

# A Result of Vapnik with Applications

Martin Anthony

Department of Statistical and Mathematical Sciences

London School of Economics

Houghton Street

London WC2A 2AE, U.K.

John Shawe-Taylor

Department of Computer Science

Royal Holloway and Bedford New College

Egham Surrey TW20 0EX, U.K.

Revised version, 23 November 1991

## **Abstract**

A new proof of a result due to Vapnik is given. Its implications for the theory of PAC learnability are discussed, with particular reference to the learnability of functions taking values in a countable set. An application to the theory of artificial neural networks is then given.

# 1 Introduction

In this paper we give a simple combinatorial proof of a bound, due to Vapnik, on the probability of relative deviation of frequencies from probabilities for a class of events. This result has applications in the theory of PAC learning, introduced by Valiant [13] and developed by many researchers [4, 5]. We discuss these applications, and describe how the theory of learnability may be extended from learning sets to learning functions, following work of Haussler [6] and Natarajan [8]. We consider functions with finite or countably infinite range, generalising the definition of the VC dimension [15]. We do not deal here with functions which take values in Euclidean space, as we feel these are best analysed using the elegant theory described in [7] of functions taking values in arbitrary metric spaces. In the final section, we apply the results to a problem in artificial neural networks. Haussler [7], Baum and Haussler [2], and Natarajan [8] have obtained upper bounds on a sample size guaranteeing valid generalisation in networks of certain types. We obtain a bound on the (generalised) VC dimension of a feedforward linear threshold net with multiple outputs, extending the result of Baum and Haussler for such networks with only one output, and obtaining a bound which is a linear factor better than the result obtained for this case by applying a more general result of Natarajan.

## 2 A result of Vapnik

In this section, we prove a special case of a result due to Vapnik [14] concerning the uniform relative deviation over a class of events of relative frequencies from probabilities.

Suppose that  $(S, \Sigma, \nu)$  is a probability space and that  $\mathcal{C} \subseteq \Sigma$ . We assume that if  $S$  is countable, then  $\Sigma$  consists of all subsets of  $X$ . In general, certain measure-theoretic restrictions must be placed on both  $\Sigma$  and  $\mathcal{C}$ . We refer here to [9], [7], [14]. For  $\mathbf{s} = (s_1, \dots, s_n) \in S^n$ , we let  $I(\mathbf{s}) = \{s_i : 1 \leq i \leq n\}$ . The integer  $\Delta_{\mathcal{C}}(\mathbf{s})$  is defined to be the number of distinct sets of the form  $A \cap I(\mathbf{s})$  where  $A$  runs through  $\mathcal{C}$ , and  $\Delta_{\mathcal{C}}(n)$  is defined to be the maximum over all  $\mathbf{s} \in S^n$  of  $\Delta_{\mathcal{C}}(\mathbf{s})$ . We say that a subset  $\mathcal{A} = \{A_1, \dots, A_t\}$  of  $\mathcal{C}$  is a complete set of distinct representatives (CSDR) of  $\mathcal{C}$  for  $\mathbf{s}$  if  $t = \Delta_{\mathcal{C}}(\mathbf{s})$ , if  $1 \leq i \neq j \leq t$  implies  $A_i \cap I(\mathbf{s}) \neq A_j \cap I(\mathbf{s})$ , and if any set of the form  $A \cap I(\mathbf{s})$  where  $A \in \mathcal{C}$  is equal to  $A_i \cap I(\mathbf{s})$  for some  $i$  between 1 and  $t$ . The *relative frequency* of  $A \in \mathcal{C}$  on  $\mathbf{s} = (s_1, s_2, \dots, s_n)$  is defined to be

$$\mathbf{P}_{\mathbf{s}}(A) = \frac{1}{n} |\{i : s_i \in A\}|.$$

We prove the following.

**Theorem 2.1** With the above definitions, for  $\eta > 0$ ,

$$\nu^m \left\{ \mathbf{x} \in S^m : \exists A \in \mathcal{C} \text{ s.t. } \frac{\nu(A) - \mathbf{P}_{\mathbf{x}}(A)}{\sqrt{\nu(A)}} > \eta \right\} \leq 4 \Delta_{\mathcal{C}}(2m) \exp\left(-\frac{1}{4}\eta^2 m\right).$$

**Proof** Let

$$Q = \left\{ \mathbf{x} \in S^m : \exists A \in \mathcal{C} \text{ s.t. } \frac{\nu(A) - \mathbf{P}_{\mathbf{x}}(A)}{\sqrt{\nu(A)}} > \eta \right\},$$

$$R = \left\{ \mathbf{xy} \in S^{2m} : \exists A \in \mathcal{C} \text{ s.t. } \mathbf{P}_{\mathbf{y}}(A) - \mathbf{P}_{\mathbf{x}}(A) > \eta \sqrt{\mathbf{P}_{\mathbf{xy}}(A)} \right\}.$$

