Lower Bounds on the VC-Dimension of Smoothly Parametrized Function Classes

Wee Sun Lee
Dept. of Systems Engineering,
RSISE, Aust. Nat. University,
Canberra, ACT 0200, Australia.
WeeSun.Lee@anu.edu.au

Peter L. Bartlett
Dept. of Systems Engineering,
RSISE, Aust. Nat. University,
Canberra, ACT 0200, Australia.
Peter.Bartlett@anu.edu.au

Robert C. Williamson
Department of Engineering,
Australian National University,
Canberra, ACT 0200, Australia.
Bob.Williamson@anu.edu.au

June 20, 1997

Abstract

We examine the relationship between the VC-dimension and the number of parameters of a thresholded smoothly parametrized function class. We show that the VC-dimension of such a function class is at least k if there exists a k-dimensional differentiable manifold in the parameter space such that each member of the manifold corresponds to a different decision boundary. Using this result, we are able to obtain lower bounds on the VC-dimension proportional to the number of parameters for several thresholded function classes including two-layer neural networks with certain smooth activation functions and radial basis functions with a gaussian basis. These lower bounds hold even if the magnitudes of the parameters are restricted to be arbitrarily small. In Valiant's probably approximately correct learning framework, this implies that the number of examples necessary for learning these function classes is at least linear in the number of parameters.

1 Introduction

Smoothly parametrized functions are often used as classification functions by thresholding the outputs to create binary valued functions. This is done because differentiability allows the use of gradient based algorithms in learning the functions. Examples of frequently used smoothly parametrized functions include feedforward neural networks with sigmoidal activation functions such as tanh and radial basis functions with a gaussian basis.

In considering the number of examples necessary to learn these functions, we utilise Valiant's probably approximately correct (PAC) framework [Valiant 1984]. In this framework, for any desired target function and any probability distribution of examples, the learning algorithm is required to produce with high probability a hypothesis that classifies most of the randomly chosen examples correctly. It has been shown in [Blumer *et al.* 1989] that the number of examples necessary and sufficient for PAC-learning a function class is proportional to a combinatorial dimension known as the Vapnik-Chervonenkis (VC-) dimension of the function class.

Definition 1 *Let F be a class of $\{0,1\}$-valued functions defined on a set X. A finite set $S \subset X$*

is said to be shattered if for any subset S^+ of S, there is an $f \in F$ such that $f(x) = 1$ for all $x \in S^+$ and $f(y) = 0$ for all $y \in S \backslash S^+$. The VC-dimension of F is the cardinality of the largest subset of X which is shattered by F.

Bounds on the VC-dimension of several specific parametrized function classes are known. For example, the class of thresholded functions formed from a vector space of real valued functions of dimension d is known to have VC-dimension d [Dudley 1978]. This includes functions formed from linear combinations of linearly independent *fixed* basis functions such as polynomials and radial basis functions with fixed bases. Other classes with known bounds include certain neural networks with threshold activation functions with VC-dimension $\Theta(W \log W)$ [Baum and Haussler 1989, Maass 1992, Sakurai 1993] and networks with piecewise polynomial activation functions which have VC-dimension $O(W^2 \log q)$ where W is the number of parameters and q is the number of pieces [Maass 1993, Goldberg and Jerrum 1993].

In this paper, we consider thresholded smoothly parametrized function classes. We give general lower bounds on the VC-dimension of a thresholded smoothly parametrized function class in terms of the number of "useful" parameters. Obviously, function classes can be parametrized in such a way that a lot of the parameters are redundant. For example, a neural network where the activation functions are linear has VC-dimension no more than the number of inputs plus one regardless of the number of parameters used. Similarly, a network with a $\tanh(x)$ activation function but only one unit in each hidden layer has the same decision boundary as a linear function regardless of the number of layers used. We show that if there exists a k-dimensional differentiable manifold in the parameter space such that each member of the manifold corresponds to a different decision boundary, then the VC-dimension is at least k.

Using this result, we find lower bounds for the VC-dimension of some two layer neural networks. Certain two- and three-layer feedforward neural networks with $\tanh(x)$ activation function are known to have VC-dimension at least $\Omega(W \log W)$. This bound is obtained simply by letting the $\tanh(x)$ network approximate a network with the same architecture, but threshold activation function (with VC-dimension $\Theta(W \log W)$ [Baum and Haussler 1989, Maass 1992, Sakurai 1993]) when the weights are large enough. However, if the inputs and weights are bounded such that the input-output map is "nearly linear", these techniques will no longer provide the bounds. Notable increases in performance have also been observed in experiments when the norm of the weights is minimized along with the empirical error [Hertz *et al.* 1991]. Heuristic explanations for this include suggestions that the VC-dimension of such networks decreases significantly and approaches that of a linear classifier (see for example [Boser *et al.* 1992]) as the weights are constrained to be small. We give a lower bound on the VC-dimension proportional to the number of weights even when the inputs and weights are restricted to arbitrary small open sets around the origin. This shows that the VC-dimension of the network does *not* approach the VC-dimension of a linear classifier as the allowable size of the weights is reduced. In Valiant's PAC framework, the number of examples required for learning this class of neural networks remains at least proportional to the number of weights regardless of bounds on the size of the weights and inputs.

