Lower Bounds on the Vapnik-Chervonenkis Dimension of Multi-layer Threshold Networks

Peter L. Bartlett
Department of Systems Engineering
Australian National University
PO Box 4, Canberra 2601
AUSTRALIA
(plb101@syseng.anu.edu.au)

Abstract

We consider the problem of learning in multi-layer feed-forward networks of linear threshold units. We show that the Vapnik-Chervonenkis dimension of the class of functions that can be computed by a two-layer threshold network with real inputs is at least proportional to the number of weights in the network. This result also holds for a large class of two-layer networks with binary inputs, and a large class of three-layer networks with real inputs. In Valiant's probably approximately correct learning framework, this implies that the number of examples necessary for learning in these networks is at least linear in the number of weights. This bound is within a log factor of the upper bound.

1 INTRODUCTION

Neural networks have been widely used for pattern classification problems. This paper addresses the question, 'How many training examples are necessary for satisfactory learning performance in a multi-layer feed-forward neural network used for classification?' We consider only single-output networks, and therefore two-class classification problems. We assume that the examples are generated randomly, and say that the trained network is approximately correct if it correctly classifies a random example with high probability. We require that, for almost all sequences of training examples, the trained network will be approximately correct, for any desired target function and any probability distribution of examples. This is known as 'probably approximately correct' (or pac) learning [1]. In this framework, the problem of learning consists of choosing an accurate hypothesis from some hypothesis class, such as the class of functions that can be computed on a particular network architecture. In [2], Blumer *et al.* show that the number of examples necessary and sufficient for learning is proportional to a combinatorial dimension of the hypothesis class known as the Vapnik-Chervonenkis dimension.

Definition 1 *Suppose X is a set and H is a class of functions from X to $\{0,1\}$.*

We say that H **shatters** *a finite subset $S \subseteq X$ if, for each of the $2^{|S|}$ possible classifications of the points in S there is a function in H that can perform the classification,*

$$|\{\{x \in S : h(x) = 1\} : h \in H\}| = 2^{|S|}.$$

The **Vapnik-Chervonenkis (VC-) dimension** *of H is the size of the largest shattered subset of X,*

$$\mathrm{VCdim}(H) = \max\{m : \exists S \subseteq X \ |S| = m \text{ and } H \text{ shatters } S\}.$$

If this set has no largest element, we say that $\mathrm{VCdim}(H) = \infty$.

The VC-dimension of the class of functions that can be computed in a feed-forward network is not known precisely. In [3], Baum and Haussler give upper and lower bounds for particular network architectures. They show that, for an arbitrary feed-forward network consisting of N linear threshold units and W weights, the VC-dimension is no more than $2W \log_2 eN$, where e is the base of the natural logarithm. This and Blumer *et al.*'s result show that $O(W \log N)$ training examples provide enough information for pac learning in a feed-forward network.

Baum and Haussler also give a lower bound on the VC-dimension for completely connected two-layer threshold networks. (A completely connected feed-forward network has connections between every pair of units in adjacent layers.) They show that for a network of this kind with k_0 real-valued inputs and k_1 first-layer units, the VC-dimension is at least $2\lfloor k_1/2 \rfloor k_0$ (see also [4]). For a completely connected two-layer network with binary inputs, they show that the VC-dimension is again $\Omega(W)$.

In this paper, we give lower bounds on the VC-dimension for a number of two- and three-layer architectures. In particular, we extend Baum and Haussler's lower bound to arbitrary two-layer networks with real inputs, we improve on their lower bound for binary-input two-layer networks, and we present lower bounds for some completely connected three-layer networks. In all cases, the VC-dimension is $\Omega(W)$.

The remainder of this section defines multi-layer threshold networks, and presents some notation. Section 2 introduces defining sets, which simplify the construction of a shattered set for a class of network functions. In Section 3 and 4, we give results for two- and three-layer networks respectively. Section 5 discusses some extensions to this work. Some of the results in this paper were announced (without proofs) in a note [5].

1.1 NOTATION

We consider networks of processing units in layered, feed-forward architectures.

Definition 2 *A* **feed-forward network architecture** *is a directed graph (U, C), where U is a set, $C \subset U \times U$, and U and C satisfy the following constraints.*

1. *The set U can be partitioned into $L+1$ nonempty, disjoint sets U_i, $U = \bigcup_{i=0}^{L} U_i$, where L is a positive integer.*
2. *If $(u, v) \in C$, where $u \in U_i$ and $v \in U_j$, then $i < j$.*
3. *For all u in $U - U_L$, there is a $v \in U$ such that $(u, v) \in C$.*
4. *For all u in U_i (where $i > 0$), there is a v in U_{i-1} with $(v, u) \in C$.*

We say that (U, C) is an L-layer network architecture.

