Quantifying Generalization in
Linearly Weighted Neural Networks

(Short title: **Quantifying Generalization**)

Martin Anthony[1] and Sean B. Holden[2]

Abstract

The Vapnik-Chervonenkis dimension has proven to be of great use in the theoretical study of generalization in artificial neural networks. The 'probably approximately correct' learning framework is described and the importance of the VC dimension is illustrated. We then investigate the VC dimension of certain types of linearly weighted neural networks. First, we obtain bounds on the VC dimensions of radial basis function networks with basis functions of several types. Secondly, we calculate the VC dimension of polynomial discriminant functions defined over both real and binary-valued inputs.

[1]Mathematics Department, The London School of Economics and Political Science (University of London), Houghton Street, London WC2A 2AE, UK, email: `anthony@vax.lse.ac.uk`

[2]Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK, email: `sbh@eng.cam.ac.uk`

Contents

1 Linearly weighted neural networks

In this article we are interested in the study of two specific neural networks, taken from a very simple and extremely effective class of networks called *linearly weighted neural networks* (LWNNs). We are interested in using these networks to solve the standard two-class pattern classification problem, where as usual we assume that a sequence of labelled training examples is available with which we can train a network. We concern ourselves only with pattern classification problems; we do not consider the use of neural networks for tasks such as function approximation.

A LWNN computes a function $f_{\mathbf{w}} : \mathbf{R}^n \to \{0, 1\}$ given by

$$f_{\mathbf{w}}(\mathbf{x}) = \rho[w_0 + w_1\phi_1(\mathbf{x}) + \cdots + w_m\phi_m(\mathbf{x})], \tag{1}$$

where $\mathbf{w}^T = [\ w_0 \quad w_1 \quad \cdots \quad w_m\ ]$ is a vector of weights, the *basis functions* $\phi_i : \mathbf{R}^n \to \mathbf{R}$ are arbitrary, fixed functions and the function ρ is defined as

$$\rho(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We define the class $\mathcal{F}_n^\Phi$ of functions computed by the network in the obvious manner as

$$\mathcal{F}_n^\Phi = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbf{R}^{m+1}\} \tag{3}$$

where $\Phi = \{\phi_1, \ldots, \phi_m\}$ is the set of basis functions being used.

Networks of this general form have been studied extensively since the early 1960s; see, for example, Nilsson [29]. The general class of LWNNs described contains various popular network types as special cases, the most notable probably being the modified Kanerva model [36], regularization networks [32], and the two networks which we consider here: the radial basis function networks (RBFNs) introduced by Broomhead and Lowe [9] and the polynomial discriminant functions (PDFs) [12].

In the case of RBFNs we use a set of m basis functions of the form

$$\phi_i(\mathbf{x}) = \phi(\|\mathbf{x} - \mathbf{y}_i\|) \tag{4}$$

where $\mathbf{y}_i \in \mathbf{R}^n$ is a fixed *center*, $\|.\|$ is the Euclidean norm and $\phi : \mathbf{R}^+ \cup \{0\} \to \mathbf{R}$ is a fixed function. These networks are discussed in detail in section 3, where we also consider more general RBFNs. In the case of PDFs the basis functions are formed as products of elements of the input vector $\mathbf{x}$; for example,

$$\phi_i(\mathbf{x}) = x_1^2 x_2^2 x_{n-1}^5. \tag{5}$$

These networks are discussed in full in section 4.

A simple interpretation of the way in which LWNNs operate is available. Input vectors are mapped into an *extended space*[3] using the basis functions; *extended vectors* in the new space are of the form

$$\tilde{\mathbf{x}}^T = [\ \phi_1(\mathbf{x}) \quad \phi_2(\mathbf{x}) \quad \cdots \quad \phi_m(\mathbf{x})\ ]. \tag{6}$$

The aim here is to produce extended vectors in such a way that the classification problem is a *linearly separable* one in the extended space, as clearly training the network by choosing a suitable $\mathbf{w}$ now corresponds to choosing a *hyperplane* (in the extended space) which correctly divides the extended vectors. Several fast training algorithms are therefore available; see for example [14].

The reader may be surprised that we consider networks of the form of equation 1 — are these networks not completely outperformed by multilayer perceptrons? The answer is in fact a definite *no*; these networks have proved to be highly successful in practice and we believe that any casual dismissal of this type of network, although quite common, is definitely misguided. We do not discuss this issue at length here: however, the reader is referred to Broomhead and Lowe [9], Niranjan and Fallside [30], Lowe [24], Renals and Rohwer [38], Kreßel *et al.* [22] and Boser *et al.* [8] for examples of the use of RBFNs, PDFs and other linearly weighted neural networks in practical applications. A complete review is given in Holden [20].

2 The Vapnik-Chervonenkis dimension and the theory of generalization

In this section we introduce the VC dimension and the growth function and give a brief review of the associated computational learning theory in order to illustrate the importance of these parameters. A comprehensive review of the use of the VC dimension in the theory of neural networks is given in Anthony [1] and in Holden [19].

A given neural network computes a class $\mathcal{F}$ of functions $f_{\mathbf{w}} : \mathbf{R}^n \rightarrow \{0, 1\}$, the actual function computed depending on the specific weight vector used.

Definition 1 *We define the* hypothesis $h_{\mathbf{w}}$ *associated with a function* $f_{\mathbf{w}}$ *as the*

[3] We use this term as typically $m > n$.

subset of $\mathbf{R}^n$ *for which* $f_{\mathbf{w}}(\mathbf{x}) = 1$*,*

$$h_{\mathbf{w}} = \{\mathbf{x} \in \mathbf{R}^n \mid f_{\mathbf{w}}(\mathbf{x}) = 1\}. \tag{7}$$

The hypothesis space H *computed by the network is the set*

$$H = \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathbf{R}^W\} \tag{8}$$

of all hypotheses where W *is the total number of weights used by the network. (In the case of LWNNs we have* $W = m + 1$*.)*

2.1 The VC dimension

The VC dimension can be regarded as a measure of the 'capacity' of a network, or of the 'expressive power' of its hypothesis space. It was introduced, along with the growth function by Vapnik and Chervonenkis [43] in their study of the uniform convergence of relative frequencies to probabilities and has recently become important in machine learning. The reasons for its importance in this field are presented below.

Definition 2 *Given a finite set* $S \subseteq \mathbf{R}^n$ *and some function* $f_{\mathbf{w}} \in \mathcal{F}$ *we define the* dichotomy (S^+, S^-) of S induced by $f_{\mathbf{w}}$ *to be the partition of* S *into the disjoint subsets* S^+ *and* S^- *where* $S^+ \cup S^- = S$ *and* $\mathbf{x} \in S^+$ *if* $f_{\mathbf{w}}(\mathbf{x}) = 1$*,* $\mathbf{x} \in S^-$ *if* $f_{\mathbf{w}}(\mathbf{x}) = 0$*.*

Definition 3 *Given a hypothesis space* H *and finite* $S \subseteq \mathbf{R}^n$ *we define* $\triangle_H(S)$ *as the set*

$$\triangle_H(S) = \{h \cap S \mid h \in H\}. \tag{9}$$

We say that S *is* shattered *by* H *if* $\triangle_H(S) = 2^S$ *where* 2^S *is the set of all subsets of* S*.*

Note that in equation 9 in this definition, each $h \cap S$ induces a dichotomy on S, and $\triangle_H(S)$ is the set of dichotomies induced on S by H. The growth function and the VC dimension are now defined as follows.

Definition 4 (Growth Function) *The growth function is defined on the set of positive integers as*

$$\triangle_H(i) = \max_{S \subseteq \mathbf{R}^n, |S|=i} (\mid \triangle_H(S) \mid). \tag{10}$$

Definition 5 (Vapnik-Chervonenkis dimension) *The VC dimension $\mathcal{V}(H)$ of a hypothesis space H is the largest integer i such that $\triangle_H(i) = 2^i$, or infinity if no such i exists.*

The growth function thus tells us the maximum number of different dichotomies induced by $\mathcal{F}$ for any set of i points, and the VC dimension tells us the size of the largest set of points shattered by H. Note that due to the close relationship between H, $\mathcal{F}$ and the actual neural network with which we are dealing, we can refer to the growth function and the VC dimension of $\mathcal{F}$ and of the neural network and can define the quantities $\triangle_{\mathcal{F}}(S)$, $\triangle_{\mathcal{F}}(i)$ and $\mathcal{V}(\mathcal{F})$ in the obvious manner.