Previous results bounding the VC-dimension of neural networks from below that we are aware of hold only for certain networks with activation functions which can approximate the threshold function [Baum and Haussler 1989, Bartlett 1993, Maass 1992, Sakurai 1993]. We are able to give lower bounds proportional to the number of weights for neural networks with a large class of analytic activation functions which are not necessarily able to approximate the threshold function. These techniques also give a lower bound for radial basis functions with a gaussian basis when the centers are adaptable. This bound is approximately n (where n is the input dimension) times better than the best previous lower bound, which follows from the bound in [Anthony and Holden 1993] for gaussian radial basis functions with fixed centers.

2 Parametrized Function Classes

A function $f: A \times X \to \mathbf{R}$ can be used to form a parametrized function class F by letting each parameter $a \in A$ define a function in the class F. An example of this is an artificial neural network where A is the set of weights and X is the set of inputs. Such functions are used as classification functions by thresholding the output.

Definition 2 *Let A be an open subset of $\mathbf{R}^m$ and X be an open subset of $\mathbf{R}^n$. Let $f: A \times X \to \mathbf{R}$ be some (fixed) continuous function. We use f to define decision regions D_a^+ and D_a^- by*

$$\begin{aligned} D_a^+ &:= \{x \in X: f(a,x) > 0\}, \\ D_a^- &:= \{x \in X: f(a,x) < 0\} \end{aligned}$$

The region of input space where the function is positive is separated from the region where it is negative by the decision boundary.

Definition 3 *The* boundary *of $a \in A$ in the open set $X \subset \mathbf{R}^n$, denoted* $\mathrm{bdy}(a)$ *is defined by*

$$\mathrm{bdy}(a) := X \setminus (D_a^+ \cup D_a^-) = \{x \in X : f(a,x) = 0\}.$$

For thresholded real valued function classes, the following definition of the VC-dimension is useful.

Definition 4 *Let* $\theta: \mathbf{R} \rightarrow \{0,1\}$ *be defined by* $\theta(x) = 1$ *if* $x > 0$, *and* $\theta(x) = 0$ *otherwise. Let* Z *be some set and let* F *be a class of functions from* Z *to* $\mathbf{R}$. *The thresholded function class formed from the function class* F *is the class* $F_\theta = \{\theta \circ f : f \in F\}$.

Let $x = (x_1, \ldots, x_m) \in \mathbf{R}^m$, *and let* $\Theta(x) = (\theta(x_1), \ldots, \theta(x_m))$. *Let* $T \subset \mathbf{R}^m$, *and write* $\Theta(T) = \{\Theta(x): x \in T\}$. *For any sequence,* $z = (z_1, \ldots, z_m) \in Z^m$, *let* $F_{|z} = \{(f(z_1), \ldots, f(z_m)): f \in F\}$. *The VC-dimension of* F_θ *is*

$$\mathrm{VCdim}(F_\theta) := \sup\{m: \exists\ z \in Z^m, |\Theta(F_{|z})| = 2^m\}.$$

We will need the inverse function theorem and implicit function theorem from calculus (see [Spivak 1965]).

Definition 5 *For* $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, $f(a_1, \ldots, a_n) = (f_1(a), \ldots, f_m(a))^T$, *let* $Df(a)$ *denote the Jacobian matrix of* f *at* a, *where the Jacobian matrix is the* $m \times n$ *matrix:*

$$Df(a) = \begin{bmatrix} \frac{\partial f_1(a)}{\partial a_1} & \frac{\partial f_1(a)}{\partial a_2} & \cdots & \frac{\partial f_1(a)}{\partial a_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m(a)}{\partial a_1} & \frac{\partial f_m(a)}{\partial a_2} & \cdots & \frac{\partial f_m(a)}{\partial a_n} \end{bmatrix}$$

For $f: \mathbf{R}^m \times \mathbf{R}^n \rightarrow \mathbf{R}^m$, *let* $D_a f(a,b)$ *denote the* $m \times m$ *matrix comprising the first* m *columns of the Jacobian matrix at* (a,b) *and* $D_b f(a,b)$ *denote the* $m \times n$ *matrix comprising the last* n *columns of the Jacobian matrix at* (a,b).

Theorem 6 (Inverse Function Theorem)

Let $a \in \mathbf{R}^n$ and suppose that $f: \mathbf{R}^n \to \mathbf{R}^n$ is continuously differentiable in an open set containing a, and $\det(Df(a)) \neq 0$. Then there is an open set V containing a and an open set W containing $f(a)$ such that $f: V \to W$ has a continuous inverse $f^{-1}: W \to V$ which is differentiable and for all $y \in W$ satisfies $Df^{-1}(y) = [Df(f^{-1}(y))]^{-1}$.

