On the Capabilities of Higher-Order Neurons: A Radial Basis Function Approach

Michael Schmitt

Lehrstuhl Mathematik und Informatik, Fakultät für Mathematik
Ruhr-Universität Bochum, D–44780 Bochum, Germany
http://www.ruhr-uni-bochum.de/lmi/mschmitt/
mschmitt@lmi.ruhr-uni-bochum.de

Abstract

Higher-order neurons with k monomials in n variables are shown to have Vapnik-Chervonenkis (VC) dimension at least $nk + 1$. This result supersedes the previously known lower bound obtained via k-term monotone disjunctive normal form (DNF) formulas. Moreover, it implies that the VC dimension of higher-order neurons with k monomials is strictly larger than the VC dimension of k-term monotone DNF. The result is achieved by introducing an exponential approach that employs Gaussian radial basis function (RBF) neural networks for obtaining classifications of points in terms of higher-order neurons.

1 Introduction

Higher-order neurons are simple but powerful extensions of linear neuron models. They introduce the concept of nonlinearity by incorporating monomials, that is, products of input variables, as an intermediate step—effectively a hidden layer—of computation. A common way of specifying a higher-order neuron is to determine the number of input variables and the number of monomials, but leaving the order of the monomials open. Thus, one obtains a powerful neuron model that is capable of computing with unlimited degrees of nonlinearity.

The computational capabilities that arise from this model, however, are far from being completely understood. One of the well-studied theoretical notions that are used to characterize the computational richness of neural networks is the Vapnik-Chervonenkis (VC) dimension. More generally, the VC dimension of a function class quantifies its classification capabilities (Vapnik and Chervonenkis, 1971): It indicates the cardinality of the largest set for which all possible binary-valued classifications can be obtained using functions from the class. The VC dimension is also well founded as a measure for the sample complexity of learning (see, e.g., Anthony and Bartlett, 1999): It yields estimates for the number of examples needed by a learning algorithm to output functions that have low generalization errors.

We establish here a new lower bound on the VC dimension of higher-order neurons: We show that the higher-order neuron with k monomials in n variables has VC dimension at least $nk + 1$. The largest lower bound that has been known previously is derived from the lower bound for Boolean formulas in k-term monotone disjunctive normal form (DNF), that is, disjunctions of at most k monomials without negations. This bound has been obtained by Littlestone (1988). In particular, Littlestone has shown that the class of k-term monotone l-DNF formulas (i.e., with monomials containing at most l variables) has VC dimension at least $lk\lfloor\log(n/m)\rfloor$, where $l \leq m \leq n$, and $k \leq \binom{m}{l}$. Using $l = n/4$ and $m = n/2$, for instance, this yields the lower bound $nk/4$ for the VC dimension of k-term monotone DNF and, hence, of higher-order neurons with k monomials, where k has to satisfy the given constraints.

The new bound that we provide here for higher-order neurons supersedes this previous bound in a threefold way: First, it improves the bound from k-term monotone DNF formulas in value. Second, it releases k from the constraints through n in that the new bound holds for every n and k—in particular, for values of k that are larger than the number of monotone monomials. Finally, $nk + 1$ is even larger than the VC dimension of the class of k-term monotone DNF formulas itself: We show that the difference between both dimensions is larger than $k\log(k/e) + 1$.

So far, a considerable number of results and techniques for VC dimension bounds have been provided in the context of real valued function classes (see, e.g., Bartlett and Maass, 2003, and the references there). For specific subclasses that can be computed by higher-order neurons, tight bounds have been calculated: Karpinski and Werther (1993) have shown that univariate polynomials with at most k terms have a VC dimension[1] proportional to k. Further, the VC dimension of the class of monomials over the reals is equal to n (see Ehrenfeucht et al., 1989; Schmitt, 2002c, for lower and upper bound, respectively). There is also a VC dimension result known for n-variate d-degree polynomials (see, e.g., Ben-David and Lindenbaum, 1998): This class has VC dimension equal to $\binom{n+d}{d}$. However, as the class contains polynomials consisting of $\binom{n+d}{d}$ terms and the bound k on the number of monomials restricts the number of variables in terms of k, this result entails for higher-order neurons (without a constraint on the degree) a

[1]Strictly speaking, Karpinski and Werther (1993) studied a related notion, the so-called pseudo-dimension. Following their methods, it is not hard to obtain this result for the VC dimension (see also Schmitt, 2002a).

lower bound not better than the bound due to Littlestone (1988).