We now give three examples, which, since we shall use them later, we present as lemmas. Consider first the class of functions of the form

$$\mathcal{F} = \{f_{\mathbf{w}}(\mathbf{x}) = \rho[w_0 + w_1 x_1 + \cdots + w_n x_n] \mid \mathbf{w} \in \mathbf{R}^{n+1}\}, \tag{11}$$

known as the class of *linear threshold functions* (LTFs). The following result is well-known; a proof may be found in Wenocur and Dudley [45], for example.

Lemma 6 *When $\mathcal{F}$ is the class of linear threshold functions, $\mathcal{V}(\mathcal{F}) = n + 1$. Furthermore, the class of* homogeneous LTFs— *those linear threshold functions for which $w_0 = 0$—has VC dimension n* □

As a more interesting example, consider the class of feedforward networks of LTFs having W weights and thresholds and N computation nodes. A full definition of this type of network is given in [5]; it corresponds to the standard multilayer perceptron network. Note that such networks are generally not LWNNs. The upper bound in the following result is proved by Baum and Haussler [5], and the lower bound by Maass [25, 26].

Lemma 7 *Let $\mathcal{F}_{W,N}$ of functions computed by a feedforward network of LTFs having W weights and thresholds and N computation nodes. Then*

$$\mathcal{V}(\mathcal{F}_{W,N}) \leq 2W \log_2(eN). \tag{12}$$

Furthermore, there are such networks having VC dimension $\Omega(W \log_2 W)$. Thus the upper bound is asymptotically optimal up to a constant. □

Various other bounds on the VC dimension for specific networks in this class can be found in Bartlett [6]; see also [1].

Consider again the definition of $\mathcal{F}_n^\Phi$, the class of functions computed by the networks we will consider, given in equation 3. We have the following result [21, 15]. (See also the theorem of Dudley in section 4.)

Lemma 8 *Regardless of the functions in* Φ,

$$\mathcal{V}(\mathcal{F}_n^\Phi) \leq m + 1. \tag{13}$$

Furthermore, if the set of functions $\Phi = \{1, \phi_1, \ldots, \phi_m\}$ *is linearly independent, then equality holds in equation 13.* □

As is easily seen, if we let $\Phi' = \{\phi_1, \phi_2, \ldots, \phi_m\}$ then $\mathcal{V}(\mathcal{F}_n^{\Phi'}) \leq m$.

There is a well-known result, commonly known as *Sauer's lemma*, which, given the VC dimension of some class $\mathcal{F}$ of functions, provides an upper bound on the growth function.

Lemma 9 (Sauer [40], Blumer *et al.* **[7])** *Given a class* $\mathcal{F}$ *of functions for which* $\mathcal{V}(\mathcal{F}) = d \geq 0$ *and* $d < \infty$,

$$\triangle_{\mathcal{F}}(k) \leq \Psi(d, k) = 1 + \sum_{i=1}^{d} \binom{k}{i}, \tag{14}$$

where $k \geq 1$. *When* $k \geq d \geq 1$,

$$\Psi(d, k) < \left(\frac{ek}{d}\right)^d. \tag{15}$$

For finite $\mathcal{V}(\mathcal{F})$ a further useful result, from [7] (see [3]) is that either $\triangle_{\mathcal{F}}(k) = 2^k$ or,

$$\triangle_{\mathcal{F}}(k) \leq k^{\mathcal{V}(\mathcal{F})+1}. \tag{16}$$

Clearly if $\mathcal{V}(\mathcal{F}) = \infty$ then $\triangle_{\mathcal{F}}(k) = 2^k$ for all k.

2.2 Using the growth function and the VC dimension to analyze generalization

Consider a neural network which computes a class $\mathcal{F}$ of functions. We can regard the process of training this network as a process of trying to find some $f_{\mathbf{w}} \in \mathcal{F}$

which is a 'good approximation' to a given target function f_T on a set of training examples. Let $\mathbf{x} \in \mathbf{R}^n$ be chosen at random according to some arbitrary (but fixed) probability distribution P on $\mathbf{R}^n$. We define $\pi_{f_\mathbf{w}}$ to be the probability that $f_\mathbf{w}$ agrees with the target function on an example chosen at random according to the distribution P; that is,

$$\pi_{f_\mathbf{w}} = \Pr\left[f_\mathbf{w}(\mathbf{x}) = f_T(\mathbf{x})\right], \tag{17}$$

where the probability is taken over all possible examples $\mathbf{x}$. Let

$$T_k = ((\mathbf{x}_1, f_T(\mathbf{x}_1)), \ldots, (\mathbf{x}_k, f_T(\mathbf{x}_k)))$$

be a sequence of k training examples where the inputs $\mathbf{x}_i$ are picked independently according to P and define $\upsilon_{f_\mathbf{w}}$ to be the fraction of the inputs in T_k which are classified correctly by $f_\mathbf{w}$,

$$\upsilon_{f_\mathbf{w}} = \frac{1}{k}\ \{i : f_\mathbf{w}(\mathbf{x}_i) = f_T(\mathbf{x}_i)\}\ .$$

When we train a neural network we choose a particular vector of weights $\mathbf{w}$ on the basis of the value of $\upsilon_{f_\mathbf{w}}$, and we thus need to know whether $\upsilon_{f_\mathbf{w}}$ converges to $\pi_{f_\mathbf{w}}$ in a uniform way for all $f_\mathbf{w} \in \mathcal{F}$ as k becomes large. If this is not the case then we may end up choosing a function $f_\mathbf{w}$ for which the value of $\pi_{f_\mathbf{w}}$ is in fact relatively low. An inequality derived in [42] yields a bound on the probability that there is a function $f_\mathbf{w} \in \mathcal{F}$ for which $\pi_{f_\mathbf{w}}$ and $\upsilon_{f_\mathbf{w}}$ differ significantly. Specifically, given a particular value of α,

$$\Pr\left[\text{there is } f_\mathbf{w} \in \mathcal{F} \text{ s.t. } \upsilon_{f_\mathbf{w}} - \pi_{f_\mathbf{w}} > \alpha\sqrt{1-\pi_{f_\mathbf{w}}}\right] \le 4\triangle_\mathcal{F}(2k)\exp\left(\frac{-\alpha^2 k}{4}\right). \tag{18}$$

Here, the probability referred to is the distribution, over all sequences T_k of k training examples, obtained by choosing each of the k examples independently at random from $\mathbf{R}^n$ according to the probability distribution P. (The result quoted here is based on a slight improvement on the original result of Vapnik; see Anthony and Shawe-Taylor [4].) Now, by equation 16, if $\mathcal{V}(\mathcal{F})$ is finite then $\triangle_\mathcal{F}(k)$ is bounded above by a polynomial function of k and thus, since $\exp\left(\frac{-\alpha^2 k}{4}\right)$ decays exponentially in k, we can make the right-hand side of equation 18 arbitrarily small by choosing k large enough. Furthermore, equation 18 provides a bound on the rate of convergence which is independent of the particular probability distribution P and the particular target function f_T. The usefulness of this result will soon become apparent. Roughly speaking, it tells us that, given any δ between 0 and 1, then provided k is larger than some constant which does not depend on either f_T or P, the following holds with probability at least $1-\delta$: for any target function f_T and for any probability distribution on the examples, $\pi_{f_\mathbf{w}}$

and $v_{f_{\mathbf{w}}}$ are close for a randomly chosen sample. The VC dimension influences the speed of convergence of the quantity on the right-hand side of equation 18 and, consequently, the size of k required to guarantee a particular generalization performance.

2.3 VC dimension and computational learning theory

The above discussion illustrates one reason for the importance of the VC dimension and growth function in the analysis of generalization in neural networks and other systems. In their analysis of valid generalization in general feedforward networks of LTFs, Baum and Haussler [5] used a modified form of *Probably Approximately Correct (PAC) learning theory*, introduced by Blumer *et al.* [7] and based on the work of Valiant [41], to relate network size to generalization ability. This work has recently been extended to a class of networks described in section 1 by Holden and Rayner [21], to networks with more than one output node by Anthony and Shawe-Taylor [4], and to networks with real-valued outputs by Haussler [18] . In this section we give a brief introduction to the formalism.