Theorem 7 (Implicit Function Theorem)

Suppose $f: \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R}^m$ is continuously differentiable on an open set containing (a, b) and $f(a, b) = 0$. Let $\mathcal{M}$ be the $m \times m$ matrix $D_a f(a, b))$. If $\det(\mathcal{M}) \neq 0$, there is an open set $B \subset \mathbf{R}^n$ containing b and an open set $A \subset \mathbf{R}^m$ containing a, with the following property: for each $x \in B$, there is a unique $y \in A$ such that $f(y, x) = 0$. The function $h: x \mapsto y$ is differentiable.

We want to relate the VC-dimension of the thresholded function class to the number of parameters which are not redundant in some sense. Hence, we will consider subsets of the parameter space which form differentiable manifolds in the space such that each member of the manifold defines a different boundary. For our purposes, C^1-manifolds will be sufficient.

Definition 8

1. *If U and V are open sets in $\mathbf{R}^n$, a continuously differentiable function $h: U \to V$ with a continuously differentiable inverse $h^{-1}: V \to U$ is called a* diffeomorphism.

2. *A subset M of $\mathbf{R}^n$ is called a k-*dimensional manifold *(in $\mathbf{R}^n$) if for every point $x \in M$, there is an open set U containing x, an open set $V \subset \mathbf{R}^n$, and a diffeomorphism $h: U \to V$ such that $h(U \cap M) = V \cap (\mathbf{R}^k \times \{0\}) = \{y \in V : y^{k+1} = \cdots = y^n = 0\}$.*

3. *If for all* $a_1, a_2 \in M, a_1 \neq a_2 \Rightarrow \mathrm{bdy}(a_1) \neq \mathrm{bdy}(a_2)$ *then we say that* M *has unique decision boundaries.*

Let the function class F be $\{f(a, \cdot) : a \in A\}$, where A is an open subset of $\mathbf{R}^m$ and f is continuously differentiable. Let $g: A \times X^m \to \mathbf{R}^m$ be defined by $g(a, x_1, \ldots, x_m) = (f(a, x_1), \ldots, f(a, x_m))^T$. For a fixed x, define $g_x(a) = g(a, x)$. The next lemma is a simple consequence of the Inverse Function Theorem.

Lemma 9 *Let* φ *be any diffeomorphism from an open subset of* A *to a subset of* $\mathbf{R}^m$ *and* $\psi_x = g_x \circ \varphi^{-1}$*. If* $\mathrm{VCdim}(F_\theta) < k$*, where* F_θ *is the thresholded function class formed from* F*, then for every* $a \in A$*, for every* $x \in X^m$*,* $g_x(a) = 0 \Rightarrow \mathrm{rank}(D\psi_x(b)) < k$*, where* $b = \varphi(a)$*.*

Proof. Suppose $\mathrm{VCdim}(F_\theta) < k$ but there is an $a \in A$, an $x \in X$ and a diffeomorphism $\varphi(a) = b$ such that $g(\varphi^{-1}(b), x) = 0$ but $\mathrm{rank}(D\psi_x(b)) \geq k$. Because the rank is at least k, we can choose a submatrix of k linearly independent rows from the matrix $D\psi_x(b)$ corresponding to k examples from X. We can now choose k linearly independent columns from the previous submatrix corresponding to k parameters. Let $x' \in X^k$ be the k components of x corresponding to the k linearly independent rows of $D\psi_x(b)$ we picked. Let $b' \in \mathbf{R}^k$ be the k components of b corresponding to the k columns. Define $h_{x'}: \mathbf{R}^k \to \mathbf{R}^k$ so that $h_{x'}(b')$ comprises the k components of $\psi_x(b)$ corresponding to those k rows (the appropriate components of b are fixed). Then the $k \times k$ matrix is $Dh_{x'}(b')$. By the Inverse Function Theorem, $h_{x'}$ has a continuous inverse at $h_{x'}(b') = 0$. Note that a function $y: X \to Y$ is continuous if and only if the inverse image of any open subset of Y is an open subset of X. Since the inverse of $h_{x'}$ at $h_{x'}(b') = 0$ is continuous, the inverse image of an open subset in $\mathbf{R}^k$ containing b' is an open subset containing $h_{x'}(b') = 0$.

So we can pick 2^k points, one from each orthant in a small enough open subset containing $h_{x'}(b') = 0$. The inverse of $h_{x'}$ at these 2^k points gives us the 2^k points in the neighbourhood of b'. Applying φ^{-1} on the appropriate points gives us the 2^k functions in A required to get $|H(F_{|x})| = 2^k$. So the VC-dimension must be at least k, contradicting $\mathrm{VCdim}(F_\theta) < k$. $\square$

From the previous lemma, it is obvious that to show that the VC-dimension is at least k, all we have to do is to pick a parameter and k points from its decision boundary such that the rank of the corresponding Jacobian matrix is k. (This technique was used in [Bartlett 1993] to give lower bounds on the VC-dimension of neural networks with threshold activation functions.) However, this may not be easy to do. The following theorem gives conditions which may be easier to check in some cases.