There has been previous work that established techniques for deriving lower bounds on the VC dimension for quite general types of real-valued function classes. Building on results by Lee et al. (1995), Erlich et al. (1997) provide powerful means for obtaining lower bounds for parameterized function classes[2]. An essential requirement for using these techniques, however, is that the function class is "smoothly" parameterized, a fact that does not apply to the exponents of polynomials. The lower bound method of Koiran and Sontag (1997) for various types of neural networks, generalized by Bartlett et al. (1998) to neural networks with a given number of layers, cannot be employed for higher-order neurons either. This technique is constrained to networks where each neuron computes a function with finite limits at infinity, a property monomials do not have. Further, Koiran and Sontag (1997) designed a lower bound method for networks consisting of linear and multiplication gates. However, the way these networks are constructed—with layers consisting of products of linear terms[3]—does not give rise to higher-order neurons, even when the number of layers is restricted.

We provide a completely new approach to the derivation of lower bounds on the VC dimension of higher-order neurons. First, we establish the lower bound $nk + 1$ on the VC dimension of a specific type of radial basis function (RBF) neural network (see, e.g., Haykin, 1999). The networks considered here have k Gaussian[4] units as computational elements and satisfy certain assumptions with respect to the input domain and the values taken by the parameters. The bound for these networks improves a result of Erlich et al. (1997) in combination with Lee et al. (1995) who established the lower bound $n(k-1)$ for RBF networks[5] with restrictions neither on inputs nor on parameters. Then we use our result for RBF networks to obtain the lower bound on the VC dimension of higher-order neurons.

[2]A parameterized function class is given in terms of a function having two types of variables: input variables and parameter variables. The function class is obtained by instantiating the parameter variables with, in general, real numbers. Neural networks are prominent examples for parameterized function classes.

[3]Precisely, such a layer uses products of the form $\prod_{i=1}^{l}(x - a_i)$ where it is crucial that there is no bound on l.

[4]named after the German mathematician Carl Friedrich Gauss (1777–1855)

[5]These results and the one presented here concern RBF networks with uniform width, that is, where all units have equal width. This constraint is not a fundamental restriction since, as shown by Park and Sandberg (1991), these networks still have universal approximation capabilities. As far as the VC dimension is concerned, however, better lower bounds are known for more general types of RBF networks (Schmitt, 2002b).

Thus, RBF networks open a new way to assess the classification capabilities of higher-order neurons. This Gaussian RBF approach has also proven to be helpful in a different context dealing with the roots of univariate polynomials (Schmitt, 2004).

Higher-order neurons are a special case of a particular type of neural networks, the so-called product unit neural networks (Durbin and Rumelhart, 1989). It immediately follows from the bound for higher-order neurons established here that the VC dimension of product unit neural networks with n input nodes and one layer of k hidden nodes (that is, nodes that are neither input nor output nodes) is at least $nk + 1$.

Concerning upper bounds for the VC dimension of higher-order neurons there are two relevant results: The bound $O(n^2k^4)$ due to Karpinski and Macintyre (1997) is the smallest bound known for higher-order neurons with unlimited degree (see also Schmitt, 2002c). The higher-order neuron with n variables, k monomials, and degree no larger than d has VC dimension no more than $2nk \log(9d)$ (Schmitt, 2002c). The derivation of the new lower bound not only narrows the gap between upper and lower bounds, but gives also rise to subclasses of degree-restricted higher-order neurons for which the bound is optimal up to the factor $2 \log(9d)$.

We introduce definitions and notation in Section 2. Section 3 provides geometric constructions that are required for the derivations of the main results presented in Section 4. Finally, in Section 5, we compare the new bound with the upper bound for k-term monotone DNF formulas and show that the new bound exceeds the VC dimension of this class.