2.3.1 Standard PAC learning

Consider a neural network having a hypothesis space H. We define a *concept class* C in a similar manner as a set of subsets of $\mathbf{R}^n$. In general we also impose some further restrictions on C and H, details of which can be found in [7]; these are rather technical and do not introduce problems for the neural networks likely to be used in practice. The concept class C may or may not be equal to the hypothesis space H. Now, given a target concept $c_T \in C$, training corresponds to choosing a weight vector $\mathbf{w}$ such that the hypothesis $h_{\mathbf{w}}$ is a good approximation to c_T.

Once again we have a sequence $T_k = ((\mathbf{x}_1, o_1), \ldots, (\mathbf{x}_k, o_k))$ of k training examples where the inputs $\mathbf{x}_i$ are drawn independently from an arbitrary distribution P on $\mathbf{R}^n$ and o_i is equal to 1 if $\mathbf{x}_i \in c_T$ and 0 otherwise. We define a *learning function* for C as a function which given a T_k for large enough k and any $c_T \in C$ will return a hypothesis $h_{\mathbf{w}} \in H$ which is, with high probability, a good approximation to c_T. Formally, the *error* of a hypothesis $h_{\mathbf{w}}$ with respect to c_T and P is the probability, over $\mathbf{R}^n$, according to P of the symmetric difference $h_{\mathbf{w}} \triangle c_T$. Given small, specified ϵ and δ, we demand that there is some k, which does not depend on either the probability distribution P or on c_T, such that the hypothesis $h_{\mathbf{w}}$

produced by the learning function satisfies

$$\Pr[\text{Error of } h_{\mathbf{w}} > \epsilon] \leq \delta. \tag{19}$$

The probability referred to here is that distribution on all possible sequences of k training examples which results when each of the k examples is chosen from $\mathbf{R}^n$ according to the distribution P, independently of the other examples. It is worth emphasizing again that we require there to be a suitable k which depends *only* on ϵ and δ. The *sample complexity* of the learning function is the smallest value of k guaranteed to achieve this, and any concept class for which there is such a learning function is said to be *uniformly learnable*.

An important result proved in [7] is that C is uniformly learnable if and only if $\mathcal{V}(C)$ is finite. An account of PAC learning theory can be found in [3].

2.3.2 Extended PAC learning

Some shortcomings of PAC learning as described above should immediately be apparent. In this formalism there is no satisfactory way in which to deal with a sequence T_k containing misclassifications. There is also no way in which to deal with a target concept which has been defined in a stochastic manner — a common assumption in pattern recognition — rather than with a deterministic concept c_T.

PAC learning is extended in [7] in such a way that T_k is generated by drawing examples independently from an arbitrary distribution P' on $\mathbf{R}^n \times \{0, 1\}$. The error with respect to P' of a function $f_{\mathbf{w}}$ computed by a neural network is then defined as

$$\Pr[f_{\mathbf{w}}(\mathbf{x}) \neq o], \tag{20}$$

where the probability is over all $(\mathbf{x}, o)$, with respect to P'. Note that, in general, we might not have a deterministic target concept, as given some $\mathbf{x} \in \mathbf{R}^n$ both $(\mathbf{x}, 1)$ and $(\mathbf{x}, 0)$ may have non-zero probability. Note, however, that a (deterministic) target concept, together with a probability distribution P on $\mathbf{R}^n$, may be represented as such a distribution P' (see [7, 1]). Thus, the present model encompasses the basic model. We are now able to model the situation in which examples in T_k are generated as in the standard PAC learning model but where $\mathbf{x}_i$ or o_i are subsequently modified by some random process.

In a similar manner to that described above, this extended PAC learning formalism requires us to search for a hypothesis $h_{\mathbf{w}} \in H$ which, with high probability, is a good approximation to a particular stochastic target concept. In order to illustrate the importance of the growth function and the VC dimension in the theory we state the following theorem, which follows from the same general result of Vapnik as does equation 18. Some measurability conditions on the class $\mathcal{F}$ of functions computed by the network must be satisfied; see Blumer *et al.* [7] and Pollard [33] for details; these are once again never a cause for concern in practice. For applications of this theorem see Baum and Haussler [5], Holden and Rayner [21] and Anthony and Shawe-Taylor [4]. Before stating the theorem, it is useful to introduce some notation. For $f_{\mathbf{w}} \in \mathcal{F}$, and for $T_k = ((\mathbf{x}_1, o_1), (\mathbf{x}_2, o_2), \ldots, (\mathbf{x}_k, o_k)) \in (\mathbf{R}^n \times \{0,1\})^k$, we define $v_{f_{\mathbf{w}}}$ to be

$$v_{f_{\mathbf{w}}} = \frac{1}{k} \left|\{i : f_{\mathbf{w}}(\mathbf{x}_i) = o_i\}\right|,$$

the fraction of examples $(\mathbf{x}, o)$ in the sample which 'agree' with $f_{\mathbf{w}}$. Further, let $\pi_{f_{\mathbf{w}}} = P'\{(\mathbf{x}, o) : f_{\mathbf{w}}(\mathbf{x}) = o\}$. Thus, $v_{f_{\mathbf{w}}}$ is a sample-based estimate of $\pi_{f_{\mathbf{w}}}$. The following result enables us to bound the probability that a sample is misleading, in the sense that $v_{f_{\mathbf{w}}}$ is large, yet $\pi_{f_{\mathbf{w}}}$ is quite substantially smaller. More specifically, given two numbers γ and ϵ between 0 and 1, we should like it to be the case that with high probability, if the 'agreement' of a function on a random sample satisfies $v_{f_{\mathbf{w}}} > 1 - (1-\gamma)\epsilon$, then the 'true agreement', $\pi_{f_{\mathbf{w}}}$ satisfies $\pi_{f_{\mathbf{w}}} > 1 - \epsilon$. The following result is a consequence of the result of Vapnik described in equation 18.

Theorem 10 (Vapnik [42], Blumer *et al.* **[7])** *Consider the class $\mathcal{F}$ of functions $f_{\mathbf{w}} : \mathbf{R}^n \to \{0,1\}$ and a sequence T_k of examples as described above. Let γ and ϵ be such that $0 < \gamma, \epsilon \leq 1$ and define $\mathcal{P}$ as the probability that* every *function $f_{\mathbf{w}} \in \mathcal{F}$ such that $v_{f_{\mathbf{w}}} > 1 - (1-\gamma)\epsilon$ also satisfies $\pi_{f_{\mathbf{w}}} > 1 - \epsilon$. Then $\mathcal{P}$ satisfies*

$$\mathcal{P} > 1 - 4\triangle_{\mathcal{F}}(2k) \exp\left(\frac{-\gamma^2 \epsilon k}{4}\right). \tag{21}$$

This theorem is important because, clearly, if we can find an upper bound on the growth function of the network, for example by finding its VC dimension and applying Sauer's lemma, then we can say something about its ability to generalize. Specifically, if our network can be trained to classify correctly a fraction $1-(1-\gamma)\epsilon$ of the k training examples, the probability that its error is less than ϵ is at least $\mathcal{P}$. This is exactly the type of analysis carried out in [5, 21], and as the growth function and VC dimension tend to depend quite specifically on the *size* of the network measured in terms of, for example, the total number of parameters adapted during training, this type of analysis generally allows us to relate the size of a network to the number of examples on which the network should be trained in order to obtain valid generalization with high probability. We remark that such analysis is independent of the particular learning function or learning algorithm being used; in this sense, Theorem 10 may appear to be stronger than is necessary in practice. Nonetheless, there are results [7, 17, 3] showing that, no matter what learning function is used, the required number of training samples for PAC learning must still be bounded below by a quantity depending on the VC dimension.

3 Radial basis function networks

Radial basis function networks in their most general form (when used for classification, rather than function approximation) compute functions $f_{\mathbf{w}} : \mathbf{R}^n \to \{0, 1\}$ where,

$$f_{\mathbf{w}} = \rho[\overline{f}_{\mathbf{w}}(\mathbf{x})]. \tag{22}$$

The function $\overline{f}_{\mathbf{w}} : \mathbf{R}^n \to \mathbf{R}$ is of the form,

$$\overline{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} \lambda_i \phi(\|\mathbf{x} - \mathbf{y}_i\|) + \sum_{i=1}^{q} \theta_i \psi_i(\mathbf{x}) \text{ where } q \leq p \tag{23}$$

in which $\mathbf{w}^T = [\ \lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_p \quad \theta_1 \quad \theta_2 \quad \cdots \quad \theta_q\ ]$ is a vector of weights, $\mathbf{y}_i \in \mathbf{R}^n$ are *centers* of the basis functions, $\phi : \mathbf{R}^+ \cup \{0\} \to \mathbf{R}$, $\|.\|$ is a norm, which in this article is assumed to be the Euclidean norm, and $\{\psi_i \mid i = 1, \ldots, q\}$ is a basis of the vector space $\pi_{d-1}(\mathbf{R}^n)$ of algebraic polynomials from $\mathbf{R}^n$ to $\mathbf{R}$ of degree at most $(d - 1)$ for some specified d.