The function $I : U \to 2^U$ is defined by $I(u) = \{v \in U : (v, u) \in C\}$. A **completely connected** *feed-forward architecture is a feed-forward architecture that also satisfies $I(u) = U_{i-1}$ for all $u \in U_i$ and all $i \in \{1, 2, \ldots, L\}$.*

In this definition, U is the set of units in the network, U_0 is the set of network **input units** (the other units are called **processing units**), U_l is the set of units in the l-th layer from the inputs ($0 \leq l \leq L$), and L is the number of layers of processing units. We write $U_l = \{u_1^l, u_2^l, \ldots, u_{k_l}^l\}$ for $0 \leq l \leq L$, where k_l is the number of units in layer l. The processing units in layer L are called **output units**. In this paper, we are interested in networks with a single output unit, $k_L = 1$.

The set C describes the connections between units in the network. Condition 2 describes the feed-forward requirement. Condition 3 forbids redundant units. Condition 4 requires every non-input unit to have some connection from another unit, and ensures that we cannot shift a unit into a lower layer. It can be shown that there is a unique partition $\{U_l\}$ of U that satisfies these four conditions. Because of this, we can refer to the number of layers L, the sets U_i of units in a layer, and the sizes k_i of a layer ($i = 0, 1, \ldots, L$) without explicitly defining them, since they are uniquely determined by the architecture (U, C).

Definition 3 *A* **feed-forward threshold network** *$N = (U, C, w, \theta)$ is a feed-forward architecture (U, C), together with* **weights** *$w : C \to \mathbb{R}$ and* **thresholds** *$\theta : U - U_0 \to \mathbb{R}$. Given an input vector, $x = (x_1, x_2, \ldots, x_{k_0}) \in \mathbb{R}^{k_0}$, the function $o : \mathbb{R}^{k_0} \times U \to \mathbb{R}$ expresses the values computed by each unit as follows. Each input unit u_i^0 is associated with a real value x_i, and each processing unit computes a thresholded weighted sum of its inputs. That is,*

$$o(x, u_i^0) = x_i,$$

and, if $u \in U - U_0$,

$$o(x, u) = \mathcal{H}\left(\sum_{v \in I(u)} o(x, v) w(v, u) - \theta(u)\right)$$

where $\mathcal{H}$ is the Heaviside function ($\mathcal{H}(\alpha)$ is 1 if $\alpha \geq 0$ and 0 otherwise). The output of the network is the value computed by the output unit u_1^L, so the **network function**, *$F_N : \mathbb{R}^{k_0} \to \mathbb{R}$, is $F_N(x) = o(x, u_1^L)$. Let $T_{\mathbb{R}}(A)$ (respectively $T_{\mathbb{B}}(A)$) be the class of functions that a feed-forward threshold network with architecture $A = (U, C)$ and real-valued ($\{0, 1\}$-valued) inputs can implement.*

In this paper, we are interested in lower bounds on the Vapnik-Chervonenkis dimension of $T_{\mathbb{R}}(A)$ and $T_{\mathbb{B}}(A)$, for various architectures A.

We often need to refer to the weights and threshold of a unit as a vector. Let w_u be the real vector consisting of the threshold $\theta(u)$ and weights $w(v, u)$ associated with the unit u (where some suitable ordering is maintained). Let $w_{\bar{u}}$ be a vector consisting of the concatenation of the vectors $\{w_v : v \in U - U_0 - \{u\}\}$ (again with some suitable ordering on the elements). Let the function $F_u(w_u', w_{\bar{u}}'; \cdot) : \mathbb{R}^{k_0} \to \mathbb{R}$ be the network function when the weight vectors w_u and $w_{\bar{u}}$ are replaced by w_u' and $w_{\bar{u}}'$ respectively.

For a positive integer m, vector $y \in \mathbb{R}^m$, and $\epsilon > 0$, define

$$B_\epsilon(y) = \{z \in \mathbb{R}^m : \|z - y\| < \epsilon\},$$

where $\|x\| = \sqrt{\sum_{i=1}^m x_i^2}$ for $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$.

2 DEFINING SETS AND THE VC-DIMENSION

Let $N = (U, C, w, \theta)$ be a feed-forward threshold network with k_0 input units.