Then we claim that  $\nu^m(Q) \leq 4\nu^{2m}(R)$  for  $m > 2/\eta^2$ . Suppose  $\mathbf{x} \in Q$ , so that there is  $C \in \mathcal{C}$  with  $\nu(C) - \mathbf{P}_{\mathbf{x}}(C) > \eta\sqrt{\nu(C)}$ . Since  $\mathbf{P}_{\mathbf{x}}(C) \geq 0$ , this implies  $\nu(C) > \eta^2$ . Now suppose  $m > 2/\eta^2$  and  $\mathbf{y} \in S^m$  is such that  $\mathbf{P}_{\mathbf{y}}(C) > \nu(C)$ . If (noting that  $\mathbf{P}_{\mathbf{y}}(C) > 0$  implies the denominator is positive)

$$F = \frac{\mathbf{P}_{\mathbf{y}}(C) - \mathbf{P}_{\mathbf{x}}(C)}{\sqrt{\mathbf{P}_{\mathbf{xy}}(C)}},$$

then some simple calculus shows that  $F > \eta$ . It is known that, since  $m > 2/\eta^2 \geq 2/\nu(C)$ ,  $\mathbf{P}_{\mathbf{y}}(C) > \nu(C)$  with probability at least  $1/4$ , and we therefore have  $\nu^{2m}(R) \geq \frac{1}{4}\nu^m(Q)$ . The claim follows.

Fix  $\mathbf{z} = (z_1, \dots, z_{2m}) \in S^{2m}$ , let  $\mathcal{A}(\mathbf{z}) = \{A_1, \dots, A_t\}$  be a CSDR of  $\mathcal{C}$  for  $\mathbf{z}$  and let

$$R^i = \left\{ \mathbf{xy} \in S^{2m} : \mathbf{P}_{\mathbf{y}}(A_i) - \mathbf{P}_{\mathbf{x}}(A_i) > \eta \sqrt{\mathbf{P}_{\mathbf{xy}}(A_i)} \right\}.$$

Let  $\Lambda$  be the ‘‘swapping’’ subgroup of the symmetric group of degree  $2m$ , used in this context by Pollard [9]. This is the group generated by the transpositions  $(i, m+i)$  for  $1 \leq i \leq m$ .  $\Lambda$  has a natural group action on  $S^{2m}$ ; given  $\mathbf{s} = (s_1, \dots, s_{2m})$  and  $\tau \in \Lambda$ , we define

$$\tau\mathbf{s} = (s_{\tau(1)}, \dots, s_{\tau(2m)}).$$

Denote by  $\Theta^i(\mathbf{z})$  the number of permutations  $\tau$  in  $\Lambda$  for which  $\tau\mathbf{z}$  belongs to  $R^i$ , and define  $\Theta^i(2m)$  to be the maximum over all  $\mathbf{z} \in S^{2m}$  of this parameter. Because the action of  $\Lambda$  is measure preserving (with respect to the product measure  $\nu^{2m}$ ), it can be shown easily (see, for example, [4, 15]) that

$$\nu^{2m}(R^i) \leq \frac{\Theta^i(2m)}{|\Lambda|}.$$

We bound this latter quantity. Following [7], for each  $1 \leq j \leq 2m$ , we let  $X_j = 1$  if  $z_j \in A_i$  and  $X_j = 0$  otherwise and, for  $1 \leq j \leq m$ , we let  $Y_j$  be the random variable which equals

$X_j - X_{m+j}$  with probability  $1/2$  and  $X_{m+j} - X_j$  with probability  $1/2$ . Let  $P$  be the uniform distribution on  $\Lambda$ . Then,

$$\begin{aligned} \frac{\Theta^i(\mathbf{z})}{|\Lambda|} &= P \left\{ \tau \in \Lambda : \sum_{j=1}^m (X_{\tau^{-1}(m+j)} - X_{\tau^{-1}(j)}) > \eta \left( \frac{m}{2} \sum_{j=1}^{2m} X_j \right)^{1/2} \right\} \\ &= \text{Prob} \left\{ \sum_{j=1}^m Y_j > \eta \left( \frac{m}{2} \sum_{j=1}^{2m} X_j \right)^{1/2} \right\}. \end{aligned}$$

By Hoeffding's inequality [9], this probability is bounded by

$$\exp \left( - \frac{\eta^2 m \sum_{j=1}^{2m} X_j}{4 \sum_{j=1}^m (X_j - X_{m+j})^2} \right) \leq \exp \left( - \frac{1}{4} \eta^2 m \right).$$