2 Definitions

A *higher-order neuron* (or *sigma-pi unit*) *with k monomials in n variables* computes the functions

$$a_0 + a_1 x_1^{b_{1,1}} \cdots x_n^{b_{1,n}} + \cdots + a_k x_1^{b_{k,1}} \cdots x_n^{b_{k,n}} \tag{1}$$

with real coefficients $a_0, \ldots, a_k$ (the *output weights* of the neuron with *bias* a_0) and nonnegative integer exponents $b_{1,1}, \ldots, b_{k,n}$. The coefficients and the exponents are the adjustable parameters of the neuron. For given n and k, the function class constituted by all functions in (1) is also known as the class of k-sparse n-variate polynomials. If the exponents $b_{1,1}, \ldots, b_{k,n}$ are allowed to be arbitrary real numbers, each monomial becomes a *product unit* and we obtain a *product unit neural network* with one hidden layer of k product units.

We use bold symbols to indicate vectors, such as $\boldsymbol{x}$ and $\boldsymbol{c}_i$; non-bold symbols are reserved for scalar variables. Let $\|\cdot\|$ denote the Euclidean norm. A *radial basis function neural network* (*RBF network*, for short) with n input nodes computes functions that map from $\mathbb{R}^n$ to $\mathbb{R}$ and can be written as

$$w_0 + w_1 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_1\|^2}{\sigma^2}\right) + \cdots + w_k \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_k\|^2}{\sigma^2}\right),$$

where k is the number of RBF units. This particular type of network is also known as a *Gaussian RBF network.* Each exponential term corresponds to the function computed by a Gaussian RBF unit with *center* $\boldsymbol{c}_i \in \mathbb{R}^n$, where n is the number of variables, and *width* $\sigma \in \mathbb{R} \setminus \{0\}$. The width is a network parameter that we assume here to be equal for all units, that is, we consider RBF networks with uniform width. Further, $w_0, \ldots, w_k$ are the *output weights* and w_0 is also referred to as the *bias* of the network.

The *Vapnik-Chervonenkis (VC) dimension* of a class $\mathcal{F}$ of real-valued functions is defined via the notion of shattering: A set $S \subseteq \mathbb{R}^n$ is said to be *shattered* by $\mathcal{F}$ if every dichotomy of S is induced by $\mathcal{F}$, that is, if for every pair (S^-, S^+), where $S^- \cap S^+ = \emptyset$ and $S^- \cup S^+ = S$, there is some function $f \in \mathcal{F}$ such that

$$\mathrm{sgn} \circ f(S^-) \subseteq \{0\} \quad \text{and} \quad \mathrm{sgn} \circ f(S^+) \subseteq \{1\}.$$

Here $\mathrm{sgn} : \mathbb{R} \to \{0, 1\}$ denotes the sign function, satisfying $\mathrm{sgn}(x) = 1$ if $x \geq 0$, and $\mathrm{sgn}(x) = 0$ otherwise. The VC dimension of $\mathcal{F}$ is then defined as the cardinality of the largest set shattered by $\mathcal{F}$. (It is said to be infinite if there is no such set.) The VC dimension of a neuron or a neural network is equated with the VC dimension of the class of functions computed by the neuron or the neural network, respectively.

Finally, we make use of the geometric notions of a ball and a hypersphere. A *ball* in $\mathbb{R}^n$ is given in terms of a center $\boldsymbol{c} \in \mathbb{R}^n$ and a radius $\rho \in \mathbb{R}$ as the set

$$B(\boldsymbol{c}, \rho) \quad = \quad \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| \leq \rho\}.$$

A *hypersphere* is the set of points on the surface of a ball, that is,

$$S(\boldsymbol{c}, \rho) \quad = \quad \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| = \rho\}.$$

3 Ancillary Constructions

In the following we provide the geometric constructions that are the basis for the main result in Section 4. The idea pursued here is to represent classifications of sets using unions of balls, where a point is classified as positive if and only if it is contained in some ball. In order for being shattered, the sets are chosen to satisfy a certain condition of independence with respect to the positions of their elements: The points are required to lie on hyperspheres such that each hypersphere is maximally determined by the set of points. In other words, removing any point increases the set of possible hyperspheres that contain the reduced set. The following definition makes this notion of independence precise.