Definition 4 *A set $S = \{x_1, x_2, \ldots, x_n\} \subset X$ (where X is $\mathbb{R}^{k_0}$ or $\mathbb{B}^{k_0}$) is a* **defining set** *for unit $u \in U - U_0$ in the network N if*

1. *We can classify the elements of S arbitrarily by perturbing the weights and threshold of unit u,*

$$\forall \epsilon > 0, \forall (b_1, b_2, \ldots, b_n) \in \{0,1\}^n, \exists w_u^* \in B_\epsilon(w_u),$$
$$\forall i \in \{1, 2, \ldots, n\}, F_u(w_u^*, w_{\bar{u}}; x_i) = b_i.$$

2. *Perturbing the weights and threshold of units other than u will not affect the classification of the points in S,*

$$\forall x \in S, \exists \epsilon > 0, \forall w_u^* \in B_\epsilon(w_u), \exists b \in \{0,1\},$$
$$\forall w_{\bar{u}}^* \in B_\epsilon(w_{\bar{u}}), F_u(w_u^*, w_{\bar{u}}^*; x) = b.$$

To find a lower bound on the VC-dimension of a class of functions, we need to prove the existence of a shattered set. The following theorem shows that the problem of constructing such a set for the classes $T_{\mathbb{R}}(A)$ and $T_{\mathbb{B}}(A)$ can be decomposed by separately constructing defining sets for units in the architecture A.

Theorem 5 *Let $A = (U, C)$ be a feed-forward network architecture, and $N = (U, C, w, \theta)$ a threshold network. If there is a set of processing units $V \subseteq U - U_0$ and a finite defining set S_u for each unit u in V, then*

$$\mathrm{VCdim}(T(A)) \geq \sum_{u \in V} |S_u|,$$

where $T(A)$ is $T_{\mathbb{R}}(A)$ or $T_{\mathbb{B}}(A)$, if the defining sets S_u are subsets of $\mathbb{B}^{k_0}$ or $\mathbb{R}^{k_0}$ respectively.

Proof Since S_u is a defining set for each u in V, for every $x \in \bigcup_{u \in V} S_u$ there is an ϵ that satisfies the criterion given in Condition 2 of Definition 4. Since the S_u are finite, we can always choose a $\delta > 0$ smaller than all of these ϵ's.

Suppose we choose a desired classification for each $x \in \bigcup_{u \in V} S_u$. For each unit $u \in V$, choose $w_u^* \in B_\delta(w_u)$ so that $F_u(w_u^*, w_{\bar{u}}; x)$ is the desired classification for each x in S_u. Condition 1 of the definition of a defining set ensures that this is always possible. By Condition 2, when we apply these changes simultaneously to the weights of all units, the classification of a point $x \in S_u$ is not affected by the modification of weights of units other than u, so all points are classified as desired. Since we can use this argument for any desired classifications of the points in $\bigcup_{u \in V} S_u$, this set is shattered by $T(A)$. □

Theorem 5 can be improved slightly using the idea of an oblivious point.

Definition 6 *Let $N = (U, C, w, \theta)$ be a feed-forward threshold network that computes a mapping from X to $\{0,1\}$, where X is a set. A point $x \in X$ is an* **oblivious point** *for N if the classification of x is unaffected by sufficiently small perturbations of the weights of N.*

Corollary 7 *Let $A = (U, C)$ be a feed-forward network architecture, and $N = (U, C, w, \theta)$ a threshold network with an oblivious point. If there is a set of processing units $V \subseteq U - U_0$ and a finite defining set S_u for each u in V, then*

$$\mathrm{VCdim}(T(A)) \geq \sum_{u \in V} |S_u| + 1.$$

Proof As in the proof of Theorem 5, we can produce any desired classification of the points in $\bigcup_{u \in V} S_u$ by perturbing the weights of the units in V.

Consider the network $\bar{N} = (U, C, \bar{w}, \bar{\theta})$, where $\bar{w}$ and $\bar{\theta}$ are identical to w and θ, but the weights and thresholds of the output unit, u_1^L, are negated, so $\bar{w}_{u_1^L} = -w_{u_1^L}$. Each set S_u is also a defining set for u in $\bar{N}$. Let y be an oblivious point for N. It must be classified differently by the networks N and $\bar{N}$. Therefore $\bigcup_{u \in V} S_u \cup \{y\}$ is shattered by $T(A)$. □

Consider a network N with real-valued inputs. If N has a first-layer unit $u \in U_1$ with a nonempty defining set and $w_u \neq 0$, then N has an oblivious point: choose a point close to an element of the defining set and not on u's separating hyperplane.

3 TWO LAYER NETWORKS

The following theorem improves the lower bound given in [3] for two-layer completely connected networks with binary inputs.

Theorem 8 *Let k_0 and k_1 be positive integers, and let A_{k_0,k_1} be the two-layer completely connected architecture with k_0 input units, k_1 first layer units, and a single output unit. For the class $T_{\mathbb{B}}(A_{k_0,k_1})$ defined in Definition 3, we have*

$$\mathrm{VCdim}(T_{\mathbb{B}}(A_{k_0,k_1})) \geq k_0 \min\left(k_1, \frac{2^{k_0}}{k_0^2/2 + k_0/2 + 1}\right) + 1.$$

Notice that the number of weights W in these networks is $(k_0 + 2)k_1 + 1$, so the VC-dimension is $\Omega(W)$.