Networks of this type were originally introduced by Broomhead and Lowe [9], whose work should be consulted for further details. Their main motivation was that, as we shall see below, the networks have a sound theoretical basis in interpolation theory. The networks can also be regarded as a special case of the regularization networks introduced by Poggio and Girosi [32] and thus have a theoretical justification in terms of standard regularization theory. Networks of this general type have been shown to perform well in comparison to many available alternatives — see, for example, Niranjan and Fallside [30] — and training algorithms are available which are considerably faster than hidden layer backpropagation; see for example Moody and Darken [28] and Chen *et al.* [11].

It is usual in practice not to include the polynomial terms in the network, so that the network computes functions,

$$f_{\mathbf{w}}(\mathbf{x}) = \rho \left[\sum_{i=1}^{p} \lambda_i \phi(\|\mathbf{x} - \mathbf{y}_i\|) \right]. \tag{24}$$

A single constant offset term λ_0 is often added to the summation, but is omitted here.

In this section we investigate the VC dimension of radial basis function networks, using various standard choices for the basis function ϕ. We will mostly be interested in networks where the centers $\mathbf{y}_i$ are fixed, although we briefly mention networks with variable centers in section 3.3. Our proof technique relies on the interpolation properties of the functions $\overline{f}_{\mathbf{w}}$, and in particular on the use of two well known theorems due to Micchelli [27].

3.1 Interpolation and Micchelli's theorems

Why use functions of the form of equation 22? Broomhead and Lowe [9] introduced RBFNs on the basis that functions of the form of $\overline{f}_{\mathbf{w}}$ had previously proved very useful in the theory of multivariable interpolation (a review is given by Powell [34, 35]; see also [32] on which our review is based).

Consider the problem of finding a function $g : \mathbf{R}^n \to \mathbf{R}$ which is a member of a given class of functions $\mathcal{G}$ and which exactly interpolates a set T_k of k examples,

$$T_k = \{(\mathbf{x}_1, o_1), \ldots, (\mathbf{x}_k, o_k)\} \tag{25}$$

where $\mathbf{x}_i \in \mathbf{R}^n$ are distinct vectors and $o_i \in \mathbf{R}$ can be chosen arbitrarily. This means that g must satisfy

$$g(\mathbf{x}_i) = o_i \text{ for } i = 1, \ldots, k. \tag{26}$$

Now let $\mathcal{G}'$ denote the class of functions $\mathcal{G}' = \{\rho \circ g \mid g \in \mathcal{G}\}$ where $\circ$ denotes function composition. Assume we have a *particular* set $S_k = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ of k points where $\mathbf{x}_i \in \mathbf{R}^n$, and we form a corresponding set T_k having arbitrary o_i. Now clearly if we can prove that given such a set T_k there exists, regardless of the o_i used, a $g \in \mathcal{G}$ which performs the interpolation, then $\mathcal{V}(\mathcal{G}') \geq k$. This is because, given any particular dichotomy (S_k^+, S_k^-) of the initial set S_k, we simply pick o_i to be an arbitrary positive quantity when $\mathbf{x}_i \in S_k^+$ and an arbitrary negative quantity when $\mathbf{x}_i \in S_k^-$. As there is a $g \in \mathcal{G}$ which interpolates the corresponding T_k, the corresponding $g' = \rho \circ g$ induces the required dichotomy, and as this applies to *any* dichotomy, $\mathcal{G}'$ shatters S_k.

The functions $\overline{f}_{\mathbf{w}}$ are useful because it is always possible to interpolate k points in such a set S_k using a function

$$\overline{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{k} \lambda_i \phi(\|\mathbf{x} - \mathbf{x}_i\|) + \sum_{i=1}^{q} \theta_i \psi_i(\mathbf{x}), \text{ where } q \leq k, \tag{27}$$

regardless of the values chosen for o_i, provided ϕ satisfies some simple conditions which we discuss below. The class of functions $\mathcal{G}$ is now simply

$$\mathcal{G} = \{\overline{f}_{\mathbf{w}} \mid \mathbf{w} \in \mathbf{R}^{k+q}\}, \tag{28}$$

where $\overline{f}_{\mathbf{w}}$ is as defined in equation 27. Notice that in equation 27 the original centers $\mathbf{y}_i$ of $\overline{f}_{\mathbf{w}}$ have been made to correspond to the points $\mathbf{x}_i$. Notice also that when using functions $\overline{f}_{\mathbf{w}}$ in this manner the constraints of equation 26 give us a set of k linear equations for $(k+q)$ coefficients. The remaining degrees of freedom are fixed by requiring that,

$$\sum_{i=1}^{k} \lambda_i \psi_j(\mathbf{x}_i) = 0 \text{ where } j = 1, \ldots, q. \tag{29}$$

A sufficient condition on ϕ for the existence of an interpolating function of the form of equation 27 is that $\phi \in \mathbf{P}_d(\mathbf{R}^n)$ where $\mathbf{P}_d(\mathbf{R}^n)$ is the set of *strictly conditionally positive definite (SCPD) functions of order* d, defined as follows.

Definition 11 *Suppose* h *is a continuous function on* $[0, \infty)$. *This function is strictly conditionally positive definite of order* $d \geq 1$ *on* $\mathbf{R}^n$ *if for any* k *distinct points* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *in* $\mathbf{R}^n$ *and* $c_1, \ldots, c_k \in \mathbf{R}$ *(not all* 0*) where*

$$\sum_{i=1}^{k} c_i \psi(\mathbf{x}_i) = 0 \tag{30}$$

for all $\psi \in \pi_{d-1}(\mathbf{R}^n)$, *the quadratic form* $\sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j h(\|\mathbf{x}_i - \mathbf{x}_j\|)$ *is positive. The function is SCPD of order* 0 *if the form* $\sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j h(\|\mathbf{x}_i - \mathbf{x}_j\|)$ *is positive definite.*

Let $\mathbf{P}_d$ be the set of functions which are in $\mathbf{P}_d(\mathbf{R}^n)$ over any $\mathbf{R}^n$,

$$\mathbf{P}_d = \bigcap_{n \geq 1} \mathbf{P}_d(\mathbf{R}^n). \tag{31}$$

Note that for all non-negative integers d, $\mathbf{P}_d \subseteq \mathbf{P}_{d+1}$. An important theorem due to Micchelli provides us with a simple means of determining whether a function ϕ is in $\mathbf{P}_d$, and hence whether it is a suitable basis function for use in forming $\overline{f}_{\mathbf{w}}$. We first need to define a *completely monotonic* function.

Definition 12 *A function* h *is* completely monotonic *on* $(0, \infty)$ *if* $h \in C^\infty(0, \infty)$ *and its sequence of derivatives is such that*

$$(-1)^i h^{(i)}(x) \geq 0 \tag{32}$$

for $x \in (0, \infty)$ *and* $i = 0, 1, 2, \ldots$.

Theorem 13 (Micchelli [27], Dyn and Micchelli [16]) *If a function $h(r)$ is continuous on $[0,\infty)$, $h(r^2) \in C^\infty(0,\infty) \cap C[0,\infty)$ and $(-1)^d h^{(d)}$ is completely monotonic on $(0,\infty)$ but not constant, then $h(r^2)$ is in $\mathbf{P}_d$.*

Now, consider the special case in which we attempt to interpolate the data in T_k using

$$\overline{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{k} \lambda_i \phi(\|\mathbf{x} - \mathbf{x}_i\|). \tag{33}$$

The function $\rho \circ \overline{f}_{\mathbf{w}}$ now corresponds to the networks most often used in practice. The interpolation is possible provided we can find a solution to the set of equations

$$\mathbf{o} = \begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_k \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1k} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{k1} & \phi_{k2} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix} = \boldsymbol{\phi}\boldsymbol{\lambda} \tag{34}$$

where $\phi_{ij} = \phi(\|\mathbf{x}_i - \mathbf{x}_j\|)$, so that

$$\boldsymbol{\lambda} = \boldsymbol{\phi}^{-1}\mathbf{o}. \tag{35}$$

It is possible to show (see Powell [35]) that $\boldsymbol{\phi}$ is nonsingular if ϕ is SCPD of order 0, or if ϕ is SCPD of order 1 and $\phi(0) \leq 0$. Thus, in some cases theorem 13 will tell us whether a particular ϕ can be used successfully. An alternative sufficient condition also exists for this special case, again proved by Micchelli.