This holds for each  $1 \leq i \leq t$ .  $R$  is the union of the sets  $R^i$  ( $1 \leq i \leq t$ ) and  $t \leq \Delta_{\mathcal{C}}(2m)$ . Therefore

$$\nu^{2m}(R) \leq \Delta_{\mathcal{C}}(2m) \exp \left( - \frac{1}{4} \eta^2 m \right).$$

Since  $\nu^m(Q) \leq 4 \nu^{2m}(R)$ , this proves the theorem for  $m > 2/\eta^2$ . The bound of the theorem holds trivially for values of  $m$  less than this.  $\square$

This bound is a slight improvement on that of Vapnik, replacing the constant 8 in Vapnik's result by 4.

### 3 VC dimension and learnability

Vapnik's result provides non-trivial bounds as  $m$  tends to infinity only if the *growth function*  $\Delta_{\mathcal{C}}(m)$  grows subexponentially with  $m$ . This can be guaranteed in many cases by an elegant theory due to Vapnik and Chervonenkis [15]. We say that the collection  $\mathcal{C}$  of sets has finite *VC dimension*  $d$  if  $\Delta_{\mathcal{C}}(d) = 2^d$  but  $\Delta_{\mathcal{C}}(d+1) < 2^{d+1}$  and that it has infinite VC dimension if it does not have finite VC dimension. When  $\mathcal{C}$  has finite VC dimension  $d$ , a result of Sauer [11] shows that for all  $m > d$ ,

$$\Delta_{\mathcal{C}}(m) < \left( \frac{em}{d} \right)^d.$$

Thus, if  $\mathcal{C}$  has finite VC dimension, the right hand side of the expression in Theorem 2.1 tends to zero as  $m$  tends to infinity, and it does so at a rate which can be bounded independently of  $\nu$ .

The problem of deciding what sample size is necessary for valid generalisation in computational models of learning has recently received much attention, particularly in the context of Valiant's PAC learning. In this framework, we are given a set of inputs and a hypothesis space of functions from the inputs to  $\{0, 1\}$ . There is assumed to be a (usually fixed but unknown) probability distribution on the inputs, and the aim is to find a good approximation to a particular target concept from the hypothesis space, given only a random sample of training examples and the value of the target concept on those examples. It is often stipulated that this be carried out using polynomially bounded time or space resources, but we do not address this aspect here.

Formally, the input space is a probability space  $(X, \Sigma, \mu)$  and the hypothesis space  $H$  is a set of measurable functions from  $X$  to  $\{0, 1\}$ . The target concept  $c$  is assumed to be one of the functions from  $H$ . In the simplest form of the standard framework, it is shown that if  $H$  has finite VC dimension, then there is a sample size such that any hypothesis from  $H$  consistent with the target concept on that many examples is likely to be a good approximation to the target [5, 7, 14, 12, 1]. However, in any real learning situation, where there is a *learning algorithm* for producing the hypothesis supposed to approximate the target, it is unrealistic to assume that the hypothesis produced is consistent with the target on all of the training sample. It is more reasonable to assume only that the hypothesis is consistent with the target on a large proportion of the training sample. To account for this, to allow the possibility of classification errors during training, and to allow for ill-defined or *stochastic* concepts, the theory has been extended [4] to discuss not the learnability of functions from  $X$  to  $\{0, 1\}$  with an underlying distribution  $\mu$ , but instead the learnability of probability distributions on the set  $S = X \times \{0, 1\}$ . We remark that any function  $c$  from  $X$  to  $\{0, 1\}$  together with an underlying distribution  $\mu$  can be realised as a probability measure  $\nu$  on  $S$ , as we show later. We make the following definitions.

Suppose that  $\nu$  is some probability measure on  $S = X \times \{0, 1\}$ . We define the *actual error* (with respect to  $\nu$ ) of  $h \in H$  to be

$$\text{er}_\nu(h) = \nu\{(x, a) : a \neq h(x)\}.$$

A *sample* of length  $m$  of  $\nu$  is a sequence  $\mathbf{x}$  of  $m$  points of  $S$ , randomly drawn according to the distribution  $\nu$ . For  $h \in H$ , the *observed error* of  $h$  on sample  $\mathbf{x} = ((x_1, a_1), \dots, (x_m, a_m))$  is

$$\text{er}_\mathbf{x}(h) = \frac{1}{m} |\{i : h(x_i) \neq a_i\}|.$$

The VC dimension of a set of  $\{0, 1\}$ -valued functions is defined in the obvious way: if  $H$  is such a set of functions, the VC dimension of  $H$  is defined to be the VC dimension of the collection  $\mathcal{S}$  of supports of the functions in  $H$ . In this case, given  $\mathbf{x} = (x_1, \dots, x_m) \in X^m$ ,  $\Delta_{\mathcal{S}}(\mathbf{x})$  equals the number of distinct vectors of the form  $(h(x_1), \dots, h(x_m))$ , as  $h$  runs through  $H$ .