Definition. *A set $Q \subseteq \mathbb{R}^n$ of at most $n+1$ points is* in general position for hyperspheres *if the system of equalities*

$$\|\boldsymbol{p} - \boldsymbol{c}\| \;=\; \eta, \qquad \textit{for all } \boldsymbol{p} \in Q, \tag{2}$$

in the variables $\boldsymbol{c} = (c_1, \ldots, c_n)$ and η has a solution and, for every $\boldsymbol{q} \in Q$, the solution set is a proper subset of the solution set of the system

$$\|\boldsymbol{p} - \boldsymbol{c}\| \;=\; \eta, \qquad \textit{for all } \boldsymbol{p} \in Q \setminus \{\boldsymbol{q}\}. \tag{3}$$

Given a set of points that satisfies this definition and lies on a hypersphere, we next want to find a ball such that one of the points lies outside of the ball while the other points are on its surface. We show that this can be done, provided that the set is in general position for hyperspheres. Moreover, the ball can be chosen with the center and the radius as close as possible to the center and the radius of the hypersphere that contains all points.

Lemma 1. *Suppose that $Q \subseteq \mathbb{R}^n$ is a set of at most $n+1$ points in general position for hyperspheres and let $\boldsymbol{q} \in Q$. Further, let $\boldsymbol{c} \in \mathbb{R}^n, \eta \in \mathbb{R}$ be a solution of the system*

$$\|\boldsymbol{p} - \boldsymbol{c}\| \;=\; \eta, \qquad \textit{for all } \boldsymbol{p} \in Q. \tag{4}$$

Then, for every $\varepsilon > 0$, there exists a solution $\boldsymbol{c}(\varepsilon) \in \mathbb{R}^n, \eta(\varepsilon) \in \mathbb{R}$ of the system

$$\begin{aligned} \|\boldsymbol{p} - \boldsymbol{c}(\varepsilon)\| &= \eta(\varepsilon), \qquad \textit{for all } \boldsymbol{p} \in Q \setminus \{\boldsymbol{q}\}, \\ \|\boldsymbol{q} - \boldsymbol{c}(\varepsilon)\| &> \eta(\varepsilon) \end{aligned}$$

satisfying

$$\|\boldsymbol{c} - \boldsymbol{c}(\varepsilon)\| < \varepsilon \quad \textit{and} \quad |\eta - \eta(\varepsilon)| < \varepsilon.$$

Proof. Without loss of generality, we may assume that $\eta > 0$. (If $\eta = 0$ then we have $|Q| = 1$, and the statement is trivial.) Since $\boldsymbol{c}$ and η solve the system (4), $\boldsymbol{c}$ and $\vartheta = \eta^2 - \|\boldsymbol{c}\|^2$ are a solution of the system

$$\|\boldsymbol{p}\|^2 - 2\boldsymbol{p}\boldsymbol{c} \;=\; \vartheta, \qquad \text{for all } \boldsymbol{p} \in Q. \tag{5}$$

Because Q is in general position for hyperspheres, the solution set of the system (5) is a proper subset of the solution set of the system

$$\|\boldsymbol{p}\|^2 - 2\boldsymbol{p}\boldsymbol{c} \;=\; \vartheta, \qquad \text{for all } \boldsymbol{p} \in Q \setminus \{\boldsymbol{q}\}. \tag{6}$$

According to facts from linear algebra, there exist $\boldsymbol{a} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that for every $\lambda \neq 0$, we have with $\boldsymbol{c} + \lambda\boldsymbol{a}$ and $\vartheta + \lambda\alpha$ a solution of the system (6) that does not solve the system (5). For a given $\varepsilon > 0$, choose $\lambda(\varepsilon) \in \mathbb{R} \setminus \{0\}$ such that $|\lambda(\varepsilon)|$ is sufficiently small to satisfy

$$\|\lambda(\varepsilon)\boldsymbol{a}\| < \varepsilon \quad \text{and} \quad \left|\sqrt{\vartheta + \|\boldsymbol{c}\|^2} - \sqrt{\vartheta + \lambda(\varepsilon)\alpha + \|\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}\|^2}\,\right| < \varepsilon. \quad (7)$$