A d-packing of the k_0-cube, $\{0,1\}^{k_0}$, is a subset of the k_0-cube in which every pair of vertices is at least Hamming distance d apart. We will use the following lower bound on the size of a maximal d-packing (see, for example, [6]).

Lemma 9 *If $P(n, d)$ is the largest integer i such that there is a d-packing of the n-cube that contains i vectors, then $P(n, d) \geq 2^n / \left(\sum_{j=0}^{d-1} \binom{n}{j}\right)$. In particular,*

$$P(k_0, 3) \geq \frac{2^{k_0}}{k_0^2/2 + k_0/2 + 1}.$$

Proof We can construct a suitable d-packing as follows. Begin with an empty set T, and a set $L = \{0,1\}^n$ of 'legal' vertices. At each step, add to T any vertex v from L, and remove from L all vertices that are within Hamming distance d of v. There are $\sum_{j=0}^{d-1} \binom{n}{j}$ vertices at Hamming distance less than d from v (including v), so at each step there are no more than $\sum_{j=0}^{d-1}\binom{n}{j}$ vertices removed from L. Continuing in this way until L is empty, we construct a d-packing T, of size at least $2^n/\left(\sum_{j=0}^{d-1}\binom{n}{j}\right)$. $\Box$

Proof (of Theorem 8) If $k_1 \leq P(k_0, 3)$, there is a 3-packing $T \subset \{0,1\}^{k_0}$, with $|T| = k_1$. Suppose $T = \{t_1, \ldots, t_{k_1}\}$. For each t_i, $i = 1, \ldots, k_1$, consider the set S_i of all k_0 vertices at Hamming distance 1 from t_i. Choose the weights and threshold of the first-layer unit u_i^1 so that its hyperplane passes through all points in the set S_i. Since T is a 3-packing, these k_1 hyperplanes do not intersect inside $[0,1]^{k_0}$. Choose the signs of the weights and threshold of the first-layer units so that the output of u_i^1 is 0 for the point t_i. If the output unit implements the AND function, then each set S_i is a defining set for u_i^1. Clearly, the point t_1 is an oblivious point, so we have

$$\mathrm{VCdim}(T_{\mathbb{B}}(A_{k_0,k_1})) \geq k_0 k_1 + 1.$$

If $k_1 > P(k_0, 3)$, we can use only $P(k_0, 3)$ of the k_1 planes in this way. Using the same argument, we can construct $P(k_0, 3)$ defining sets, each containing k_0 points, and we can find an oblivious point. This gives the second term in the minimum. $\Box$

Notice that this bound on $\mathrm{VCdim}(T_{\mathbb{B}}(A_{k_0,k_1}))$ is $\Omega(W)$ if k_1 is sufficiently small.

The following lemma gives a lower bound for two-layer networks with real inputs and arbitrary connectivity.

Lemma 10 *Consider a two-layer feed-forward architecture $A = (U, C)$ with $k_1 > 0$ first layer units, $u_1^1, u_2^1, \ldots, u_{k_1}^1$, all connected to a single output unit. Suppose the unit u_i^1 is connected to $n_i > 0$ input units, for $i = 1, 2, \ldots, k_1$. For this architecture,*

$$\mathrm{VCdim}\left(T_{\mathbb{R}}(A)\right) \geq \sum_{i=1}^{k_1} n_i + 1.$$

Proof Suppose there are k_0 input units. Label the input and first-layer units as $U_0 = \{u_1^0, u_2^0, \ldots, u_{k_0}^0\}$ and $U_1 = \{u_1^1, u_2^1, \ldots, u_{k_1}^1\}$ respectively. For the first-layer unit u_i^1, consider the set $I(u_i^1)$ of input units connected to u_i^1,

$$I(u_i^1) = \{v \in U : (v, u) \in C\}.$$

We will consider separately those units that are connected only to a single input unit ($|I(u_i^1)| = 1$). Define $V_1 = \{u_i^1 : |I(u_i^1)| = 1\}$ and $V_{>1} = \{u_i^1 : |I(u_i^1)| > 1\}$.