Using Theorem 2.1, we can obtain the following learnability result, from [4].

**Theorem 3.1** Let  $H$  be a hypothesis space of  $\{0, 1\}$ -valued functions defined on an input space  $X$ . Let  $\nu$  be any probability measure on  $S = X \times \{0, 1\}$ , let  $0 < \epsilon < 1$  and let  $0 < \gamma \leq 1$ . Then the probability (with respect to the product measure  $\nu^m$ ) that, for  $\mathbf{x} \in S^m$ , there is some hypothesis from  $H$  such that

$$\text{er}_\nu(h) > \epsilon \quad \text{and} \quad \text{er}_{\mathbf{x}}(h) \leq (1 - \gamma)\text{er}_\nu(h)$$

is at most

$$4 \Delta_H(2m) \exp\left(-\frac{1}{4}\gamma^2\epsilon m\right).$$

**Proof** For  $h \in H$ , define the error set  $E_h$  of  $h$  to be  $E_h = \{(x, a) \in S : h(x) \neq a\}$  and take  $\mathcal{C}$  to be the collection  $\mathcal{C} = \{E_h : h \in H\}$  of error sets. The result follows easily from Theorem 2.1, observing that  $\mathbf{P}_{\mathbf{x}}(E_h) = \text{er}_{\mathbf{x}}(h)$ ,  $\nu(E_h) = \text{er}_\nu(h)$  and  $\text{VCdim}(\mathcal{C}) = \text{VCdim}(H)$ .  $\square$

It is often required that the probability described in the statement of the above theorem be less than some prescribed value  $\delta$  so that, taking a large enough sample, one can guarantee that with high probability, any hypothesis with small observed error has small actual error. This condition is made precise in the following result, which provides such a sample size and improves a bound of [4].

**Proposition 3.2** Let  $0 < \epsilon, \delta < 1$  and  $0 < \gamma \leq 1$  and let  $\nu$  be any distribution on  $S = X \times \{0, 1\}$ . If  $H$  has finite VC dimension  $d$ , then there is  $m_0 = m_0(\epsilon, \delta, \gamma)$  such that if  $m > m_0$  then, for  $\mathbf{x} \in S^m$ , with probability at least  $1 - \delta$  (with respect to the product measure  $\nu^m$ ),

$$\text{er}_{\mathbf{x}}(h) \leq (1 - \gamma)\epsilon \implies \text{er}_\nu(h) \leq \epsilon.$$

A suitable value of  $m_0$  is

$$m_0 = \frac{1}{\gamma^2\epsilon(1 - \sqrt{\epsilon})} \left( 4 \log\left(\frac{4}{\delta}\right) + 6d \log\left(\frac{4}{\gamma^{2/3}\epsilon}\right) \right),$$

where  $\log$  denotes natural logarithm.

**Proof** The proof uses Sauer's inequality:  $H$  has finite VC dimension  $d$  and therefore, for  $2m \geq d$ ,

$$\Delta_H(2m) < \left(\frac{2em}{d}\right)^d.$$

We show that if  $m > m_0$ , then

$$4 \left(\frac{2em}{d}\right)^d \exp\left(-\frac{1}{4}\gamma^2\epsilon m\right) \leq \delta,$$

from which the result will follow on using Theorem 3.1.

Now,

$$\begin{aligned} & 4 \left( \frac{2em}{d} \right)^d \exp \left( -\frac{1}{4} \gamma^2 \epsilon m \right) \leq \delta \\ \iff & \log 4 + d \log 2 + d + d \log m - d \log d - \frac{1}{4} \gamma^2 \epsilon m \leq \log \delta \\ \iff & \frac{1}{4} \gamma^2 \epsilon m \geq \log \left( \frac{4}{\delta} \right) + d \log 2 - d \log d + d + d \log m. \end{aligned}$$

Now we use the fact that for any  $\alpha > 0$  and for any  $m$ ,  $\log m \leq (-\log \alpha - 1) + \alpha m$ . (This is easily proved by elementary calculus.) Choosing  $\alpha = \epsilon \sqrt{\epsilon} \gamma^2 / 4d$  and substituting into the above, we see that the desired inequality holds when  $m > m_0$ .  $\square$

Instead of considering just  $\{0, 1\}$ -valued functions, we should like to consider functions taking values in some finite or countably infinite set. The same sorts of upper bounds on sufficient sample size in terms of a parameter (which we continue to call the VC dimension) that quantifies in some sense the “expressive power” of the space of functions can be obtained. For consistency, we want the notion of VC dimension for a space of functions to reduce to the straightforward definition of VC dimension when the range space has only two elements. Various definitions have been proposed.

