 1.3 the other h nodes are called hidden nodes labeled as $N_1, N_2, \ldots, N_h$, and
   - 1.4 these nodes are sometimes called units;
2. an input function and an activation function for each hidden and output node, where
   - 2.1 the input function, which is denoted by $N_i(w_{i_1}, \ldots, w_{i_k}; x_1, \ldots, x_n; y_1, \ldots, y_{h_i})$, of node N_i operates on the netowrks inputs $x_1, \ldots, x_n$, the output values $y_1, \ldots, y_{h_i}$ of nodes connected via in-edges, and adjustable parameters $w_{i_1}, \ldots, w_{i_k}$ (usually called weights) of the node N_i, where k is the number of parameters of the node and h_i is the number of the in-edges of N_i; and
   - 2.2 the activation function $\alpha_i : \mathcal{R} \to \mathcal{R}$ operates on the value of N_i and gives the output value y_i of the node N_i.
3. the function of the network, which is denoted by $f_G(w_1, \ldots, w_W; x_1, \ldots, x_n)$, or simply $f_G(\mathbf{w}; \mathbf{x})$, is the composition of all the functions N_i.

To simplify the following, we use $N_i(\mathbf{w}; \mathbf{x}; \mathbf{y})$ to stand for $N_i(w_{i_1}, \ldots, w_{i_k};\ x_1, \ldots, x_n;\ y_1, \ldots, y_{h_i})$.

In many applications the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$ is used as the activation function. A standard sigmoidal network uses the inner product of weights and $x_1, \ldots, x_n; y_1, \ldots, y_{h_i}$ plus a threshold term as the input function, and the sigmoid function as the activation function. In sigma-pi networks the input function is a polynomial of $x_1, \ldots, x_n; y_1, \ldots, y_{h_i}$ with the weights as coefficients and the sigmoid function as the activation function. A radial basis function (RBF) network uses elliptic quadric functions as the input function and $\exp(x)$ as the activation function.

We suppose in this paper that the networks are used for classification tasks; that is, the output of the network is considered to be 1 if the real output is positive and 0 otherwise. This is equivalent to applying a simple threshold function $\mathcal{H}$ on the output value of the output node, where $\mathcal{H}$ is defined as $\mathcal{H}(t) = 1$ if $t \geq 0$ and 0 otherwise. Hence we can define the VC-dimension of a function class $\{f_G(\mathbf{w}; \cdot) \mid \mathbf{w} \in \mathcal{R}^W\}$ as that of $\mathcal{H}(f_G)$.

From the above definitions, the VC-dimension of the network, that is, the VC-dimension of the function set $\{f_G(\mathbf{w}; \cdot) \mid \mathbf{w} \in \mathcal{R}^W\}$, is

$$\max\{N \mid 2^N \leq \max_{\mathbf{x}_1, \ldots, \mathbf{x}_N} |\{\langle \mathcal{H}(f_G(\mathbf{w}; \mathbf{x}_1)), \ldots, \mathcal{H}(f_G(\mathbf{w}; \mathbf{x}_N))\rangle \mid \mathbf{w} \in \mathcal{R}^W\}|.$$

Since f_G is continuous,

$$|\{\langle \mathcal{H}(f_G(\mathbf{w}; \mathbf{x}_1)), \ldots, \mathcal{H}(f_G(\mathbf{w}; \mathbf{x}_N))\rangle \mid \mathbf{w} \in \mathcal{R}^W\}|$$
$$\leq \text{ "the number of connected components in } \mathcal{R}^W - \bigcup_{i=1}^{N} \{\mathbf{w} \mid f_G(\mathbf{w}; \mathbf{x}_i) = 0\}", \quad (1)$$

except in rare cases such as when there exists $\mathbf{w}_0$ such that $f_G(\mathbf{w}_0; \mathbf{x}_{i_j}) = 0$ $(1 \leq j \leq l)$ and for any $\mathbf{w}$ in any neighborhood of $\mathbf{w}_0$ there exists $1 \leq j \leq l$ such that $f_G(\mathbf{w}; \mathbf{x}_{i_j}) > 0$. To simplify the following argument

we will assume (see below) that N_{h+1} is a linear unit so that the cases never occur. Another way to avoid the cases is to use the argument Goldberg and Jerrum used ([3]).