Consider the first-layer units $u_i^1 \in V_{>1}$. Place their weight vectors on the unit hypersphere in $\mathbb{R}^{k_0}$ so that they are all distinct, and set their thresholds to -1. The decision boundary of each unit is a hyperplane in $\mathbb{R}^{k_0}$ that touches the unit hypersphere at some point. For each unit u_i^1, choose n_i points in general position[1] close to this point and on the unit's hyperplane. Choose the signs of the weights and thresholds of these units so that each unit classifies the origin as 0. To ensure that each set of n_i points forms a defining set for the unit u_i^1, we will need to choose the weights of the output unit appropriately.

Consider the remaining first-layer units, those in V_1. Let S_j be the set of first-layer units that have connections only from input unit u_j^0, $S_j = \{u_i^1 : I(u_i^1) = \{u_j^0\}\}$. For each input unit u_j^0 with $|S_j| > 0$, place $|S_j|$ distinct points on the axis

$$a_j = \left\{x = (x_1, x_2, \ldots, x_{k_0}) \in \mathbb{R}^{k_0} : x_i = 0, i \neq j\right\}.$$

Ensure that these points are classified as 1 by all but one of the units in $V_{>1}$ (this is always possible, since we can always choose the hyperplanes of the units in $V_{>1}$ so that their intersections do not intersect the axis a on which these points lie). For each of these points, use one of the $|S_j|$ units to define a hyperplane perpendicular to the axis a and passing through that point. Choose the signs of the weight and threshold so that the origin is classified as 0 by all of the units in S_j.

We now have two requirements for the placement of the output unit's hyperplane in the k_1-cube. It must separate the origin of the k_1-cube from all vertices with only bit i set, where $|I(u_i^1)| > 1$, and it must separate some pair of vectors of the form $(b_1, b_2, \ldots, b_{i-1}, 0, b_{i+1}, \ldots, b_{k_1})$ and $(b_1, b_2, \ldots, b_{i-1}, 1, b_{i+1}, \ldots, b_{k_1})$, where $|I(u_i^1)| = 1$. So we need a hyperplane to pass through k_1 mutually orthogonal edges of the k_1-cube. We can always choose k_1 linearly independent points, each on one of these edges, and pass our second layer plane through these points. Then we have a defining set of size n_i for each first-layer unit u_i^1. The origin is an oblivious point, so Corollary 7 implies $\mathrm{VCdim}\left(T_{\mathbb{R}}(A)\right) \geq \sum_{i=1}^{k_1} n_i + 1$. $\Box$

We can improve this bound for two-layer networks with direct connections from input units to the output unit.

Lemma 11 *Let $A = (U, C)$ be a feed-forward network architecture with a single output unit, u_i^L, and no connections from input units to u_i^L. Let $N = (U, C, w, \theta)$ be a threshold network with k_0 real-valued inputs. Suppose there is a set $\mathcal{S}$ of nonempty subsets of $\mathbb{R}^{k_0}$, and each set in $\mathcal{S}$ is a defining set for a non-output processing unit $u \in U - U_0 - \{u_1^L\}$. Suppose that there is an open region $R \subseteq \mathbb{R}^{k_0}$ that satisfies*

1. *For all $a \in \mathbb{R}$, if $a > 1$ then $x \in R \Rightarrow ax \in R$.*
2. *There is a $b \in \{0,1\}$ so that, for all sufficiently small perturbations of the weights of units in the network, each point x in R has $F_N(x) = b$.*

[1] A set S of points in $\mathbb{R}^n$ is in general position if no subset of S containing $k+1$ points lies on a $(k-1)$-dimensional hyperplane, for $k \in \{1, 2, \ldots, n\}$.

3. R has non-zero content (Lebesgue measure).

Now, consider an architecture $A' = (U, C')$, with $C' = C \cup D$, where $D \subseteq (U_0 \times \{u_1^L\})$ is a set of direct connections from the input units to the output unit.

There is a threshold network $N' = (U, C', w', \theta')$ and a set $S_0 \subseteq R$, so that each set in $\mathcal{S}$ remains a defining set for a unit $u \in U - U_0$, and S_0 is a defining set for u_1^L. Furthermore, $|S_0| = |D|$.

Proof Sketch Leave the weights and thresholds of N' the same as those in N, and choose the weights of the direct connections so that the network's output changes on a hyperplane in the region R. Provided these direct weights are sufficiently small (and Condition 1 allows us to choose them as small as desired), the classifications of the points in defining sets will not be changed. A set of $|D|$ points in general position on that hyperplane in R will constitute a defining set for u_1^L in N. $\square$

For the network described in the proof of Lemma 10, we can choose a set that satisfies the conditions of Lemma 11 (any unbounded cell with nonparallel bounding hyperplanes contains such a set), so we have

Theorem 12 *If $A = (U, C)$ is a two-layer feed-forward architecture with m connections from the input units to other units, then* $\mathrm{VCdim}(T_{\mathbb{R}}(A)) \geq m + 1$.