We adopt a definition of Haussler [6], defining the VC dimension of a space of functions from a set  $X$  to a countable set  $Y$  to be the VC dimension of the collection of *graphs* of the functions. For any  $h \in H$ , the graph  $\mathcal{G}(h)$  of  $h$  is

$$\mathcal{G}(h) = \{(x, h(x)) : x \in X\},$$

and the *graph space* of  $H$  is  $\mathcal{G}(H) = \{\mathcal{G}(h) : h \in H\}$ . Then the VC dimension of  $H$  is defined to be the VC dimension of the space  $\mathcal{G}(H)$ .

We can describe this in another way. For  $\mathbf{y} = (y_1, \dots, y_m) \in Y^m$ , let  $I_{\mathbf{y}} : Y^m \rightarrow \{0, 1\}^m$  be defined by

$$I_{\mathbf{y}}((z_1, \dots, z_m)) = (a_1, \dots, a_m), \quad \text{where } a_i = 1 \iff y_i = z_i.$$

For  $\mathbf{x} = (x_1, \dots, x_m) \in X^m$  and  $h \in H$ , define  $\mathbf{x}^*(h) = (h(x_1), \dots, h(x_m))$ . This defines a mapping  $\mathbf{x}^*$  from  $H$  to  $Y^m$ . For each  $\mathbf{y} \in Y^m$ , the composition  $I_{\mathbf{y}} \circ \mathbf{x}^*$  is a mapping from  $H$  to the finite set  $\{0, 1\}^m$ . We define  $\Pi_H(\mathbf{x})$  to be the maximum, as  $\mathbf{y}$  ranges over  $Y^m$ , of  $|I_{\mathbf{y}} \circ \mathbf{x}^*(H)|$ , the cardinality of the image of  $H$  under  $I_{\mathbf{y}} \circ \mathbf{x}^*$ . Further, we let  $\Pi_H(m)$  be the maximum of  $\Pi_H(\mathbf{x})$  over all  $\mathbf{x} \in X^m$ . Then  $\Pi_H(m) = \Delta_{\mathcal{G}(H)}(m)$ , and therefore the VC dimension of  $H$  (is either infinite, or) is the largest integer  $d$  such that  $\Pi_H(d) = 2^d$ . Notice that for finite  $Y$ ,

$$\Pi_H(\mathbf{x}) \leq |\mathbf{x}^*(H)| \leq \Delta_H(m),$$

where  $\Delta_H(m)$  is the maximum over all  $\mathbf{x} \in X^m$  of  $|\mathbf{x}^*(H)|$ .

It is easy to see that if  $Y = \{0, 1\}$ , this notion of VC dimension coincides with the standard one. With this extended definition of VC dimension, we can apply the previous learnability results, Theorem 3.1 and Proposition 3.2. We note that, as earlier, we consider probability distributions on the set  $X \times Y$  rather than functions from  $X$  to  $Y$  with underlying probability distributions on  $X$ . However, every pair  $(c, \mu)$  where  $c \in H$  and  $\mu$  is a probability measure on  $X$  can be realised by a probability measure  $\nu = \nu(c, \mu)$  on the product  $\sigma$ -algebra  $\Sigma \times 2^Y$ . To see this, note that if  $Y = \{y_n\}_{n=1}^\infty$  is an enumeration of  $Y$ , then the product  $\sigma$ -algebra  $\Sigma \times 2^Y$  consists precisely of the sets of the form

$$\bigcup_{n=1}^{\infty} A_n \times \{y_n\},$$

where each  $A_n$  belongs to  $\Sigma$ . We then define

$$\nu \left( \bigcup_{n=1}^{\infty} A_n \times \{y_n\} \right) = \sum_{n=1}^{\infty} \mu \left( c^{-1}(y_n) \cap A_n \right).$$

It is easily verified that  $\nu$  is a probability measure such that for any  $A \in \Sigma$ ,

$$\nu \{(x, c(x)) : x \in A\} = \mu(A) \quad \text{and} \quad \nu \{(x, y) : x \in A, y \neq c(x)\} = 0.$$