Theorem 10 *Let A be an open subset of $\mathbf{R}^m$, X be an open subset of $\mathbf{R}^n$ and $f : A \times X \to \mathbf{R}$ be a continuously differentiable function (in all of its arguments). Let $F := \{f(a, \cdot) : a \in A\}$. If there exists a k-dimensional manifold $M \subset A$ which has unique decision boundaries, then $\mathrm{VCdim}(F_\theta) \geq k$.*

Proof. VC-dimension is always greater than or equal to zero so the case $k = 0$ is trivial.

Assume the manifold with unique boundaries M is of dimension $k \geq 1$ but $\mathrm{VCdim}(F_\theta) < k$. Recall that $g : A \times X^m \to \mathbf{R}^m$ is defined by $g(a, x_1, \ldots, x_m) = (f(a, x_1), \ldots, f(a, x_m))^T$, $g_x(a) = g(a, x)$ and $\psi_x = g_x \circ \varphi^{-1}$ where φ is an appropriate diffeomorphism that defines the manifold M at a. Then Lemma 9 implies that for every $a \in M$ and $b = \varphi(a)$, for every $x \in X^m$, $g(a, x) = 0 \Rightarrow \mathrm{rank}(D\psi_x(b)) < k$. Let $b = (w, 0)$ and $y_x : \mathbf{R}^k \to \mathbf{R}^m$ be defined by $y_x(\xi) = \psi_x(\xi, 0)$. Then $\mathrm{rank}(Dy_x(w)) < k$ as well.

Pick $a \in M$ and $x \in X^m$ such that $g(a,x) = 0$ and $r = \text{rank}(Dy_x(w)) < k$ is the largest possible. The rank r is greater than zero, because if it were not so, the boundary could not change as we change w. By permuting x, w and $y_x(w)$, x becomes x', w becomes $w' = (c,d)$, $y_x(w)$ becomes $y(w') = (\alpha(w'), \beta(w'))$, and $Dy_x(w)$ becomes

$$Dy_{x'}(w') = \begin{bmatrix} D_c\alpha(c,d) & D_d\alpha(c,d) \\ D_c\beta(c,d) & D_d\beta(c,d) \end{bmatrix}$$

where α and c each have r components and the $r \times r$ matrix $D_c\alpha(c,d)$ is nonsingular. By the Implicit Function Theorem, with $\alpha(c,d) = 0$, there exists an open set E around d, and a continuously differentiable function $h: E \to \mathbf{R}^r$ such that $\alpha(h(d),d) = 0$ for each $d \in E$.

We will now show that $\beta(h(d),d)$ is also zero for each $d \in E$. Because $Dy_{x'}(w')$ is of rank r, there exists some matrix K such that $D_c\beta(c,d) = KD_c\alpha(c,d)$ and $D_d\beta(c,d) = KD_d\alpha(c,d)$. Differentiating $\beta(h(d),d)$ with respect to d by the chain rule:

$$\begin{aligned} D_d\beta(h(d),d) &= D_{h(d)}\beta D_d h + D_d\beta \\ &= KD_{h(d)}\alpha D_d h + KD_d\alpha \\ &= K(D_{h(d)}\alpha D_d h + D_d\alpha) \\ &= K(D_d\alpha(h(d),d)) \\ &= 0 \end{aligned}$$

(because $\alpha(h(d),d) = 0$ for all $d \in E$). Since $\beta(w') = 0$, $\beta(h(d),d) = 0$ for all $d \in E$.

Let $\beta(w') = (f'(w',x_1'), \ldots, f'(w',x_{m-r}'))^T$. We can substitute an arbitrary x' from $\text{bdy}(w')$ for one of these x_i' without increasing the rank of $Dy_{x'}(w')$ because by hypothesis we have picked the x_i''s which give the largest rank. Hence the rank of Jacobian matrix of the function after substitution remains r. Differentiating $\beta(w')$ (with the new x_i') using the chain rule as above, we find that $f'(w'', x_i') = 0$ for any $w'' \in \mathrm{graph}(h)$, where $\mathrm{graph}(h) = \{(d, h(d)) : d \in E\}$. Since we have picked x_i' arbitrarily from $\mathrm{bdy}(w')$, $\mathrm{bdy}(w') \subseteq \mathrm{bdy}(w'')$ for any $w'' \in \mathrm{graph}(h)$.