In the following we suppose that the input function is a polynomial and the activation function $\alpha(x)$ of hidden nodes is a path-connected solution of a first order differential equation with rational function coefficients and can be represented in an implicit funciton form of ln and polynomials. For example, the function α that satisfies $\ln g_1(\alpha)+g_2(\alpha) = \ln g_3(x)+g_4(x)$ and $d\alpha/dx = ((1/g_3)(dg_3/dx)+dg_4/dx)/((1/g_1)(dg_1/d\alpha)+dg_2/d\alpha)$ is legitimate for the activation function, where g_i's are rational functions. The solution of $d\alpha/dx = 1-\alpha^2$ or $d\alpha/dx = \alpha$ is legitimate. Clearly these functions are definable in *exp*-**RA**, which we will use to follow Warren's argument. However, the solution of $d\alpha/dx = 1+\alpha^2$ is not legitimate; if we restrict the solution function to be defined only on a closed interval then we could apply the same argument in this paper ([9]).

To simplify the following arguments, we assume that the activation function of the output node is the identity function and the input function for the output node is linear. We also assume that the hidden nodes are numbered 1 through h according to their proximity to the input node; the node i is never in the node j's downstream when $i < j$.

3. Partition Number

Following the idea of Goldberg and Jerrum [3], we reduce the problem of finding the VC-dimension to a problem of finding the partition number, that is, the number of path-connected areas whose boundaries are the nullsets of the output functions of the network for certain inputs:

$$f_G(\mathbf{w};\mathbf{x}_i) = 0 \qquad (1 \le i \le N) \tag{2}$$

Let us restate Theorem 1 by Warren ([14]).

Theorem 1. Let M be a connected topological n-manifold, and let $M_1,\ldots,M_b$ 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, $1 \le j \le n$, is either empty or is an $(n-j)$-manifold that has finitely many components and is locally flat in the intersection of any $j-1$ of the given M_i;
(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.
Under these hypotheses the set $M - \bigcup_{i=1}^m M_i$ has at most $\sum_{j=0}^n b_j$ connected components.

We need some other notations to continue.

Notation. $\mathcal{N}(f)$ is the set $\{x \mid f(x) = 0\}$.

Definition. An intersection $\bigcap_{i=1}^m \mathcal{N}(p_i)$ is called *regular* if at each point x of the intersection the gradients of $p_1,\ldots,p_m$ at x are linearly independent.

Definition. $\mathcal{N}(p_1),\ldots,\mathcal{N}(p_m)$ are said to form a *regular configuration* if all intersections $\bigcap_{i=1}^j \mathcal{N}(p_{k_i})$ for $j = 1,\ldots,m$ are regular.

We are heavily dependent upon the next theorem (see [9] and its references).

Theorem 2. Sets definable in terms of $\mathcal{R}, +, \cdot, <, \exp$ and Boolean operations admit finite decompositions into topological manifolds.

In this section, we set $f_i = f_G(\mathbf{w};\mathbf{x}_i)$. The following three lemmas establish the perturbation, which correspond to lemmas by Warren in [14] (the names of the corresponding lemmas are shown in parentheses).

Lemma. (Lemma 2.2) If $V = \bigcap_{i=1}^m \mathcal{N}(f_i)$ is regular or if $V = \mathcal{R}^W$ then for all but finitely many real numbers α, $V(\alpha) = \mathcal{N}(f_i - \alpha) \cap V$ is regular.

Lemma. (Lemma 2.3) Let $S = \bigcap_{i=1}^m \mathcal{N}(f_i)$. For any $f(\mathbf{w}) = f_G(\mathbf{w};\mathbf{x}_0)$, there is a $\beta > 0$ such that whenever $0 \le \epsilon \le \beta$, every topological component of the set $E = \mathcal{R}^W - S \cup \mathcal{N}(f)$ contains one of the sets $E_\epsilon = \mathcal{R}^W - S \cup \mathcal{N}(f-\epsilon) \cup \mathcal{N}(f+\epsilon)$.