It is obvious that the second inequality can be met due to the fact that the equation $\sqrt{\vartheta + \|\boldsymbol{c}\|^2} = \eta$ holds, which we get from the definition of ϑ, and the assumption $\eta > 0$. Since $\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}$ and $\vartheta + \lambda(\varepsilon)\alpha$ solve (6) but not (5), it follows that

$$\|\boldsymbol{q}\|^2 - 2\boldsymbol{q}(\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}) \;\neq\; \vartheta + \lambda(\varepsilon)\alpha,$$

which, using $\|\boldsymbol{q}\|^2 - 2\boldsymbol{q}\boldsymbol{c} = \vartheta$ from (5), is equivalent to

$$-2\lambda(\varepsilon)\boldsymbol{q}\boldsymbol{a} \;\neq\; \lambda(\varepsilon)\alpha.$$

Due to this inequality, we can choose the (not yet specified) sign of $\lambda(\varepsilon)$ such that

$$-2\lambda(\varepsilon)\boldsymbol{q}\boldsymbol{a} \;>\; \lambda(\varepsilon)\alpha.$$

Again with $\|\boldsymbol{q}\|^2 - 2\boldsymbol{q}\boldsymbol{c} = \vartheta$, it follows that

$$\|\boldsymbol{q}\|^2 - 2\boldsymbol{q}(\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}) \;>\; \vartheta + \lambda(\varepsilon)\alpha,$$

and, therefore,

$$\|\boldsymbol{q} - (\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a})\|^2 \;>\; \vartheta + \lambda(\varepsilon)\alpha + \|\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}\|^2.$$

Hence, defining

$$\boldsymbol{c}(\varepsilon) = \boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a} \quad \text{and} \quad \eta(\varepsilon) = \sqrt{\vartheta + \lambda(\varepsilon)\alpha + \|\boldsymbol{c} + \lambda(\varepsilon)\boldsymbol{a}\|^2},$$

we obtain $\|\boldsymbol{q} - \boldsymbol{c}(\varepsilon)\| > \eta(\varepsilon)$. Furthermore, the inequalities (7) imply that $\|\boldsymbol{c} - \boldsymbol{c}(\varepsilon)\| < \varepsilon$ and $|\eta - \eta(\varepsilon)| < \varepsilon$ hold as claimed. $\square$

We now apply the previous result to show that any dichotomy of a given set of points can be obtained using balls. As the set may generally be a subset of some larger set, we also ensure that the balls do not enclose any additional point. Further, we guarantee that this can be done with all centers remaining positive, a condition that will turn out to be useful in the following section. We say here that a vector is positive, if all its components are larger than zero.

Lemma 2. *Let $Q \subseteq \mathbb{R}^n$ be a set of n points in general position for hyperspheres and let $P \subseteq \mathbb{R}^n$ be a finite set with $Q \subseteq P$. Assume further that there exists a positive center $\boldsymbol{c} \in \mathbb{R}^n$ and a radius $\eta \in \mathbb{R}$ such that*

$$\begin{aligned} Q &\subseteq S(\boldsymbol{c}, \eta), \\ P \cap B(\boldsymbol{c}, \eta) &= Q. \end{aligned}$$

Then for every $R \subseteq Q$ there exists a positive center $\boldsymbol{d} \in \mathbb{R}^n$ and a radius $\zeta \in \mathbb{R}$ such that

$$\begin{aligned} R &\subseteq S(\boldsymbol{d}, \zeta), \\ P \cap B(\boldsymbol{d}, \zeta) &= R. \end{aligned}$$

Proof. Clearly, it is sufficient to consider sets R that are proper subsets of Q. Without loss of generality, we may assume that $|R| = |Q| - 1$. The general case then follows inductively. Suppose that $\boldsymbol{q} \in Q$ and let $R = Q \setminus \{\boldsymbol{q}\}$. According to Lemma 1, for every $\varepsilon > 0$ there exist $\boldsymbol{c}(\varepsilon), \eta(\varepsilon)$ satisfying

$$\|\boldsymbol{p} - \boldsymbol{c}(\varepsilon)\| = \eta(\varepsilon), \qquad \text{for all } \boldsymbol{p} \in Q \setminus \{\boldsymbol{q}\}, \tag{8}$$