**Theorem 3.3** Let  $0 < \epsilon < 1$  and  $0 < \gamma \leq 1$ . Suppose  $H$  is a hypothesis space of functions from an input space  $X$  to a countable set  $Y$ , and let  $\nu$  be any probability measure on  $S = X \times Y$ . Then the probability (with respect to  $\nu^m$ ) that, for  $\mathbf{x} \in S^m$ , there is some  $h \in H$  such that

$$\text{er}_\nu(h) > \epsilon \quad \text{and} \quad \text{er}_{\mathbf{x}}(h) \leq (1 - \gamma)\text{er}_\nu(h)$$

is at most

$$4 \Pi_H(2m) \exp \left( -\frac{\gamma^2 \epsilon m}{4} \right).$$

Furthermore, if  $H$  has finite VC dimension  $d$ , this quantity is less than  $\delta$  for

$$m > m_0(\epsilon, \delta, \gamma) = \frac{1}{\gamma^2 \epsilon (1 - \sqrt{\epsilon})} \left( 4 \log \left( \frac{4}{\delta} \right) + 6d \log \left( \frac{4}{\gamma^{2/3} \epsilon} \right) \right).$$

**Proof** The proof of this is similar to the proof of the analogous results, Theorem 3.1 and Proposition 3.2, for hypothesis spaces of  $\{0, 1\}$ -valued functions. Define the error set  $E_h$  of  $h \in H$  by

$$E_h = \{(x, y) \in X \times Y : h(x) \neq y\}.$$

Observe that  $E_h = (X \times Y) \setminus \mathcal{G}(h)$ . Let  $\mathcal{C} = \{E_h : h \in H\}$  be the collection of error sets. Suppose that  $h, g \in H$  and let

$$\mathbf{s} = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)) \in (X \times Y)^m.$$

Defining

$$I(\mathbf{s}) = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\},$$

and using the fact that  $E_h$  is the complement of  $\mathcal{G}(h)$ , we have

$$E_h \cap I(\mathbf{s}) = E_g \cap I(\mathbf{s}) \iff I(\mathbf{s}) \setminus \mathcal{G}(h) = I(\mathbf{s}) \setminus \mathcal{G}(g) \iff \mathcal{G}(h) \cap I(\mathbf{s}) = \mathcal{G}(g) \cap I(\mathbf{s}).$$

But the number of distinct sets of the form  $\mathcal{G}(h) \cap I(\mathbf{s})$  obtained as  $h$  ranges through all of  $H$  is, by definition,  $\Delta_{\mathcal{G}(H)}(\mathbf{s})$ . It follows that

$$\Pi_H(m) = \Delta_{\mathcal{G}(H)}(m) = \Delta_{\mathcal{C}}(m).$$

The first part of the Theorem now follows on using Theorem 2.1. The second part follows as in the proof of Proposition 3.2.  $\square$

## 4 An application to artificial neural networks

Artificial neural networks [10, 3] have recently received much attention; in particular, many researchers are involved in studying the problem of training a network to compute particular functions and to generalise from examples. In this section, we describe a family of such networks, and apply the preceding theory, extending results of [2].

A feedforward neural network is an ordered pair  $\mathcal{N} = (G, \mathcal{F})$ , where  $G = (V, E)$  is a directed acyclic graph, and  $\mathcal{F}$  is a finite set of *activation functions*.  $V$  is the disjoint union of a set  $I$  of *input nodes* and a set  $C$  of *computation nodes*, and  $O \subseteq C$  is a set of *output nodes*. Further, there is a *bias node*  $n_0 \in I$ . The number of input nodes will be denoted  $s + 1$  and the number of output nodes  $t$ . The underlying graph  $G$  is such that all computation nodes are connected to the bias node, and the input nodes have zero in-degree; that is,  $E \subseteq (C \cup I) \times C$  and  $\{n_0\} \times C \subseteq E$ . The computation nodes are labelled with the integers 1 to  $n = |C|$  in such a way that if  $(i, j) \in E$  then  $j > i$ . This can be accomplished since  $G$  is acyclic. We denote by  $d(j)$  the in-degree of computation node  $j$ .

Associated with computation node  $j$  is the set of states  $\Omega_j = \mathcal{R}^{d(j)}$ . We let  $\Omega^{(k)}$  denote the product  $\Omega^{(k)} = \Omega_1 \times \dots \times \Omega_k$ , and denote  $\Omega^{(n)}$  simply by  $\Omega$  (this is the set of all states of the network). Any  $\omega \in \Omega$  can be decomposed as  $\omega = \omega_1 \omega_2 \dots \omega_n$ . Given such a decomposition, we denote by  $w^k$  the vector  $\omega_1 \omega_2 \dots \omega_k$ .