Lemma. (Lemma 2.4) Given $f_1,\ldots,f_m$, one can choose $\epsilon_1,\ldots,\epsilon_m$ such that $\mathcal{N}(f_1-\epsilon_1),\mathcal{N}(f_1+\epsilon_1),\ldots,\mathcal{N}(f_m-\epsilon_m),\mathcal{N}(f_m+\epsilon_m)$ form a regular configuration and such that every connected component of $E = \mathcal{R}^W - \bigcup_{i=1}^m \mathcal{N}(f_i)$ contains one of $F = \mathcal{R}^W - \bigcup_{i=1}^m \mathcal{N}(f_i-\epsilon_i) \cup \mathcal{N}(f_i+\epsilon_i)$.

The proofs for the above lemmas are similar to those by Warren in [14] except for the first lemma in which Theorem 2 is used instead of referring to properties of the complex projective space.

Using the next lemma (the proof is the same as Warren [14]), we deduce Theorem 3.

Lemma. (Lemma 2.1) If $M_1, \ldots, M_b$ are the connected components of several $\mathcal{N}(f_1), \ldots, \mathcal{N}(f_m)$ which form a regular configuration, then the $(b+1)$-tuple $(\mathcal{R}^W; M_1, \ldots, M_b)$ satisfies the hypotheses of Theorem 1.

Theorem 3. The number of connected components, $N_{cc}(\cdot)$ of $\mathcal{R}^W - \bigcup_{i=1}^m \mathcal{N}(f_i)$ is bounded from above by

$$\sum_{j=1}^{W} \binom{m}{j} 2^j \max_{i_k\text{'s and } \alpha_i\text{'s} > 0} \{N_{cc}(\textstyle\bigcap_{k=1}^j \mathcal{N}(p_{i_k})) \mid p_{i_k} \text{ is either } f_{i_k} + \alpha_{i_k} \text{ or } f_{i_k} - \alpha_{i_k}\}.$$

This theorem is a direct consequence of the above lemmas and the fact that $\mathcal{N}(f_i + \alpha) \cap \mathcal{N}(f_i - \alpha) = \phi$.

4. Upper bounds

In this section we show an explicit upper bound for $N_{cc}(\bigcap_{i=1}^m \mathcal{N}(f_i))$ where $f_i(\mathbf{w}) = f_G(\mathbf{w}; \mathbf{x}_i)$. Note that the α's which appeared in the previous section are absorbed into the activation function of the output node of G in this section.

The theory of fewnomials is crucial in this section. The name "fewnomial" is from Kushnirenko who first posed the problem of estimating from above the number of isolated nondegenerate roots in the positive orthant in $\mathcal{R}^n$ of a polynomial system $P_1 = \cdots = P_n = 0$ via the number of monomials appearing in the system ([7]). Khovanskii has extensively studied the problem and obtained many interesting results. We use the following in this paper (new terminology is defined below):

Theorem 4. Let us consider an ordered system of Pfaff equations $\alpha_1 = \cdots = \alpha_q = 0$ in $\mathcal{R}^N$, in which all 1-forms have polynomial coefficients. For $1 \leq j \leq q$, let all coefficients of the form $\beta_j = \alpha_j \wedge \cdots \wedge \alpha_1$ be polynomials of degree $\leq k_j$. Then the manifold determined by the nonsingular system of polynomial equations $f_1 = \cdots = f_m = 0$ on each separating solution of the system of Pfaff equations is homotopy equivalent to a cell complex with at most $p_1 \cdots p_n \cdot \varphi_1 \cdots \varphi_q$ cells, where, for $1 \leq i \leq m$, p_i is the degree of the polynomial f_i, $n = N - q$, $p_{m+1} = \cdots = p_n = k_q + \sum_{i=1}^m (p_i - 1) + 1$, for $1 \leq j < q$, $\varphi_j = 2^{q-j} \sum_{i=1}^n (p_i - 1) + \sum_{l=1}^{q-j} 2^{(l-1)} k_{j+l} + k_j + 1$, and $\varphi_q = \sum_{i=1}^n (p_i - 1) + k_q + 1$. Assume also that all numbers k are at most equal to M. Then, the number is bounded from above by

$$2^{q(q-1)/2} p_1 \cdots p_m \cdot S^{n-m}[(n-m+1)S - (n-m)]^q,$$

where $S = \sum_{i=1}^m (p_i - 1) + M + 1$.