Theorem 14 (Micchelli [27]) *If h is continuous on $[0,\infty)$, positive on $(0,\infty)$, and has a first derivative that is completely monotonic but not constant on $(0,\infty)$, then for any set of k vectors $\mathbf{x}_i \in \mathbf{R}^n$ where n is arbitrary,*

$$(-1)^{k-1} \det h(\|\mathbf{x}_i - \mathbf{x}_j\|^2) > 0. \tag{36}$$

Now, clearly, if we choose a suitable function ϕ such that $\phi(\sqrt{r})$ satisfies the conditions in theorem 14 it is not possible that $\det(\boldsymbol{\phi}) = 0$, which implies that $\boldsymbol{\phi}$ must be nonsingular and consequently that there exists a suitable weight vector $\boldsymbol{\lambda}$ regardless of the actual values for o_i used.

In summary, provided we use a basis function ϕ chosen using the relevant conditions given in theorem 13 or theorem 14, then our radial basis function network,

Form of basis function	Type of basis function
$\phi_{LIN}(r) = r$	Linear
$\phi_{CUB}(r) = r^3$	Cubic
$\phi_{TPS}(r) = r^2 \ln r$	Thin plate spline
$\phi_{MQ}(r) = (r^2 + c^2)^\beta, c \in \mathbf{R}^+, 0 < \beta < 1$	Generalized Multiquadric
$\phi_{IMQ}(r) = (r^2 + c^2)^{-\alpha}, c \in \mathbf{R}^+, \alpha > 0$	Generalized Inverse Multiquadric
$\phi_{GAUSS}(r) = \exp\left[-\left(\frac{r}{c}\right)^2\right], c \in \mathbf{R}^+$	Gaussian

Table 1: Standard basis functions used in radial basis function networks.

having fixed centers and as defined in equation 22 shatters the set of p vectors $\{\mathbf{x}_i\}$ which correspond to the centers $\{\mathbf{y}_i\}$ such that

$$\mathbf{x}_i = \mathbf{y}_i \text{ where } i = 1, \ldots, p. \tag{37}$$

It therefore has a VC dimension of at least p.

3.2 Networks with fixed centers

Table 1 summarizes some of the usual basis functions ϕ used in RBFNs. The use of these functions is justified by the theory introduced above [32]. Note that the parameter c is fixed — it is not adapted during training. We immediately obtain the following two corollaries.

Corollary 15 *Consider the simple RBFNs of the form,*

$$f_{\mathbf{w}}(\mathbf{x}) = \rho\left[\sum_{i=1}^{p} \lambda_i \phi(\|\mathbf{x} - \mathbf{y}_i\|)\right] \tag{38}$$

where the centers $\mathbf{y}_i$ are fixed and distinct. If ϕ is one of the functions ϕ_{LIN}, ϕ_{GAUSS}, ϕ_{MQ} or ϕ_{IMQ} then the VC dimension $\mathcal{V}(\mathcal{F})$ of the network is exactly p.

Proof: The functions ϕ_{GAUSS} and ϕ_{IMQ} are in $\mathbf{P}_0$ by theorem 13 and the functions $\sqrt{r}$ and $(r+c^2)^\beta$ where $0 < \beta < 1$ satisfy the conditions in theorem 14. This means that by the arguments given above $\mathcal{V}(\mathcal{F}) \geq p$ for all four cases. Also, from Lemma 8 we know that $\mathcal{V}(\mathcal{F}) \leq p$, and consequently we must have $\mathcal{V}(\mathcal{F}) = p$. $\square$

Corollary 16 *Consider the RBFNs of the form,*

$$f_{\mathbf{w}}(\mathbf{x}) = \rho \left[\sum_{i=1}^{p} \lambda_i \phi(\|\mathbf{x} - \mathbf{y}_i\|) + \psi(\boldsymbol{\theta}, \mathbf{x}) \right] \tag{39}$$

where $\psi(\boldsymbol{\theta}, \mathbf{x})$ *is the degree* 1 *polynomial,*

$$\psi(\boldsymbol{\theta}, \mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n, \tag{40}$$

x_i *are the elements of* $\mathbf{x}$*, and* $p \geq n+1$*. Again, the centers* $\mathbf{y}_i$ *are fixed and distinct. If* ϕ *is the function* ϕ_{CUB} *or* ϕ_{TPS} *then the VC dimension of the network obeys* $p \leq \mathcal{V}(\mathcal{F}) \leq p+n+1$*.*

Proof: By theorem 13 both ϕ_{CUB} and ϕ_{TPS} are in $\mathbf{P}_2$ and hence $\mathcal{V}(\mathcal{F}) \geq p$. From Lemma 8 we know that $\mathcal{V}(\mathcal{F}) \leq p+n+1$ and the results follows. □

3.3 Networks with variable centers

What happens to the VC dimension of a radial basis function network if we allow its centers $\mathbf{y}_i$ to adapt during training, rather than force them to remain fixed? Obviously, the results presented above provide lower bounds on the VC dimension of RBFNs having basis functions ϕ of the appropriate type. We also have the following simple result.

Corollary 17 *Consider the networks of the types mentioned in corollaries 15 and 16. If the centers* $\mathbf{y}_i$ *are allowed to adapt then the networks can all form* arbitrary *dichotomies of* any *set of* p *distinct points.*

The proof of this result is trivial — the networks can shatter the set of p points corresponding to the p centers $\mathbf{y}_i$ and these centers can now be placed anywhere. It is interesting that there is no requirement that the p points be in any kind of *general position* as is often the case in similar results for other types of network (see Cover [12]).

Corollary 17 suggests that the lower bounds suggested for the VC dimension of this type of network are unlikely to be tight. We have not been able to improve them however, and we leave as an open question whether or not it is possible to obtain a lower bound similar to that recently proved by Maass [25, 26] for certain feedforward networks, and mentioned in Lemma 7. Lee *et al.* [23] have shown that the VC dimension of a RBFN of the type considered in corollary 15 having variable centers and Gaussian basis functions with $c = 1$ is at least $pn - n$, which is (approximately) proportional to the number of variable parameters in the network. However, it is not known whether a similar result also applies for other standard basis functions such as those given in table 1. Similarly, we have not been able to prove upper bounds on the VC dimension of these networks for all but the simplest cases, such as when

$$\phi(r) = r^i \text{ where } i \text{ is even.} \tag{41}$$

Then the network computes a class of polynomial discriminant functions and the results of section 4 can be applied.

4 Polynomial discriminant functions

In this section, we discuss the polynomial discriminant functions (PDFs), determining the VC dimension in two distinct situations: when the inputs are real numbers and when the inputs are restricted to be binary-valued (that is, 0 or 1). As mentioned in Section 1, the PDFs are linearly weighted neural networks in which the basis functions compute some of the products of the entries of the input vectors. In other words,

$$\tilde{\mathbf{x}}^T = [\ \phi_1(\mathbf{x}) \quad \phi_2(\mathbf{x}) \quad \cdots \quad \phi_m(\mathbf{x})\ ], \tag{42}$$

where each ϕ_i is of the form

$$\phi_i(\mathbf{x}) = \prod_{1 \leq j \leq n} x_j^{r_j}, \tag{43}$$

for some non-negative integers r_i. We say that the PDF f is of *order at most* k when the largest degree of any of the multinomial basis functions ϕ_i used to define f is k; that is, if f is a LWNN over those basis functions in equation 43 having $\sum_{i=1}^{n} r_i \leq k$. Furthermore, the order of a PDF f is said to be precisely k when f has order at most k but not at most $k - 1$; that is, when in every representation of f in the form given in equation 1, one of the basis functions required has degree k. Thus the PDFs of order 1 are precisely the linear threshold functions of Lemma 6 and, for example, the PDFs of order 2 defined on $\mathbf{R}^3$ are of the form

$$\rho[w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_1^2 + w_5x_2^2 + w_6x_3^2 + w_7x_1x_2 + w_8x_1x_3 + w_9x_2x_3] \tag{44}$$

for some contants w_i, $(0 \leq i \leq 9)$, in which at least one of the terms of degree 2 has a non-zero coefficient. (Note that a PDF of this form, *over* $\{0,1\}^3$, can be reduced to one of degree 1 unless one of w_7, w_8, w_9 is non-zero, simply because if x_i is 0 or 1, then $x_i^2 = x_i$. This simple observation will prove useful below.)