Again, the VC-dimension is $\Omega(W)$.

4 THREE LAYER NETWORKS

Define the k_0–k_1–k_2 architecture as the three-layer, completely connected architecture with k_0 input units, k_1 first-layer units, k_2 second-layer units, and a single output unit.

Theorem 13 *Let k_0, k_1, and k_2 be positive integers. If A_{k_0,k_1,k_2} is the k_0–k_1–k_2 architecture, we have*
(a) *If $k_0 \geq k_1$,*

$$\mathrm{VCdim}(T_{\mathbb{R}}(A_{k_0,k_1,k_2})) \geq k_0k_1 + k_1 \left[\min \left(k_2, \frac{2^{k_1}}{k_1^2/2 + k_1/2 + 1} \right) - 1 \right] + 1.$$

(b) *If $1 < k_0 < k_1 \geq k_2$,*

$$\mathrm{VCdim}(T_{\mathbb{R}}(A_{k_0,k_1,k_2})) \geq k_0k_1 + \frac{k_1(k_2-1)}{2} + 1.$$

(c) *If $1 < k_0 < k_1 < k_2$,*

$$\mathrm{VCdim}(T_{\mathbb{R}}(A_{k_0,k_1,k_2})) \geq k_0k_1 + k_0 \left[\min \left(k_2, \frac{\sum_{i=0}^{k_0} \binom{k_1}{i}}{k_1^2/2 + k_1/2 + 1} \right) - 1 \right] + 1.$$

The VC-dimension is $\Omega(W)$ in cases (a) and (b), provided k_2 is not too large. In particular, if we fix the number of second layer units in a three-layer completely connected architecture, and increase the number of input and/or first-layer units, then asymptotically the VC-dimension increases at least linearly with the number of weights. There can be no analogous result if we increase the number of second-layer units only, since there is a bounded number of boolean functions of k_1 variables.

Proof (a) If $k_0 = 1$, the bound is trivially true, so suppose $k_0 > 1$. Place the k_1 first-layer hyperplanes around the unit hypersphere and choose k_0 points near the hypersphere on each hyperplane, as in the proof of Lemma 10. These points will form defining sets for the first layer units. Choose the signs of the first layer weights so that the output of all first-layer units is 0 when the input is the origin of $\mathbb{R}^{k_0}$. Consider the outputs of the first layer units. We need to ensure that the origin of the k_1-cube is classified differently from the k_1 neighbouring vertices, so we choose the weights of a second layer unit so that it classifies the origin as 0 and all other vertices of the k_1-cube as 1.

Now, since $k_1 \leq k_0$, the hyperplanes divide $\mathbb{R}^{k_0}$ into 2^{k_1} distinct cells. Find a maximal 3-packing T of the k_1-cube that includes the origin. From Lemma 9, $|T| \geq 2^{k_1}/(k_1^2/2 + k_1/2 + 1)$. If $k_2 \leq |T|$, we can find defining sets for $k_2 - 1$ second layer units as follows. Each vertex of the k_1-cube corresponds to a cell in $\mathbb{R}^{k_0}$. For every t in T (except the origin), place a point in each of the k_1 cells adjacent to the cell correponding to t. Choose the weights of a second layer unit so that its hyperplane passes through these k_1 vertices adjacent to t, and the output of the unit in response to t is 0. Let the output unit implement the AND function. Clearly, we have k_1 defining sets of size k_0, and $k_2 - 1$ defining sets of size k_1, so Corollary 7 implies the VC-dimension is at least $k_0k_1 + k_1(k_2 - 1) + 1$.

If $k_2 > |T|$, we can find defining sets in this way for only $|T| - 1$ second layer units. This gives the second term in the minimum.

(b) Consider the intersection of $\mathbb{R}^{k_0}$ and a two-dimensional plane P through the origin. Let x_1 and x_2 be two perpendicular axes in this plane. Place the weight vectors of the k_1 first-layer units on the unit circle in P so that they are all distinct and all have a positive component in the x_2 direction. Set the thresholds of these units to 1, so that each unit classifies the origin as 0. These units each define a hyperplane in $\mathbb{R}^{k_0}$ that is perpendicular to P. For each of these hyperplanes, choose k_0 points in general position on the hyperplane near its intersection with the unit hypersphere. Each of these sets of k_0 points will form a defining set for a first-layer unit. We need to ensure that the origin of the k_1-cube is classified differently from the k_1 neighbouring vertices, so choose the weights of a second-layer unit so that it classifies the origin as 0, and all other vertices of the k_1-cube as 1.