From the continuity of the components of $D_c\alpha(w')$, and hence of its determinant, $D_c\alpha(w')$ is nonsingular in a neighbourhood of w'. Choose $w'' \in \mathrm{graph}(h)$ which is also in this neighbourhood. Again, we can substitute any x'' from the boundary of w'' into $\beta(w'')$ without changing the rank of the Jacobian matrix. Since w' and w'' are both in E, we can again differentiate using the chain rule to show that $\mathrm{bdy}(w'') \subseteq \mathrm{bdy}(w')$. So w' and w'' must have the same boundary, which is a contradiction.

So if $\mathrm{VCdim}(F_\theta) < k$, any k-dimensional manifold M must contain distinct $a' = \varphi^{-1}(b')$ and $a'' = \varphi^{-1}(b'')$ such that $\mathrm{bdy}(a') = \mathrm{bdy}(a'')$, where φ is the diffeomorphism that defines M at a'.
$\square$

As a simple example we consider the VC-dimension of the linear classifier (perceptron) when the parameters are restricted to an open set.

Example 11 *Consider the linear function $f : A \times \mathbf{R}^n \to \mathbf{R}$ such that $f(a, x) = a_0 + a_1x_1 + \cdots + a_nx_n$ where A is any open subset of $\mathbf{R}^{n+1}$. Choose an open subset $A' \subset A$ such that none of the parameters, $(a_0, \ldots, a_n)$, are zero. This can always be done because A is an open set. Let M be the projection of A' onto the subspace where $a_0 \neq 0$ is constant. Then M is an n-dimensional manifold with unique decision boundaries. One way to see that the boundaries are unique is to check the intersection of the boundaries with the axes of $\mathbf{R}^n$. Using Theorem 10, the VC-dimension of the thresholded linear function class is at least n.*

3 Two Layer Neural Networks

3.1 Tanh Activation Function

We first consider finding a lower bound for the VC-dimension of a two layer neural network with $\tanh(x)$ activation functions when both the inputs and weights (parameters) are restricted to an arbitrary open subset which includes the origin.

Definition 12 *A* two layer feedforward network *with an n-dimensional input $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$, k hidden units with* tanh *activation, and weights $w = (v_{10}, \ldots, v_{kn}, w_0, \ldots, w_k) \in \mathbf{R}^W$ (where $W = kn + 2k + 1$) is a function $f : \mathbf{R}^W \times \mathbf{R}^n \to \mathbf{R}$ given by*

$$f(w, x) = w_0 + \sum_{i=1}^{k} w_i \ \tanh(v_i \cdot x + v_{i0}),$$

where $v_i = (v_{i1}, \ldots, v_{in})$ and $v_i \cdot x = \sum_{j=1}^{n} v_{ij} x_j$. The weights $w_0, v_{i0}, i = 1, \ldots, k$ are called the offsets.

By permuting the hidden units, the input-output map of the network remains unchanged. Furthermore, because the tanh activation function is odd, changing the sign of all the input weights, the bias and the output weight of a hidden unit will not change the input-output map. It has shown that any two networks with the same input-output behaviour must be related by a transformation from the finite group generated from these transformations provided that the networks are irreducible [Sussmann 1992]. A net is reducible if any of the following conditions hold:

1. $w_i = 0$ for some $i = 1, \ldots, k$;

2. there exist two different indices j_1, $j_2 \in \{1,\ldots,k\}$ such that the $|y_{j_1}(x)| = |y_{j_2}(x)|$ for all $x \in \mathbf{R}^n$ where $y_j(x) = v_j \cdot x + v_0$; or

3. $v_i = 0$ for some $i = 1,\ldots,k$.

This means that the input-output maps of the networks are essentially unique up to a finite group of transformations. Unfortunately, uniqueness of input-output map is not sufficient (although it is necessary) for uniqueness of decision boundaries. For example, multiplying the function by a constant results in a network with the same decision boundary but different parameters. Although we do not know the largest dimension for a manifold with unique boundaries in the parameter space, we can use Sussmann's result to find such a manifold of dimension not too much smaller than the number of parameters.

Theorem 13 *Let F be the class of two layer feedforward networks with k hidden units with tanh activation, input space $X := \{(x_1,\ldots,x_n) \in \mathbf{R}^n : |x_i| < C\}$ (where C is a constant greater than zero), and $k(n+2)+1$ weights restricted to an open set which includes the origin. Then* $\mathrm{VCdim}(F_\theta) \geq \mu$ *where $\mu = (k-1)(n+1)+1$ is the number of weights of a network with $n-1$ inputs and $k-1$ hidden units.*

Proof. Delete all the weights from the nth input and all the weights connected to the kth hidden unit except the weight connecting the two of them as shown in Figure 1. Let the smaller network (with weights deleted) be

$$\begin{aligned} f(w,x) &= w_0 + \sum_{i=1}^{k-1} w_i \ \tanh(v_i' \cdot x' + v_{i0}) \\ &\quad + w_k \ \tanh(v_{kn} x_n) \\ &= g(w',x') + w_k \ \tanh(v_{kn} x_n) \end{aligned}$$