Polynomial discriminators have been studied in the the context of pattern classification (see, for example, [14, 13, 29]), where the aim is to classify a given set of training data points into two categories correctly, this classification being used as a (hopefully valid) means of classifying further points. In addition, they have recently been employed in signal processing [37]. It is therefore an important problem to determine the 'power' of classification achievable by such discriminators and to quantify the sample size required for valid learning.

Much work has been done on the representation of functions by PDFs; we refer to [39, 10, 2, 31, 44].

We shall denote by $\mathcal{P}(n, k)$ the (full) class of PDFs of order at most k defined on $\mathbf{R}^n$. Thus, $\mathcal{P}(n, k)$ is the set of linearly weighted neural networks formed from all basis functions of degree at most k of the form ϕ_i given in equation 43. Further, we shall denote by $\mathcal{P}_{\mathbf{B}}(n, k)$ the (full) class of *boolean PDFs* obtained by restricting $\mathcal{P}(n, k)$ to binary-valued inputs; i.e., to $\{0, 1\}^n$. Thus $\mathcal{P}(n, k)$ is the set of $\{0, 1\}$ functions on $\mathbf{R}^n$ whose positive and negative examples are separated by some surface which can be described by a multinomial equation of degree at most k and $\mathcal{P}_{\mathbf{B}}(n, k)$ is the set of $\{0, 1\}$ functions on $\{0, 1\}^n$ (i.e, Boolean functions of n variables) whose positive and negative examples can be separated in this way. To start with, we consider only these two classes of PDFs. Later we shall discuss more restricted classes; for example, one may be interested only in PDFs over a restricted set of all basis functions ϕ_i of at most a given degree.

4.1 Further notations and definitions

Let us denote the set $\{1, 2, \ldots, n\}$ by $[n]$. We shall denote the set of all subsets of at most k objects from $[n]$ by $[n]^{(k)}$ and we shall denote by $[n]^k$ the set of all selections, in which repetition is allowed, of at most k objects from $[n]$. Thus, $[n]^k$ may be thought of as a collection of 'multi-sets'. For example, $[3]^{(2)}$ consists of the sets

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\},$$

while $[3]^2$ consists of the multisets

$$\emptyset, \{1\}, \{1, 1\}, \{2\}, \{2, 2\}, \{3\}, \{3, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}.$$

In general, $[n]^{(k)}$ consists of $\sum_{i=0}^{k} \binom{n}{i}$ sets, and $[n]^k$ consists of $\binom{n+k}{k}$ multisets. With a slight abuse of mathematical notation, $[n]^{(k)} \subseteq [n]^k$. For each $\emptyset \neq S \in [n]^k$, and for any $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$, $\mathbf{x}_S$ denotes the product of the x_i for $i \in S$ (with repetitions as required). For example, $\mathbf{x}_{\{1,2,3\}} = x_1 x_2 x_3$ and $\mathbf{x}_{\{1,1,2\}} = x_1^2 x_2$. We define $\mathbf{x}_\emptyset = 1$ for all $\mathbf{x}$.

It is clear that the basis functions ϕ_i for the PDFs may be written in the form $\phi_i(\mathbf{x}) = \mathbf{x}_S$ for some non-empty multi-set S. Therefore a function defined on $\mathbf{R}^n$ is a PDF of order at most k if and only if there are constants w_S, one for each $S \in [n]^k$, such that

$$f(\mathbf{x}) = \rho\left[\sum_{S \in [n]^k} w_S \mathbf{x}_S\right]. \tag{45}$$

Restricting attention to $\{0,1\}$ inputs, note that any term $\mathbf{x}_S$ in which S contains a repetition is redundant, simply because for $x = 0$ or 1, $x^r = x$ for all r; thus, for example, for binary inputs, $x_1 x_2^2 x_3^3 = x_1 x_2 x_3$. Therefore, we arrive at the following characterisation of $\mathcal{P}_{\mathbf{B}}(n,k)$. A function $f : \{0,1\}^n \to \{0,1\}$ is in $\mathcal{P}_{\mathbf{B}}(n,k)$ if and only if there are constants w_S, one for each $S \in [n]^{(k)}$, such that

$$f(\mathbf{x}) = \rho\left[\sum_{S \in [n]^{(k)}} w_S \mathbf{x}_S\right]. \tag{46}$$

Of course, each boolean PDF is the restriction to $\{0,1\}^n$ of a PDF; what we have emphasized here is that in the case when the inputs are restricted to be 0 or 1, some redundancy can be eliminated immediately. This last observation shows that in considering the classes $\mathcal{P}_{\mathbf{B}}(n,k)$, it suffices to use extended vectors of the form

$$\tilde{\mathbf{x}}^T = [\ \psi_1(\mathbf{x}) \quad \psi_2(\mathbf{x}) \quad \cdots \quad \psi_m(\mathbf{x})\ ], \tag{47}$$

where each ψ_i is of the form

$$\psi_i(\mathbf{x}) = \mathbf{x}_S = \prod_{i \in S} x_i, \tag{48}$$

for a non-empty subset S of at most k elements of $[n]$. The number of such S, and hence the length of these extended vectors, is $\sum_{i=1}^{k} \binom{n}{i}$. For general PDFs of order at most k, one uses the extended vectors of equation 43, of length $\binom{n+k}{k} - 1$; the entries here are $\mathbf{x}_S$ for $\emptyset \neq S \in [n]^k$.

4.2 VC dimension and independence of basis functions

As noted earlier, classification by LWNNs corresponds to classification by linear threshold functions of the extended vectors in the corresponding higherdimensional space. This is explicit in the context of PDFs and boolean PDFs from equations 45 and 46.

We shall make use of the following well-known characterisation of sets shattered by homogeneous linear threshold functions, a proof of which we include for completeness.

Lemma 18 *A subset* $S = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_s\}$ *of* $\mathbf{R}^d$ *can be shattered by the set of homogeneous linear threshold functions on* $\mathbf{R}^d$ *if and only if* S *is a linearly independent set of vectors.*

Proof: Suppose that the vectors are linearly dependent. Then at least one of the vectors is a linear combination of the others. Without loss, suppose that $\mathbf{y}_1 = \sum_{i=2}^{s} \lambda_i \mathbf{y}_i$ for some constants λ_i, $(2 \leq i \leq s)$. Let $\langle \mathbf{x}, \mathbf{y} \rangle$ denote the standard (Euclidean) inner product on $\mathbf{R}^d$. Suppose $\mathbf{w}$ is such that for $2 \leq j \leq s$, $\langle \mathbf{w}, \mathbf{y}_j \rangle > 0$ if and only if $\lambda_j > 0$. Then $\langle \mathbf{w}, \mathbf{y}_1 \rangle = \sum_{i=2}^{s} \lambda_i \langle \mathbf{w}, \mathbf{y}_i \rangle \geq 0$. It follows that there is no homogeneous linear threshold function for which $\mathbf{y}_1$ is a negative example and, for $2 \leq j \leq s$, $\mathbf{y}_j$ is a positive example if and only if $\lambda_j > 0$. That is, the set S of vectors is not shattered.

For the converse, it suffices to prove the result when $s = d$. Let $\mathbf{A}$ be the matrix whose rows are the vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_d$ and let $\mathbf{v}$ be any of the 2^d vectors with entries $1, -1$. Then $\mathbf{A}$ is nonsingular and so the matrix equation $\mathbf{A}\mathbf{w} = \mathbf{v}$ has a solution. The homogeneous linear threshold function t defined by this solution weight-vector $\mathbf{w}$ satisfies $t(\mathbf{y}_j) = 1$ if and only if entry j of $\mathbf{v}$ is 1. Thus all possible classifications of the set of vectors can be realised, and the set is shattered. □

Recall that a set $\{h_1, h_2, \ldots, h_s\}$ of functions defined on a set X is *linearly dependent* if there are constants λ_i $(1 \leq i \leq s)$, not all zero, such that, *for all* $\mathbf{x} \in X$,

$$\lambda_1 h_1(\mathbf{x}) + \lambda_2 h_2(\mathbf{x}) + \ldots + \lambda_s h_s(\mathbf{x}) = 0; \tag{49}$$

that is, if some non-trivial linear combination of the functions is the zero function on X. The following result is due to Dudley [15]; we present here a new proof based on the idea of extended vectors.