Each computation node  $j$  has associated with it an *activation function*

$$f^j : \Omega_j \times \mathcal{R}^{d(j)} \rightarrow \{0, 1\},$$

and  $\mathcal{F}$  is the set of  $n$  activation functions. For  $w \in \Omega_j$ , the function  $h_w^j$  from  $\mathcal{R}^{d(j)}$  to  $\{0, 1\}$  is given by  $h_w^j(x) = f^j(w, x)$ .  $H^j$  denotes the set of functions  $h_w^j$  where  $w$  runs through  $\Omega_j$ , and we denote  $\Delta_{H^j}(m)$  by  $\Delta_j(m)$ .

An input  $x \in \mathcal{R}^s$  to the network consists of an assignment of a real number to each non-bias input node. Further, each node has an output value of 0 or 1. The output of a node is defined recursively in terms of the outputs of the previous nodes. The output of a non-bias input node is defined to be the input on that node, and the output of  $n_0$  is always 1. The *input vector* to computation node  $j$  depends on the input  $x$  and on  $\omega^{j-1}$ , and we write it as  $\mathbf{I}_j(\omega^{j-1}, x) \in \mathcal{R}^{d(j)}$ . The output of node  $j$  is then computed as

$$f^j(\omega_j, \mathbf{I}_j(\omega^{j-1}, x)).$$

The function computed by the network when in state  $\omega \in \Omega$  is the function  $F_\omega$  from  $X = \mathcal{R}^s$  to  $\{0, 1\}^t$  whose value is the  $(0, 1)$ -vector of outputs of the output nodes. The set of all  $F_\omega$  as  $\omega$  ranges through  $\Omega$  is denoted  $F$ , and we call  $F$  the set of functions computable by  $\mathcal{N}$ .

The *output function* of the network, which describes precisely the output of each computation node, is the function

$$\sigma : \Omega \times X \rightarrow \{0, 1\}^n.$$

Entry  $i$  of  $\sigma(\omega, x)$  is 1 if and only if, when the net is in state  $\omega$  and receives input  $x$ , node  $i$  has output 1. For a sequence  $\mathbf{x} = (x_1, \dots, x_m)$  of inputs, we define  $S(\mathbf{x})$  to be the number of distinct vectors of the form

$$(\sigma(\omega, x_1), \dots, \sigma(\omega, x_m)),$$

where  $\omega$  runs through all the states in  $\Omega$ , and we define  $S(m)$  to be the maximum over all  $\mathbf{x} \in X^m$  of  $S(\mathbf{x})$ . Clearly

$$\Pi_H(m) \leq \Delta_H(m) \leq S(m).$$

We bound  $S(m)$  in the following lemma, obtaining the same bound as was obtained in [2] for the case of one output. (Indeed, the proof makes essentially the same over-estimates as were made there.)

**Proposition 4.1** With the above notation, for any positive integer  $m$ ,

$$S(m) \leq \prod_{j=1}^n \Delta_j(m).$$

**Proof** For any  $i$  between 1 and  $n$ , let  $\mathcal{N}_i$  be the subnetwork induced by the input nodes and nodes 1 to  $i$ , and let

$$\sigma_i : \Omega^{(i)} \times X \rightarrow \{0, 1\}^i$$

be the output function of  $\mathcal{N}_i$ . Further, let  $S_i(m)$  be defined for the net  $\mathcal{N}_i$  in the same way as  $S(m)$  was defined for  $\mathcal{N}$ . We claim that for any  $i$  between 1 and  $n$ ,

$$S_i(m) \leq \prod_{j=1}^i \Delta_j(m),$$

from which the result will follow, since  $S_n(m) = S(m)$ . We prove the claim by induction on  $i$ .

The base case is easily seen to be true;  $S_1(m) = \Delta_1(m)$ , since the output function in this case is exactly the output of node 1.

Assume that the claim holds for  $i = k - 1$  ( $k \geq 2$ ) and consider now the case  $i = k$ . Observe that, writing  $\omega \in \Omega^{(k)}$  as  $\omega = w^{k-1}\omega_k$ , where  $w^{k-1} \in \Omega^{(k-1)}$  and  $\omega_k \in \Omega_k$ ,

$$\sigma_k(\omega^k, x) = \sigma_k(\omega^{k-1}\omega_k, x) = (\sigma_{k-1}(\omega^{k-1}, x), f^k(\omega_k, \mathbf{I}_k(\omega^{k-1}, x))).$$