Now, consider the open subsets of P that are bounded by segments of the first-layer hyperplanes. We refer to these subsets as 'cells'. Each cell corresponds to a vertex of the k_1-cube. Order the k_1 first-layer units so that the x_1 component of $w_{u_i^1}$ is less than that of $w_{u_j^1}$ if and only if $i < j$. If we define movement in the positive x_1 direction as left-to-right, this means that the lines are ordered from left to right by the location of their intersections with the unit circle, $w_{u_i^1}$. Define the vector of first-layer unit outputs using this ordering. Each of these vectors corresponds to a cell in P. Let $D_i(k_1) \subseteq \{0,1\}^{k_1}$ be the set of first-layer unit output vectors that contain exactly i 1's. That is, a vector $y = (y_1, y_2, \ldots, y_{k_1}) \in D_i(k_1)$ has $|\{j : y_j = 1, 1 \le j \le k_1\}| = i$. We say that $y = (y_1, \ldots, y_{k_1})$ has i contiguous 1's if there is an $n \in \{1, \ldots, k_1 - i + 1\}$ such that $y_j = 1$ if and only if $n \le j \le n + i - 1$.

Claim 1 *For any positive integer k_1 and any $0 \le i \le k_1$,*

$$D_i(k_1) = \left\{ v \in \{0,1\}^{k_1} : v \text{ has } i \text{ contiguous } 1\text{'s} \right\}$$

and so

$$|D_i(k_1)| = \begin{cases} 1 & i = 0 \\ k_1 - i + 1 & 1 \le i \le k_1 \\ 0 & k_1 < i. \end{cases} \tag{1}$$

The proof of the claim is by induction on k_1. For all k_1, $D_0(k_1)$ contains only the origin of the k_1-cube, and $D_i(k_1)$ is empty for $i > k_1$. If $k_1 = 1$, $D_0(k_1) = \{0\}$, $D_1(k_1) = \{1\}$, and $D_i(k_1)$ is empty for $i > 1$. Suppose the claim is true for $k_1 = 1, 2, \ldots, m$. Add another first-layer unit to the network with threshold 1 and weight vector w that lies on the unit circle in P, to the left of $w_{u_1^1}, w_{u_2^1}, \ldots w_{u_m^1}$. The addition of this line will add $m+1$ cells to the arrangement, because the new line intersects all of the old lines. Label these cells by the vector of first-layer unit outputs. If we move to the right along the line from w, we pass through one cell from $D_0(m)$, one from $D_1(m)$, ..., and one from $D_m(m)$. A new cell is created to the left of the line in each of these cells, so $|D_i(m+1)| = |D_i(m)|+1$ for $i = 1, 2, \ldots, m+1$. Equation (1) is therefore true for $k_1 = m+1$, and hence for any k_1.

Notice that the first line we cross in moving to the right from w is the line corresponding to unit u_m^1. So all cells through which we passed, except for the cell containing w, are labelled with a vector with bit m set to 1. That is, all new cells created by the addition of the line are labelled with vectors consisting of a contiguous string of 1's, so

$$D_i(m+1) \subseteq \left\{ v \in \{0,1\}^{m+1} : v \text{ has } i \text{ contiguous 1's} \right\}.$$

This and Equation (1) prove the claim.

We construct a defining set for a second-layer unit by choosing a point in each cell corresponding to an element of $D_i(k_1)$ (for $i = 2, 3, \ldots, k_2$). Each second-layer hyperplane (except one) is made to pass through all elements of some set $D_i(k_1)$, and the output unit's weights are chosen so that the cells in the k_1-cube bounded by these hyperplanes are classified distinctly. By the claim above, the elements of $D_i(k_1)$ (where $i = 1, 2, \ldots, k_1$) can be written as the rows of an upper triangular matrix with diagonal elements all nonzero, so the vectors in $D_i(k_1)$ are linearly independent. This means we have $k_2 - 1$ defining sets for second-layer units, containing $|D_2(k_1)|, |D_3(k_1)|, \ldots, |D_{k_2}(k_1)|$ points respectively. Now,

$$\begin{aligned} \sum_{i=2}^{k_2} |D_i(k_1)| &= \sum_{i=2}^{k_2} (k_1 - i + 1) \\ &= (k_1 - k_2/2)(k_2 - 1) \\ &\ge k_1(k_2 - 1)/2, \end{aligned}$$

so Corollary 7 gives

$$\mathrm{VCdim}\left(T_{\mathbb{R}}(A_{k_0,k_1,k_2})\right) \ge k_0 k_1 + \frac{k_1(k_2-1)}{2} + 1.$$

To see that we can choose appropriate second- and third-layer hyperplanes, recall that each set $D_i(k_1)$ contains vertices with exactly i 1's. The weights of each second-layer unit u_i^2 can be chosen so that its hyperplane intersects the elements of $D_i(k_1)$ (for $i = 2, \ldots, k_2$). No second-layer hyperplanes intersect in $[0,1]^{k_1}$, and this arrangement of hyperplanes divides the k_1-cube into k_2+1 cells. To ensure that each cell is classified distinctly, the output unit's hyperplane must cut the k_2-cube on each of k_2 mutually orthogonal edges, and such a hyperplane can always be found.