Definition. A submanifold Γ of co-dimension q in a manifold M is called a separating solution of an ordered system $\alpha_1 = \cdots = \alpha_q = 0$ of Pfaff equations (where α_i is a 1-form on M) if there is a chain of submanifolds $M = \Gamma^0 \supset \Gamma^1 \supset \cdots \supset \Gamma^q = \Gamma$ such that, for each $i = 1, \ldots, q$, the manifold Γ_i is a solution of the Pfaff equation $\alpha_i = 0$ on the manifold Γ^{i-1}.

Theorem 4 is obtained by combining Corollary 2, Example 1, and Example 2 in §3.14 of [7].

Let us consider our case. The input-output function of the network G is expressed as:

$$y_i - \alpha_i(N_i(\mathbf{w}; \mathbf{x}; \mathbf{y})) = 0 \quad (1 \leq i \leq h+1) \tag{3}$$

where N_i is the input function and α_i is the activation function of the node N_i. Note that the output node is N_{h+1} and the network output $f_G(\mathbf{w}; \mathbf{x}) = y_{h+1}$.

The standard conversion technique developed by Khovanskii ([7]) is to replace each appearance of transcendental functions, say $\alpha_i(N_i(\mathbf{w}; \mathbf{x}; \mathbf{y}))$, in the equations by a newly introduced variable, say u_i, and add a new Pfaffian form equation, $d(u_i - \alpha_i(N_i(\mathbf{w}; \mathbf{x}; \mathbf{y}))) = 0$.

What we want to estimate is the number of connected components of $\bigcap_{i=1}^m \mathcal{N}(f_G(\mathbf{w}; \mathbf{x}_i)) \subset \mathcal{R}^W$. Theorem 4 seems to apply only to $\bigcap_{i=1}^m \mathcal{N}(f_G(\mathbf{w}; \mathbf{x}_i; \mathbf{y})) \subset \mathcal{R}^{W+h+1}$, since the intermediate variables $\mathbf{y} \in \mathcal{R}^{h+1}$ appear free in (3). This may derive results other than what we want. However, it turns out to be illusory; we can treat y_i in the same way as the new variables to be introduced to replace appearances of transcendental functions. In this way we can in effect solve the problem to bound from above: $\bigcap_{i=1}^m \{\mathbf{w} \mid \exists \mathbf{y}\, f_G(\mathbf{w}; \mathbf{x}_i; \mathbf{y}) = 0\}$.

Our original system of equations is:

$$y_{j,i} - \alpha_i(N_i(\mathbf{w}; \mathbf{x}_j; \mathbf{y}_j)) = 0 \quad (1 \leq i \leq h+1,\ 1 \leq j \leq m) \tag{4}$$

Note that in (3), $\mathbf{y}$ is an existentially quantified variable and may depend on values of $\mathbf{x}_j$, so that we have to prepare $\mathbf{y}_j$ separately for each $\mathbf{x}_j$.

During the conversion, we use $u_{j,i}$ for $N_i(\mathbf{w};\mathbf{x}_j;\mathbf{y}_j)$. Then we discard $y_{j,i}-u_{j,i}=0$ by substituting $y_{j,i}$ for $u_{j,i}$. Our converted equations consist of m polynomial equations (in fact linear equations, since we have assumed that α_{h+1} is the identity function and N_{h+1} is linear),

$$y_{j,h+1}-N_{h+1}(\mathbf{w};\mathbf{x}_j;\mathbf{y}_j)=0 \quad (1\le j\le m), \tag{5}$$