Thus, for any  $\mathbf{x} = (x_1, \dots, x_m) \in X^m$ , the number of vectors of the form

$$\left( \sigma_k(\omega^{k-1}\omega_k, x_1), \dots, \sigma_k(\omega^{k-1}\omega_k, x_m) \right)$$

as  $\omega = \omega^{k-1}\omega_k$  ranges through  $\Omega^{(k)}$  is at most  $\Delta_k(m)S_{k-1}(m)$ , and hence

$$S_k(m) \leq \Delta_k(m)S_{k-1}(m) \leq \Delta_k(m) \prod_{j=1}^{k-1} \Delta_j(m) = \prod_{j=1}^k \Delta_j(m).$$

□

We say that  $\mathcal{N}$  is a feedforward linear threshold network in the case when each activation function  $f \in \mathcal{F}$  computes the inner product of  $\omega_j$  with  $\mathbf{I}_j(\omega^{j-1}, x)$  and outputs 1 if this is positive and 0 otherwise. In this case,  $H^j$  has VC dimension  $d(j)$  and, as in [2], we have the following bound.

**Corollary 4.2** Let  $H$  be the space of functions computable by a feedforward linear threshold neural network  $\mathcal{N}$  with underlying graph  $G = (V, E)$ ,  $n$  computation nodes and possibly more than one output node. Then

$$\text{VCdim}(H) \leq 2|E| \log_2(en).$$

□

**Proof** We sketch the proof, which is as in [2]. Since the hypothesis space  $H^j$  has VC dimension  $d(j)$ , it follows by Sauer's inequality that for  $m > d(j)$ ,  $\Delta_j(m) < (em/d(j))^{d(j)}$ . By Proposition 4.1,

$$S(m) \leq \prod_{j=1}^n \left( \frac{em}{d(j)} \right)^{d(j)}.$$

As in [2], one can show that this is at most  $(nem/|E|)^{|E|}$ , which is less than  $2^m$  when  $m$  is  $2|E| \log_2(en)$ . Thus, for  $m \geq 2|E| \log_2(en)$ ,  $\Pi_H(m) \leq S(m) < 2^m$ , and it follows that

$$\text{VCdim}(H) \leq 2|E| \log_2(en).$$

□

In particular, the VC dimension of the network can be bounded independently of the number of output nodes. This result, together with Theorem 3.3, provides an upper bound on the size of training sample required for the network to give valid generalisation.

Natarajan [8] has shown that for (not necessarily feedforward) linear threshold networks with  $n$  nodes (including inputs), the VC dimension is at most of the order of  $n^3 \log n$ . The above result shows that it is at most  $n^2 \log n$  for the case of feedforward linear threshold nets with  $n$  computation nodes. This extends the result of Baum and Haussler for the one-output case.

## 5 References

- [1] Martin Anthony, Norman Biggs and John Shawe-Taylor, The learnability of formal concepts, in *COLT'90, Proceedings of the Third Annual Workshop on Computational Learning Theory, Rochester, NY, 1990*, San Mateo, Ca: Morgan Kaufmann.
- [2] Eric Baum and David Haussler, What size net gives valid generalization, *Neural Computation*, 1(1) 1989 151-160.
- [3] Norman Biggs, Combinatorics and connectionism, *Discrete Mathematics*, to appear.
- [4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler and Manfred K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM*, 36(4) (1989) 929-965.
- [5] David Haussler, Quantifying inductive bias: AI learning algorithms and Valiant's learning framework, *Artificial Intelligence*, 36 (1988) 177-221.

- [6] David Haussler, preliminary extended abstract, COLT'89.
- [7] David Haussler, *Generalizing the PAC model for neural net and other learning applications*, Technical Report UCSC-CRL-89-30, University of California Computer Research Laboratory, Santa Cruz, CA, 1989.
- [8] B. K. Natarajan, On learning sets and functions, *Machine Learning 4 (1989) 67-97*.
- [9] David Pollard, *Convergence of Stochastic Processes*, New York: Springer-Verlag, 1984.
- [10] D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing*, Cambridge, MA: MIT Press, 1986.
- [11] N. Sauer, On the density of families of sets, *J. Combinatorial Theory (A)*, *13 (1972) 145-147*.
- [12] John Shawe-Taylor, Martin Anthony and Norman Biggs, *Bounding sample size with the Vapnik-Chervonenkis dimension*, Technical Report CSD-TR-618, Department of Computer Science, Royal Holloway and Bedford New College (University of London) 1989, and to appear in *Discrete Applied Mathematics*.
- [13] Leslie G. Valiant, A theory of the learnable, *Communications of the ACM*, *27 (1984) 1134-1142*.
- [14] Vladimir N. Vapnik, *Estimation of Dependences Based on Empirical Data*, New York: Springer-Verlag, 1982.
- [15] V.N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theor. Prob. Appl.*, *16 (1971) 264-280*.