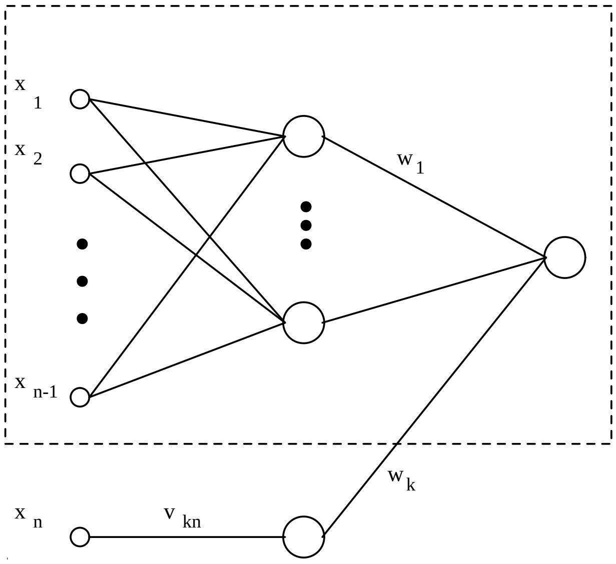

Figure 1: Network with weights deleted

where $w' = (v_{10}, \ldots, v_{k-1,n-1}, w_0, \ldots, w_{k-1})$, $v_i' = (v_{i1}, \ldots, v_{i,n-1})$ and $x' = (x_1, \ldots, x_{n-1})$. We will also fix w_k and v_{kn} to be constants. This is equivalent to working on a manifold where the codimension is the number of fixed and deleted weights.

Now, $g(w', x')$ is a network with $n-1$ inputs and $k-1$ hidden units (the network within the box in Figure 1). Since nets which are reducible are not dense anywhere in the parameter space and the number of different parameter values with the same input-output map for irreducible nets is finite, we can always find an open subset of weights such that the input-output map is unique for all the parameter values in the subset. The input-output map of $g(w', \cdot)$ is unique not only over the whole of $\mathbf{R}^{n-1}$ (as shown in [Sussmann 1992]) but also for any open set of x' values (in particular, the open set satisfying $|x_i'| < C, i = 1, \ldots, n-1$) because $g(w', \cdot)$ is an analytic function. We can also choose the open subset of weights such that the boundary exists for all weights in the set; i.e. choose an open set of $W' \times X'$ such that the outputs are in the range of $h : x_n \mapsto -w_k \tanh(v_{kn} x_n)$.

When the output of f is zero, we have $g(w', x') = -w_k \ \tanh(v_{kn} x_n)$. Fix w_k and v_{kn} so that they are not adjustable. Then the boundary of f is graph$(\tanh^{-1}(-g(w', \cdot)/w_k)/v_{kn})$. We will show that this is unique for each w' in the open set of weights.

Let R be the range of h, where $h : x_n \mapsto -w_k \ \tanh(v_{kn} \ x_n)$. Because $g(w', \cdot)$ is a continuous function, $B := g^{-1}(w', R)$ is an open set of X'. The boundary of f can only exist for $x' \in B$. Since the input-output map is unique for $g(w', \cdot)$ in domain B for each w', graph(g) is unique (for domain B) for each w'. This implies that the boundary of f is unique since $\tanh^{-1}$ is a one-to-one function.

So we have found a manifold of unique boundaries the dimension of which is the number of weights in a net with $n-1$ inputs and $k-1$ hidden units. Theorem 10 then gives the desired result. $\square$

Since a network with the standard sigmoid activation $1/(1+e^{-x})$ is equivalent to a $\tanh(x)$ net up to translation and change of coordinates of the weights, the same result holds for networks with the standard sigmoid activation. It would be interesting to know if the VC-dimension of an arbitrary open set of parameters (which does not necessarily include the origin) is also proportional to the number of parameters (when boundaries exist).

3.2 Other Activation Functions

Similar bounds can be found using the same techniques for networks with other analytic activation functions if the networks have unique input-output mappings up to a finite group of transformations when they are irreducible. It has been shown that odd activation functions which satisfy the *independence property* (**IP**) have this property [Albertini *et al.* 1993]. For networks with no offset, the *weak independence property* (**WIP**) is sufficient.

Definition 14 *The function* $\sigma: \mathbf{R} \to \mathbf{R}$ *satisfies the* independence property (IP) *if, for every positive integer* l*, for any nonzero real numbers* $b_1, \ldots, b_l$*, and for any real numbers* $\beta_1, \ldots, \beta_l$ *for which*

$$(b_i, \beta_i) \neq \pm(b_j, \beta_j) \ \forall i \neq j,$$

implies that the functions

$$1, x \mapsto \sigma(b_1 x + \beta_1), \ldots, x \mapsto \sigma(b_l x + \beta_l)$$

are linearly independent (where $x \in \mathbf{R}$*). The function* σ *satisfies the* weak independence property (WIP) *if the above linear independence property holds for all pairs* (b_i, β_i) *with* $\beta_i = 0, i = 1, \ldots, l$*.*

Obviously IP implies WIP. The following conditions for IP and WIP are from [Albertini *et al.* 1993].