and mh Pfaff equations,

$$dy_{j,i}-Q_i(f_{j,i},y_{j,i})\left(\sum_{k=1}^{W}\frac{\partial f_{j,i}}{\partial w_k}dw_k+\sum_{k=1}^{i-1}\frac{\partial f_{j,i}}{\partial y_{j,k}}dy_{j,k}\right)=0 \quad (1\le i\le h,\ 1\le j\le m), \tag{6}$$

where $f_{j,i}$ is an abbreviation of $N_i(\mathbf{w};\mathbf{x}_j;\mathbf{y}_j)$, α_i is a solution of a first-order differential equation $d\alpha_i/dx=Q_i(x,\alpha_i)$, where Q_i is a rational function of two variables, and can be represented in an implicit function form of a composition of ln and polynomials. We further eliminate the linear equations (5) by solving them for any variables and applying the solution to (6); note that the degrees of coefficient polynomials do not increase.

Note that variable y_j never appears in N_i $(i\le j)$, and that the requirements in Theorem 4 are satisfied. Hence we get the next theorem:

Theorem 5. $N_{cc}(\bigcap_{i=1}^{m}\mathcal{N}(f_G(\mathbf{w};\mathbf{x}_i))$ is bounded from above by

$$2^{mh(mh-1)/2}\cdot S^n[(n+1)S-n]^{mh},$$

where $S=m\sum_{i=1}^{h}(\deg(Q_i)+\deg(N_i)-1)+1$, where $\deg(Q_i)$ is the maximum of the degree of the numerator and denominator of Q_i. Note that for a sigmoidal neural network, $\deg(Q_i)=2$ and $\deg(N_i)=1$, and for a radial basis function network, $\deg(Q_i)=1$ and $\deg(N_i)=2$, so that $S=2mh+1$ in both cases.

Lemma 6. If $m\ge 7W+(3/2)\log C$, then $2^m>\sum_{j=1}^{W}2^j\binom{m}{j}\cdot C$.

(Proof) Using a formula ([2]): "for any $m\ge d\ge 1$, $\sum_{i=0}^{d}\binom{m}{i}\le 2(m^d/d!)\le(em/d)^d$," we get $\sum_{j=1}^{W}2^j\binom{m}{j}\cdot C<(2em/W)^W C$. Let $m=7W+(3/2)\log C+t$. Then $m-(W\log(2em/W)+\log C)>W[(\log(128/14e)+(\log C)/(2W)+t/W)-\log(1+(2\log C)/(7W)+t/(7W))]>0$ for $t\ge 0$.

Combining Theorem 5 and Lemma 6, we get the desired VC-dimension bound.

Theorem 7. When W is large enough, the VC-dimension of the sigmoidal , radial basis function, and sigma-pi networks is asymptotically bounded from above by W^2h^2. The same bound applies to any networks with polynomial input functions and activation functions which are path-connected solutions of a first-order differential equation with rational function coefficients and that can be represented in an implicit function form of a composition of the natural logarithm ln and polynomials.

5. Lower bounds

We will briefly describe lower bounds for the VC-dimension of sigmoidal, radial basis function, sigma-pi networks. Our result for the one-hidden-layer case is just a variation on the next theorem ([10]).

Theorem. For a certain set S of $N=(1/2)nh(\log h-o(\log h)-O((\log h)^2/n))$ points in general position in $\mathcal{R}^n$, any dichotomy of S can be realized by some one-hidden-layer neural network with linear threshold function units (the network has n input, h hidden and one output units), and the Lebesgue measure of a set of such point sets is not zero.

The main points of the proof of the theorem are roughly,

1. to confine any given $n-\log h$ points among $(1/2)(n-\log h)h$ points in general position in $\mathcal{R}^{n-\log h}$ (that is, to suitably adjust the $n-\log h$ parameters of a hidden unit to have value of 1 on these points and 0 on the remaining points); and
2. to separate $\log h$ points in $\mathcal{R}^{\log h}$ (*i.e.,* to output 1 on one set of points and 0 on the other) arbitrarily by adjusting $\log h$ parameters of a hidden unit.