Theorem 19 *Let* $\overline{\mathcal{H}}$ *be a vector space of real-valued functions defined on a set* X. *Suppose that* $\overline{\mathcal{H}}$ *has (vector space) dimension* d. *For any* $\overline{h} \in \overline{\mathcal{H}}$, *define the* $\{0,1\}$*-valued function* h *on* X *by*

$$h(\mathbf{x}) = \rho(\overline{h}(\mathbf{x})) = \begin{cases} 1 & \text{if } \overline{h}(\mathbf{x}) \geq 0 \\ 0 & \text{if } \overline{h}(\mathbf{x}) < 0, \end{cases} \tag{50}$$

and define

$$\mathcal{H} = \left\{ h : \overline{h} \in \overline{\mathcal{H}} \right\}. \tag{51}$$

Then the VC dimension of $\mathcal{H}$ *is* d.

Proof: Suppose that $\{\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_d\}$ is a basis for $\overline{\mathcal{H}}$ and, for $\mathbf{x} \in X$, let $\mathbf{x}^{\overline{\mathcal{H}}} = (\overline{h}_1(\mathbf{x}), \overline{h}_2(\mathbf{x}), \ldots, \overline{h}_d(\mathbf{x}))$. The subset S of X is shattered by $\mathcal{H}$ if and only if for each $S^+ \subseteq S$ there is $\overline{h} \in \overline{\mathcal{H}}$ such that $\overline{h}(\mathbf{x}) \geq 0$ if $\mathbf{x} \in S^+$ and $\overline{h}(\mathbf{x}) < 0$ if $\mathbf{x} \in S^- = S \setminus S^+$. But, since $\{\overline{h}_1, \ldots, \overline{h}_d\}$ is a basis of $\overline{\mathcal{H}}$, for any $\overline{h} \in \overline{H}$ there are constants w_i $(1 \leq i \leq d)$ such that $\overline{h} = \sum_{i=1}^{d} w_i \overline{h}_i$. Thus, equivalently, S is shattered by $\mathcal{H}$ if and only if for every subset S^+ of S, there are constants w_i such that

$$\sum_{i=1}^{d} w_i \overline{h}_i(\mathbf{x}) \begin{cases} \geq 0 & \text{if } \mathbf{x} \in S^+; \\ < 0 & \text{if } \mathbf{x} \in S^-; \end{cases} \tag{52}$$

that is, the inner product $\left\langle \mathbf{w}, \mathbf{x}^{\overline{\mathcal{H}}} \right\rangle$ is non-negative for $\mathbf{x} \in S^+$ and is negative for $\mathbf{x} \in S^-$. But this says precisely that the linear threshold function $f_{\mathbf{w}}$ given by

$$f_{\mathbf{w}}(\mathbf{x}) = \rho[w_1 x_1 + w_2 x_2 + \ldots w_d x_d] \tag{53}$$

satisfies

$$f_{\mathbf{w}}(\mathbf{x}^{\overline{\mathcal{H}}}) = 1 \Longleftrightarrow \mathbf{x} \in S^+. \tag{54}$$

It follows that the set S is shattered by $\mathcal{H}$ if and only if the set $\{\mathbf{x}^{\overline{\mathcal{H}}} \mid \mathbf{x} \in S\}$ is shattered by homogeneous linear threshold functions in $\mathbf{R}^d$. Because $\mathcal{V}(\mathcal{H})$ is the size of the largest set shattered by $\mathcal{H}$ and because Lemma 18 now shows that S cannot be shattered by $\mathcal{H}$ if $|S| > d$, it follows that $\mathcal{V}(\mathcal{H}) \leq d$. Further, by lemma 18, the VC dimension equals d if and only if there is a set $\left\{\mathbf{x}_1^{\overline{\mathcal{H}}}, \ldots, \mathbf{x}_d^{\overline{\mathcal{H}}}\right\}$ of linearly independent extended vectors in $\mathbf{R}^d$. Suppose this is not so. Then the vector subspace of $\mathbf{R}^d$ spanned by the set $\left\{\mathbf{x}^{\overline{\mathcal{H}}} \mid \mathbf{x} \in X\right\}$ is of dimension at most $d-1$ and therefore is contained in some hyperplane. Hence there are constants $\lambda_1, \lambda_2, \ldots, \lambda_d$, not all zero, such that for every $\mathbf{x} \in X$, $\sum_{i=1}^{d} \lambda_i (\mathbf{x}^{\overline{\mathcal{H}}})_i = 0$. But this means that for all $\mathbf{x} \in X$, $\sum_{i=1}^{d} \lambda_i \overline{h}_i(\mathbf{x}) = 0$, and hence the function $\sum_{i=1}^{d} \lambda_i \overline{h}_i$ is identically zero on X, contradicting the linear independence of $\overline{h}_1, \ldots, \overline{h}_d$. It follows that the VC dimension of $\mathcal{H}$ is d, as claimed. □

This theorem is very useful and has been mentioned already in earlier parts of this paper. In its statement we have denoted the domain of the class of functions by X. In the applications here, X will be either $\mathbf{R}^n$ or $\{0, 1\}^n$, for some n. For the moment, it is convenient to phrase the theorem and the next result in terms of general X. The theorem applies directly to linearly weighted neural networks as follows.

Theorem 20 *Let $\Phi = \{\phi_1, \ldots, \phi_m\}$ be a given set of basis functions defined on a set X and let $\mathcal{F}^\Phi$ be the set of linearly weighted neural networks on X based on Φ. If $\{1\} \cup \Phi$ is a linearly independent set in the vector space of real-valued functions on X, where 1 denotes the identically-1 function on X, then $\mathcal{V}(\mathcal{F}^\Phi) = m + 1$. In general $\mathcal{V}(\mathcal{F}^\Phi)$ is the maximum cardinality of a linearly independent subset of $\{1\} \cup \Phi$.*

Proof: Let $\overline{\mathcal{H}} = \mathrm{Sp}(1, \Phi)$ be the vector space of real functions on X spanned by the identically-1 function on X and the basis functions $\Phi = \{\phi_1, \phi_2, \ldots, \phi_m\}$. Then $\overline{\mathcal{H}}$ consists of all functions of the form

$$\overline{h}(\mathbf{x}) = w_0 + w_1\phi_1(\mathbf{x}) + \ldots + w_m\phi_m(\mathbf{x}) \tag{55}$$

for all possible choices of constants w_i. It is clear from this and equation 1 that

$$\mathcal{F}^\Phi = \mathcal{H}, \tag{56}$$

with the notation as in Theorem 19, so that the VC dimension of $\mathcal{F}^\Phi$ is the vector-space dimension of $\overline{\mathcal{H}} = \mathrm{Sp}(1, \Phi)$. The result follows. □

4.3 VC dimension of PDFs

We now apply the above results to the classes $\mathcal{P}(n, k)$ and $\mathcal{P}_{\mathbf{B}}(n, k)$. For $\mathcal{P}(n, k)$, the full class of PDFs of order at most k, the basis functions are given by $\phi_i(\mathbf{x}) = \mathbf{x}_S$ for $\emptyset \neq S \in [n]^k$. For $\mathcal{P}_{\mathbf{B}}(n, k)$, the basis functions can be taken to be $\psi_i(\mathbf{x}) = \mathbf{x}_S$ where $\emptyset \neq S \in [n]^{(k)}$. Let

$$\Phi(n, k) = \left\{\mathbf{x}_S \mid \emptyset \neq S \in [n]^k\right\}, \quad \Phi_{\mathbf{B}}(n, k) = \left\{\mathbf{x}_S \mid \emptyset \neq S \in [n]^{(k)}\right\}. \tag{57}$$

Proposition 21 *For all n and k, the set $\{1\} \cup \Phi(n, k)$ is a linearly independent set of real functions on $\mathbf{R}^n$.* □

The proof is omitted; see Anthony [2] for details.

Now consider $\Phi_{\mathbf{B}}(n, k)$, regarded as a set of real functions on domain $X = \{0, 1\}^n$.