$$\|\boldsymbol{q} - \boldsymbol{c}(\varepsilon)\| > \eta(\varepsilon), \tag{9}$$

$$\|\boldsymbol{c} - \boldsymbol{c}(\varepsilon)\| < \varepsilon, \tag{10}$$

$$|\eta - \eta(\varepsilon)| < \varepsilon. \tag{11}$$

Obviously, property (8) implies that $R \subseteq S(\boldsymbol{c}(\varepsilon), \eta(\varepsilon))$. Property (9) states that $\boldsymbol{q} \notin B(\boldsymbol{c}(\varepsilon), \eta(\varepsilon))$. Since the assumption $P \cap B(\boldsymbol{c}, \eta) = Q$ implies that for every $\boldsymbol{p'} \in P \setminus Q$ the constraint

$$\|\boldsymbol{p'} - \boldsymbol{c}\| > \eta$$

holds, properties (10) and (11) entail the condition

$$\|\boldsymbol{p'} - \boldsymbol{c}(\varepsilon)\| > \eta(\varepsilon)$$

for all sufficiently small ε. Thus, for any such ε we get the assertion $P \cap B(\boldsymbol{c}(\varepsilon), \eta(\varepsilon)) = R$. Further, as $\boldsymbol{c}$ is positive, property (10) ensures that $\boldsymbol{c}(\varepsilon)$ is positive for some sufficiently small ε. Hence, the claim follows for $\boldsymbol{d} = \boldsymbol{c}(\varepsilon)$, $\zeta = \eta(\varepsilon)$. □

4 VC Dimension Bound for Higher-Order Neurons

Before getting to the main result, we derive the lower bound $nk + 1$ for the VC dimension of a restricted type of RBF network. For more general RBF networks, results of Erlich et al. (1997) and Lee et al. (1995) yield the lower bound $n(k-1)$. The following theorem is stronger not only in the value of the bound, but also in the assumptions that hold: The points of the shattered set all have the same distance from the origin, the centers of the RBF units are rational numbers, and the width can be chosen arbitrarily small.

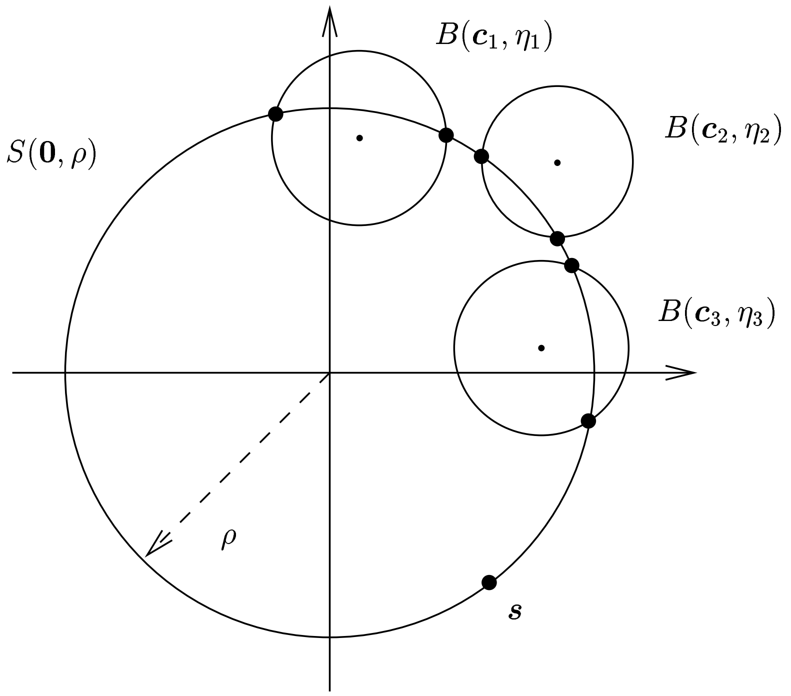

Figure 1: The points of the shattered set are chosen from the intersections of the hypersphere $S(\mathbf{0}, \rho)$ with the surfaces of pairwise disjoint balls $B(\boldsymbol{c}_i, \eta_i)$. All balls have their centers in the positive orthant. There is one additional point $\boldsymbol{s}$ not contained in any of the balls.