For the linear threshold network we used a pair of hidden units for 1 above. Clearly the same could be done for the sigmoidal and and sigma-pi networks. For the radial basis function networks, we do not need a pair for 1 above; we could do it by just one unit. Hence we have the following theorem:

Theorem 8. For a certain set S of $N=(1/2)nh(\log h-o(\log h)-O((\log h)^2/n))$ points in general position in $\mathcal{R}^n$, any dichotomy of S can be realized by some one-hidden-layer neural network with sigmoidal or sigma-pi units (the network has n input, h hidden and one output units). This means that the VC-dimension of

the set of networks is bounded from below asymptotically by $(1/2)nh\log h$ or $(1/2)W\log h$. For radial basis networks, the above N and the VC-dimension are twice as large. In any case the Lebesgue measure of a set of such point sets is not zero (in fact it is infinity).

Similar arguments can be done to get the following: (see [11] and [12]).

Theorem 9. $O(W\log h)$ is a lower bound for the VC-dimension of the sigmoidal, radial basis function, and sigma-pi networks.

6. Conclusion

We showed that the VC-dimension of a class of neural networks, which includes sigmoidal, radial basis function, and sigma-pi networks, is bounded from above asymptotically by W^2h^2, where h is the number of hidden units and W is the number of adjustable parameters, where the upper bound results extend recent results by Karpinski and Macintyre. On the other hand, the proven lower bound for the above three types of networks is only $O(W\log h)$. The gap between the two bounds is far larger than those obtained for the linear threshold network case ([10],[11], [12]). Since, as noted by Khovanskii ([7]), the upper bound he gave is quite crude when the transcendental functions are just the exponentiation function, the upper bound for the VC-dimension we gave may also be crude. We expect that $O(W\log W)$ may be an upper bound for the VC-dimension for networks of bounded depth (the unbounded case is different; we will show it in our later paper) with sigmoid function units with linear input functions.

7. Acknowledgments

We would like to thank Prof. John Shawe-Taylor for his pointing out the work of Karpinski and Macintyre, Prof. Shun-ichi Amari for his valuable comments and Ms. Mariko Sakurai for her valuable help.

8. References

[1] Bartlett, P. L. & R. C. Williamson: The VC-dimension and pseudodimension of two-layer neural networks with discrete inputs, *preprint* (1993).

[2] Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth: Learnability and the Vapnik-Chervonenkis Dimension, *Journal of the ACM*, **36**(4), 929-965 (1989).

[3] Goldberg, P. and M. Jerrum: Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers, *Proc. Sixth Annual ACM Conference on Computational Learning Theory*, 361–369 (1993).

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

[5] Karpinski, M. and A. Macintyre: Quadratic bounds for VC dimension of sigmoidal neural networks, *manuscript* (1994).

[6] Karpinski, M. and A. Macintyre: *personal communication* (1995).

[7] Khovanskii, A. G. : *Fewnomials*, Translations of Mathematical Monographs 88, AMS (1991).

[8] Maass, W. G.: Bounds for the computational power and leaning complexity of analog neural nets, *Proc. 25th Annual Symposium of the Theory of Computing*, 335-344 (1993).

[9] Macintyre, A. and E. D. Sontag: Finiteness results for sigmoidal "neural" networks, *Proc. 25th Annual Symposium of the Theory of Computing*, 325–334 (1993).

[10] Sakurai, A.: Tighter Bounds of the VC-Dimension of Three-layer Networks, *Proc. WCNN'93*, III, 540–543 (1993).

[11] Sakurai, A.: On the VC-dimension of depth four threshold circuits and the complexity of Boolean-valued functions, *Proc. ALT93 (LNAI 744)*, 251–264 (1993).

[12] Sakurai, A.: On the VC-dimension of neural networks with a large number of hidden layers, *Proceedings of NOLTA'93*, IEICE, 239–242 (1993).

[13] Vapnik, V., and A. Y. Chervonenkis : On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and Its Applications*, **16**, 264-280 (1971).

[14] Warren, H. E.: Lower bounds for approximation by nonlinear manifolds, *Trans. AMS*, **133**, 167–178 (1968).