Lemma 15 *If* σ *is a polynomial, WIP does not hold. If* σ *is odd, infinitely differentiable and* $\sigma^{(k)}(0) \neq 0$ *for an infinite number of values of* k *then* σ *satisfies the property WIP.*

Lemma 16 *Assume that* σ *is a real-analytic function, and it extends to a function* $\sigma: \mathbf{C} \to \mathbf{C}$ *analytic on a subset* $D \subset \mathbf{C}$ *of the form*

$$D = \{z \in \mathbf{C}: |\mathrm{Im}\ z| \leq \lambda\} \backslash \{z_0, \overline{z_0}\}$$

for some $\infty > \lambda > 0$*. Here* $\overline{z_0}$ *is the complex conjugate of* z_0*,* $\mathrm{Im}\ z_0 = \lambda$ *and* z_0 *and* $\overline{z_0}$ *are singularities, that is, there is a sequence* $z_n \to z_0$ *so that* $|\sigma(z_n)| \to \infty$*, and similarly for* $\overline{z_0}$*. Then,* σ *satisfies property IP.*

Functions which satisfy the property IP include the tanh function considered earlier. Most rational functions also satisfy this property.

Theorem 17 *Let F be a two layer neural network with k hidden units and an activation function which is odd, real analytic and satisfies the property IP. If the input space is $\mathbf{R}^n$, then* $\mathrm{VCdim}(F_\theta) \geq \mu$, *where μ is the number of weights in a network with $n-1$ inputs and $k-1$ hidden units. For a network with no offsets it is sufficient if the activation function is odd, real analytic and satisfies WIP.*

The proof is essentially the same as for the tanh activation function.

3.3 Radial Basis Functions

Another smoothly parametrized function class commonly used for classification is the radial basis function with a gaussian basis.

Definition 18 *A k-*term radial basis function *with n inputs, $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$, gaussian basis functions and parameters $w = (c_{11}, \ldots, c_{kn}, w_1, \ldots, w_k) \in \mathbf{R}^W$ (where $W = kn + k$) is a function $f : \mathbf{R}^W \times \mathbf{R}^n \to \mathbf{R}$ given by*

$$f(w, x) = \sum_{i=1}^{k} w_i \; e^{-\|x - c_i\|^2},$$

where $c_i = (c_{i1}, \ldots, c_{in})$ are the centers, *$w_i \neq 0$, $i = 1, \ldots, k$ and $\|\cdot\|$ denotes the Euclidean norm.*

The VC-dimension of a k-term radial basis function has been shown to be at least k [Anthony and Holden 1993]. This bound is tight when the centres are not adjustable. When the centres are adjustable, we give a lower bound of $kn - n$.

First we will need a result on uniqueness of input-output mappings similar to that for the tanh network. It is well known (see e.g. [Powell 1987]) that for any $l > 0$ and any input dimension, the functions

$$x \mapsto e^{-\|x-c_1\|^2}, \ldots, x \mapsto e^{-\|x-c_l\|^2}$$

are linearly independent provided the centres are distinct. This implies that the input-output mappings are unique up to permutation of the centers if none of the w_i's are zero.

Theorem 19 *Let F be a k-term radial basis function with gaussian basis functions. If the input space is $\mathbf{R}^n$, then $\mathrm{VCdim}(F_\theta) \geq \mu$, where $\mu = kn - n$ is the number of parameters in a $k-1$-term radial basis function with $n-1$ inputs.*

Proof. As in the proof of Theorem 13 we will work on a manifold formed by projecting onto the subspace where some parameters are either zero or constant. Set $c_{i1} = 0$ for $i \neq 1$, $c_{1j} = 0$ for $j \neq 1$ and fix w_1 and c_{11} to non-zero constants. So at a boundary we have

$$\begin{aligned} w_1 \exp - \left(x_1^2 - 2x_1 c_{11} + c_{11}^2 + \sum_{j=2}^{n} x_j^2 \right) \\ + \sum_{i=2}^{k} w_i \exp - \left(x_1^2 + \|x' - c_i'\|^2 \right) = 0 \end{aligned}$$

where $x' = (x_2, \ldots, x_n)$ and $c_i' = (c_{i2}, \ldots, c_{in})$, which implies

$$\begin{aligned} \exp\left(-x_1^2\right) \Bigg[w_1 \exp \left(- \sum_{j=2}^{n} x_j^2 - c_{11}^2 + 2c_{11}x_1 \right) \\ + \sum_{i=2}^{k} w_i \exp\left(-\|x' - c_i'\|^2\right) \Bigg] = 0 \end{aligned}$$

$$\Leftrightarrow \qquad \sum_{i=2}^{k} w_i \exp\left(-\|x' - c_i'\|^2\right) =$$

$$-w_1 \exp\left(-\sum_{j=2}^{n} x_j^2 - c_{11}^2\right) \exp(2c_{11}x_1)$$

$$\Leftrightarrow \qquad x_1 = \frac{1}{2c_{11}} \log\left[-\frac{\sum_{i=2}^{k} w_i \exp\left(-\|x' - c_i'\|^2\right)}{w_1 \exp\left(-\sum_{j=2}^{n} x_j^2 - c_{11}^2\right)}\right]$$