Proposition 22 *For all n, k with $k \leq n$, $\{1, \Phi_{\mathbf{B}}(n, k)\}$ is a linearly independent set of real functions defined on $\{0, 1\}^n$.*

Proof: Let $n \geq 1$, and suppose that for some constants α_0 and α_S, for all $\mathbf{x} \in \{0,1\}^n$,

$$A(\mathbf{x}) = \alpha_0 + \sum_{\emptyset \neq S \in [n]^{(k)}} \alpha_S \mathbf{x}_S = 0. \tag{58}$$

Set $\mathbf{x}$ to be the all-0 vector to deduce that $\alpha_0 = 0$. Let $1 \leq r \leq k$ and assume, inductively, that $\alpha_S = 0$ for all $S \subseteq [n]$ with $|S| < r$. Let $S \subseteq [n]$ with $|S| = r$. Setting $x_i = 1$ if $i \in S$ and $x_j = 0$ if $j \notin S$, we deduce that $A(\mathbf{x}) = \alpha_S = 0$. Thus for all S of cardinality r, $\alpha_S = 0$. Hence $\alpha_S = 0$ for all S, and the functions are linearly independent. □

The above two results, coupled with theorem 20, enable us to determine the VC dimensions of the classes of PDFs and boolean PDFs.

Corollary 23 *For all n, k,*

$$\mathcal{V}(\mathcal{P}(n,k)) = \binom{n+k}{k}, \tag{59}$$

and for all n, k with $k \leq n$,

$$\mathcal{V}(\mathcal{P}_{\mathbf{B}}(n,k)) = \sum_{i=0}^{k} \binom{n}{i}. \tag{60}$$

Note that if all inputs are restricted to be binary and if $m > 1$, then the VC dimension of the corresponding LWNN is lower than if the inputs are allowed to be arbitrary real numbers. We remark that the VC dimensions coincide for $m = 1$, the case of linear threshold functions.

Theorem 20 tells us a little more than this. As mentioned near the beginning of this section, one may only be interested in LWNNs based on a strict subset of the basis functions $\Phi(n,k)$. For example, the special case of RBFNs in which the centers are fixed or variable and the function ϕ is of the form $\phi(r) = r^i$ for an even positive integer i, reduces essentially to PDFs based on some of the functions $\phi_i(\mathbf{x}) = \prod_{1 \leq j \leq n} x_j^{r_j}$ as in equation 43. But since the set $\{1, \Phi(n,k)\}$ is a linearly independent set for any n, k, it follows that any LWNN based on a strict subset of m of the functions in $\cup_{k \geq 1} \Phi(n,k)$ has VC dimension $m+1$. A similar comment applies to binary-input LWNNs based on strict subsets of $\Phi_{\mathbf{B}}(n,k)$ for all n and for any $k \leq n$. These observations may be summarized as follows.

Theorem 24 *Any class of PDFs based on m of the standard basis functions $\cup_{k \geq 1} \Phi(n,k)$ has VC dimension $m+1$. Any class of boolean PDFs based on m of the standard boolean PDF basis functions $\Phi_{\mathbf{B}}(n,n)$ has VC dimension $m+1$.* □

References

[1] M. Anthony, Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. To appear in *The Computational and Learning Complexity of Neural Networks*, (ed. Ian Parberry).

[2] M. Anthony, Classification by Polynomial Surfaces, LSE Mathematics Preprint Series, LSE-MPS-39, October 1992. To appear, *Discrete Applied Mathematics*.

[3] M. Anthony and N. Biggs, *Computational Learning Theory: An Introduction*, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, Cambridge, UK, 1992.

[4] M. Anthony and J. Shawe-Taylor, *A result of Vapnik with applications*, *Discrete Applied Mathematics*), 47, 1993: 207–217.

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

[6] P.L. Bartlett. Lower Bounds on the Vapnik-Chervonenkis Dimension of Multi-Layer Threshold Networks. In *Proc. 6th Annu. Workshop on Comput. Learning Theory*, pages 144–150. ACM Press, New York, NY, 1993.

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

[8] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pages 144–152, 1992.

[9] D.S. Broomhead and D. Lowe, Multivariable functional interpolation and adaptive networks, *Complex Systems*, 2, 1988: 321–355.

[10] J. Bruck, Harmonic analysis of polynomial threshold functions, *SIAM Journ. Disc. Math.*, 3 (2), 1990: 168–177.

[11] S. Chen, S.A. Billings and P.M. Grant. Recursive hybrid algorithm for non-linear system identification using radial basis function networks. *International Journal of Control*, 55(5), 1992: 1051–1070.

[12] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Computers* 14, 1965: 326–334.

[13] L. Devroye, Automatic pattern recognition: a study of the probability of error, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(4), 1988: 530–543.

[14] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, 1973.

[15] R.M. Dudley, Central limit theorems for empirical measures, *Ann. Probability* 6, 1978: 899–929.

[16] Nira Dyn and Charles A. Micchelli, Interpolation by sums of radial functions, *Numerische Mathematik* 58, 1990: 1–9.

[17] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

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

[19] S.B. Holden, Neural networks and the VC dimension, to appear in the Proceedings of the Third IMA Conference on Mathematics in Signal Processing, December 1992.

[20] S.B. Holden, On the Theory of Generalization and Self-Structuring in Linearly Weighted Connectionist Networks. PhD dissertation, Technical report CUED/F-INFENG/TR.161 (1994), Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England, 1994.

[21] S.B. Holden and P.J.W. Rayner, Generalization and PAC learning: Some new results for the class of generalized single layer networks. To appear in IEEE Transactions on Neural Networks.

[22] U. Kreßel, J. Franke, and J. Schurmann. Polynomial classifier versus multilayer perceptron. Unpublished manuscript. Daimler-Benz AG, Reseach Center Ulm, Wilhelm-Runge-Str. 11, 7900 Ulm, F.R.Germany.

[23] Wee Sun Lee, Peter L. Bartlett and Robert C. Williamson, Lower Bounds on the VC-Dimension of Smoothly Parametrized Function Classes. Unpublished manuscript.

[24] D. Lowe. Adaptive radial basis function nonlinearities and the problem of generalization. In *Proceedings of the First IEE International Conference on Artificial Neural Networks*, pages 171–175, 1989.

[25] W. Maass. Bounds on the computational power and learning complexity of analog neural nets (extended abstract). In *Proceedings of 25th Annual ACM Symposium on the Theory of Computing*, pages 335–344. ACM Press, 1993.

[26] W. Maass. Neural nets with superlinear VC-dimension. *Neural Computation*. To appear.

[27] C.A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2, 1986: 11–22.

[28] J. Moody and C.J. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1, 1989: 281–294.

[29] N.J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.

[30] M. Niranjan and F. Fallside, Neural networks and radial basis functions in classifying static speech patterns. Technical Report CUED/F-INFENG/TR 22 (1988), Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England, 1988.

[31] R. Paturi and M. Saks, On threshold circuits for parity, In *Proceedings of 31st IEEE Symposium on Foundations of Computer Science, 1990.*

[32] T. Poggio and F. Girosi, Networks for approximation and learning, *Proceedings of the IEEE*, 78(9), 1990: 1481–1497.

[33] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[34] M.J.D. Powell, Radial basis functions for multivariable interpolation: A review. Technical Report #DAMTP 1985/NA12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, October 1985.

[35] M.J.D.Powell, The Theory of Radial Basis Function Approximation in 1990. Technical report #DAMTP 1990/NA11, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, December 1990.

[36] R.W. Prager and F. Fallside, The modified Kanerva model for automatic speech recognition, *Computer Speech and Language*, 3, 1989: 61–81.

[37] P.J.W. Rayner and M.R. Lynch, A new connectionist model based on a non-linear adaptive filter. In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, 1989.*

[38] S. Renals and R. Rohwer. Phoneme classification experiments using radial basis functions. In *Proceedings of the International Joint Conference on Neural Networks*, pages I-461–I-467, June 1989.

[39] M. Saks, Slicing the hypercube. In *Surveys in Combinatorics, 1993*, a volume of invited talks at the 1993 British Combinatorial Conference, Cambridge University Press, 1993.

[40] N. Sauer, On the density of families of sets, *Journal of Combinatorial Theory (A)*, 13, 1972: 145–147.

[41] L. G. Valiant, A theory of the learnable, *Communications of the ACM*, 27(11), 1984: 1134–1142.

[42] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

[43] V.N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 1971: 264-280.

[44] C. Wang and A.C. Williams, The threshold order of a boolean function, *Discrete Applied Mathematics*, 31, 1991: 51-69.

[45] R.S. Wenocur and R.M. Dudley, Some special Vapnik-Chervonenkis classes, *Discrete Mathematics*, 33, 1981: 313–318.