Theorem 3. *Let $n \geq 2$, $k \geq 1$, and $\rho > 0$ be given. There exists a set $P \subseteq S(\mathbf{0}, \rho) \subseteq \mathbb{R}^n$ of $nk + 1$ points and a real number $\sigma_0 > 0$ so that P is shattered by the RBF network with k hidden units, positive rational centers, and any width $0 < \sigma \leq \sigma_0$.*

Proof. Suppose that $B(\boldsymbol{c}_1, \eta_1), \ldots, B(\boldsymbol{c}_k, \eta_k)$ are pairwise disjoint balls with positive centers $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k \in \mathbb{R}^n$ such that, for $i = 1, \ldots, k$, the intersection $S(\boldsymbol{c}_i, \eta_i) \cap S(\mathbf{0}, \rho)$ is non-empty and not a single point. (An example for $n = 2$ and $k = 3$ is shown in Fig. 1.) For $i = 1, \ldots, k$, let $P_i \subseteq S(\boldsymbol{c}_i, \eta_i) \cap S(\mathbf{0}, \rho)$ be a set of n points in general position for hyperspheres. (Note that P_i is constrained to lie on two different hyperspheres. This still allows to choose P_i in general position since P_i contains n (and not $n + 1$) points, so that the set of possible centers for P_i yields a line.) Further, let $\boldsymbol{s} \in S(\mathbf{0}, \rho)$ be some point such that $\boldsymbol{s} \notin B(\boldsymbol{c}_i, \eta_i)$, for $i = 1, \ldots, k$. We claim that the set $P = \{\boldsymbol{s}\} \cup P_1 \cup \cdots \cup P_k$, which has $nk + 1$ points, is shattered by the RBF network with the postulated restrictions on the parameters.

Assume that (P^-, P^+) is some arbitrary dichotomy of P where $\boldsymbol{s} \in P^-$. (We will argue at the end of the proof that the complementary case can be treated by reversing signs.) Let (P_i^-, P_i^+) denote the dichotomy induced on P_i. By construction, every P_i satisfies

$$P_i \subseteq S(\boldsymbol{c}_i, \eta_i) \quad \text{and} \quad P \cap B(\boldsymbol{c}_i, \eta_i) = P_i.$$

Hence by Lemma 2, instantiating the set Q with P_i and the set R with P_i^+, it follows that there exist positive centers $\boldsymbol{d}_i$ and radii ζ_i such that

$$P_i^+ \subseteq S(\boldsymbol{d}_i, \zeta_i) \quad \text{and} \quad P \cap B(\boldsymbol{d}_i, \zeta_i) = P_i^+,$$

for $i = 1, \ldots, k$. Moreover, the centers $\boldsymbol{d}_i$ can be replaced by rational centers $\tilde{\boldsymbol{d}}_i$ that are sufficiently close to $\boldsymbol{d}_i$, such that every point of P lying outside the ball $B(\boldsymbol{d}_i, \zeta_i)$ is outside the ball $B(\tilde{\boldsymbol{d}}_i, \tilde{\zeta}_i)$ for some $\tilde{\zeta}_i \in \mathbb{R}$ close to ζ_i, and every point of P lying on the hypersphere $S(\boldsymbol{d}_i, \zeta_i)$ is contained in the ball $B(\tilde{\boldsymbol{d}}_i, \tilde{\zeta}_i)$. Thus, every $\boldsymbol{p} \in P$ satisfies

$$\boldsymbol{p} \in B(\tilde{\boldsymbol{d}}_i, \tilde{\zeta}_i) \quad \text{if and only if} \quad \boldsymbol{p} \in P_i^+, \tag{12}$$

for $i = 1, \ldots, k$. Clearly, since the centers $\boldsymbol{d}_i$ are positive, the rational centers $\tilde{\boldsymbol{d}}_i$ can be chosen to be positive as well.