The argument of the log function will be positive by the assumption that x lies on a boundary. Since the log function is one-to-one and $(k-1)$-term radial basis functions have unique input-output maps, the boundaries are unique where they exist. The log and exp functions are continuous, so for x_1 in an open interval, the regions of input and (adjustable) parameter space where the boundaries exist are open sets. □

4 Conclusions

We have derived a relationship between the number of "useful" parameters in a class of smooth functions and its VC-dimension. Using this relationship, we have obtained lower bounds on the VC-dimension proportional to the number of parameters for neural networks with tanh activation functions when the weights and inputs are restricted to be in an arbitrarily small open set which includes the origin. It would be interesting to know if this is also true for any open set of parameters that does not include the origin (if the decision boundaries exist for the set of parameters, otherwise the VC-dimension is trivially zero). To do that using the same techniques would require solving the boundary uniqueness problem under such conditions. We have also obtained lower bounds on the VC-dimension proportional to the number of parameters for networks with certain real analytic activation functions which are not necessarily sigmoids. For radial basis functions with a gaussian basis, we obtained bounds proportional to the number of parameters. We expect the same results to hold for other smooth radial basis functions with nonpolynomial basis [Anthony and Holden 1993, Powell 1987] but proving this using the same techniques would require proving boundary uniqueness for the functions.

5 Acknowledgements

This research was supported by the Australian Research Council and the Australian Telecommunications and Electronics Research Board. We would like to thank Adam Kowalczyk for helpful comments.

References

[Albertini *et al.* 1993] F. Albertini, E.D. Sontag and V. Maillot, "Uniqueness of weights for neural networks", preprint, 1993.

[Anthony and Holden 1993] M. Anthony and S.B. Holden, "On the Power of Polynomial Discriminators and Radial Basis Function Networks", In *Proceedings of the Sixth Workshop on Computational Learning Theory*, pp. 158-164, 1993.

[Baum and Haussler 1989] E.B. Baum and D. Haussler, "What Size Net Gives Valid Generalization?", *Neural Computation*, 1, pp. 151-160, 1989.

[Bartlett 1993] P. L. Bartlett, "Lower Bounds on the Vapnik-Chervonenkis Dimension of Multi-layer Threshold Networks", In *Proceedings of the Sixth Workshop on Computational Learning Theory*, pp. 144-150, 1993.

[Boser *et al.* 1992] B. Boser, I. Guyon and V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pp 144-152, 1992.

[Blumer *et al.* 1989] A. Blumer, A. Ehrenfeucht, D. Haussler and M.K. Warmuth, "Learnibility and the Vapnik-Chervonenkis Dimension", *Journal of the Association for Computing Machinery*, 36, no. 4, pp. 929-965, October 1989.

[Dudley 1978] R.M. Dudley, "Central Limit Theorems for Empirical Measures", *Annals of Probability*, 6, no. 6, pp. 899-929, 1978.

[Goldberg and Jerrum 1993] P. Goldberg, M. Jerrum, "Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers", In *Proceedings of the Sixth Workshop on Computational Learning Theory*, pp. 361-369, 1993.

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

[Maass 1992] W. Maass, "Neural Nets with Superlinear VC-Dimension", *Neural Computation*, 6, pp. 877-884, 1994.

[Maass 1993] W. Maass, "Agnostic PAC-Learning of Functions on Analog Neural Nets", preprint, 1993.

[Powell 1987] M.J.D. Powell, "Radial Basis Functions For Multivariable Interpolation: A Review", *Algorithms for Approximation*, eds. J.C. Mason and M.E. Cox, pp. 143-167, Clarendon Press, Oxford, 1987.

[Sakurai 1993] A. Sakurai, "Tighter Bounds of the VC-Dimension of Three-layer Networks", In *Proceedings of the World Congress on Neural Networks*, 1993.

[Spivak 1965] M. Spivak, *Calculus on Manifolds.* Benjamin Cummings, Menlo Park, 1965.

[Sussmann 1992] H.J. Sussmann, "Uniqueness of the Weights for Minimal Feedforward Nets with a Given Input-Output Map", *Neural Networks* 5, pp. 589-593, 1992.

[Valiant 1984] L. G. Valiant, "A Theory of the Learnable", *Communications of the ACM*, 27, no. 11, pp. 1134-1143, November 1984.