(c) Place the k_1 first-layer hyperplanes around the unit hypersphere and choose a second-layer unit's weights as in the proof of (a). The first layer hyperplanes divide $\mathbb{R}^{k_1}$ into $N = \sum_{i=0}^{k_0} \binom{k_1}{i}$ cells [7]. Consider the set S of corresponding vertices of the k_1-cube. Since $0 \in S$ we can find a maximal 3-packing $T \subseteq S$ that contains 0. Since there are no more than $k_1^2/2 + k_1/2 + 1$ vertices within Hamming distance 3 of any given vertex, $|T| \ge \sum_{i=0}^{k_0} \binom{k_1}{i} / (k_1^2/2 + k_1/2 + 1)$. Now, for each element of T except 0, place a point in every neighbouring cell. Since $k_1 > k_0$, there must be at least k_0 neighbouring cells. We can ensure that these are defining sets for second-layer units, so if $k_2 \le |T|$, the VC-dimension is at least $k_0 k_1 + k_0(k_2 - 1) + 1$. If $k_2 > |T|$, we can only find $|T| - 1$ defining sets in this way, In this case, the VC-dimension is at least

$$k_0 k_1 + k_0 \left(\frac{\sum_{i=0}^{k_0} \binom{k_1}{i}}{k_1^2/2 + k_1/2 + 1} - 1 \right) + 1.$$

□

For the case of completely connected three-layer threshold networks with binary inputs and few first-layer units, we can use the bound for a two-layer network (Theorem 8) to show that the VC-dimension is $\Omega(W)$.

Proposition 14 *Let k_0, k_1, k_2, and k_3 be positive integers. If A_{k_0,k_1,k_2} is the k_0-k_1-k_2 architecture, with $k_0 > k_1$ and $k_2 < 2^{k_1}/(k_1^2/2 + k_1/2 + 1)$, then*

$$\mathrm{VCdim}\left(T_{\mathbb{B}}(A_{k_0,k_1,k_2})\right) \ge k_1 max(k_0, k_2) + 1.$$

5 CONCLUSIONS

We have shown that the Vapnik-Chervonenkis dimension of the class of functions that can be computed by arbitrary two-layer and some completely connected three-layer networks with real inputs is at least proportional to the number of weights in the network. This result also applies to completely connected two-layer networks with binary inputs, and to completely connected three-layer networks with binary inputs and few first-layer units. These results, together with the VC-dimension upper bounds in [3], show that the sample size necessary and sufficient for pac learning in these networks is $\Omega(W)$ and $O(W \log N)$, where W is the number of weights in the network and N is the number of processing units.

Notice that these lower bounds apply to feed-forward networks of processing units with sigmoid transfer functions, since a sigmoid network can compute any function on a finite set that can be computed by a threshold network with the same architecture.

These upper and lower bounds are separated by a factor of $\log N$, where N is the number of processing units in the network. Recently, Maass has shown that four-layer feed-forward threshold network architectures can be constructed with VC-dimension $\Omega(W \log N)$ [8]; it is not known if the $\log N$ factor is necessary for particular classes of multi-layer architectures (for example, completely connected ones), or for networks with fewer than four layers.

It seems likely that the VC-dimension bounds for two-layer networks with binary inputs can be extended to architectures with limited connectivity.

Acknowledgements

This research was supported by the Australian Telecommunications and Electronics Research Board. Thanks to Steven Young for many helpful discussions, and to Eric Baum for valuable comments.

References

[1] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, 27, no. 11, pp. 1134–1143, November 1984.

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

[3] E. B. Baum and D. Haussler, "What size net gives valid generalization," *Neural Computation*, 1, pp. 151–160, 1989.

[4] E. B. Baum, "On the capabilities of multilayer perceptrons," *Journal of Complexity*, 4, pp. 193–215, 1988.

[5] P. L. Bartlett, "Vapnik-Chervonenkis dimension bounds for two- and three-layer networks," *Neural Computation*, 5, no. 3, pp. 353–355, 1993.

[6] A. P. Street, *Combinatorics: A First Course.* St Pierre, Canada, Charles Babbage Research Centre, 1982.

[7] T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, EC-14, pp. 326–334, 1965.

[8] W. Maass, "Bounds for the computational power and learning complexity of analog neural nets," Technische Universitaet Graz, (extended abstract), 1992.