The parameters of the RBF network are specified as follows: The i-th unit is associated with the ball $B(\tilde{\boldsymbol{d}}_i, \tilde{\zeta}_i)$. Assigned to it is $\tilde{\boldsymbol{d}}_i$ as the center and as output weight the value $\exp(\tilde{\zeta}_i^{\,2}/\sigma^2)$ (where σ will determined below) so that the unit contributes the term

$$\exp\left(\frac{\tilde{\zeta}_i^2}{\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{d}}_i\|^2}{\sigma^2}\right)$$

to the computation of the network. From assertion (12) we obtain that every $\boldsymbol{p} \in P \setminus P_i^+$ satisfies the constraint

$$\|\boldsymbol{p} - \tilde{\boldsymbol{d}}_i\| \;>\; \tilde{\zeta}_i.$$

Thus, for every sufficiently small $\sigma > 0$ and every $\boldsymbol{p} \in P \setminus P_i^+$, we achieve that

$$\exp\left(-\frac{\|\boldsymbol{p} - \tilde{\boldsymbol{d}}_i\|^2 - \tilde{\zeta}_i^2}{\sigma^2}\right) \;<\; \frac{1}{k} \tag{13}$$

is valid for $i = 1, \ldots, k$. On the other hand, for every $\boldsymbol{p} \in P_i^+$ condition (12) implies

$$\|\boldsymbol{p} - \tilde{\boldsymbol{d}}_i\| \;\leq\; \tilde{\zeta}_i,$$

which entails

$$\exp\left(-\frac{\|\boldsymbol{p}-\tilde{\boldsymbol{d}}_i\|^2-\tilde{\zeta}_i^2}{\sigma^2}\right) \;\geq\; 1 \tag{14}$$

for every $\sigma > 0$. Finally, we set the bias term equal to -1. It is now easy to see that the dichotomy (P^-, P^+) is induced by the parameter settings: If $\boldsymbol{p} \in P^-$ then, according to inequality (13), the weighted output values of the units and the bias sum up to a negative value. In the case $\boldsymbol{p} \in P^+$ we have $\boldsymbol{p} \in P_i^+$ for some i and, by inequality (14), the weighted unit i outputs value of at least 1, while the other units output positive values, so that the total network output is positive.

The construction for the case that classifies $\boldsymbol{s}$ as positive works similarly. We invoke Lemma 2 substituting P_i^- for R and derive the analogous version of assertion (12) with P_i^+ replaced by P_i^-. Then it is obvious that, if the weights defined above are equipped with negative signs and 1 is used as the bias, the network induces the dichotomy as claimed.

We observe that σ may have been chosen such that it depends on the particular dichotomy. To complete the proof, we require σ_0 to be small enough so that inequality (13) holds for $\sigma \leq \sigma_0$ on all points and dichotomies of P. ☐

We remark that one assumption of the theorem can be slightly weakened: It is not necessary to require that $\boldsymbol{s} \in S(\mathbf{0}, \rho)$. Instead, every point not contained in any of the balls $B(\boldsymbol{c}_i, \eta_i)$ can be selected for $\boldsymbol{s}$. However, the restriction is required for the application of the theorem in the following result, which is the main contribution of this article. For its proof we recall the definition of a product unit neural network in Section 2.

Theorem 4. *For every $n, k \geq 1$, the higher-order neuron with k monomials in n variables has VC dimension at least $nk + 1$.*

Proof. We first consider the case $n \geq 2$. By Theorem 3, for $\rho > 0$ let $P \subseteq \mathbb{R}^n$, $P \subseteq S(\mathbf{0}, \rho)$, be the set of cardinality $nk+1$ that is shattered by the RBF network with k hidden units and the stated parameter settings. We show that P can be transformed into a set P' that is shattered by the higher-order neuron with k monomials. The weighted output computed by unit i in the RBF network on input $\boldsymbol{p} \in P$ can be written as

$$\begin{aligned}
w_i \cdot \exp\left(-\frac{\|\boldsymbol{p}-\boldsymbol{c}_i\|^2}{\sigma^2}\right) \;&=\; w_i \cdot \exp\left(-\frac{\|\boldsymbol{p}\|^2 - 2\boldsymbol{p}\boldsymbol{c}_i + \|\boldsymbol{c}_i\|^2}{\sigma^2}\right) \\
&=\; w_i \cdot \exp\left(-\frac{\|\boldsymbol{p}\|^2 + \|\boldsymbol{c}_i\|^2}{\sigma^2}\right) \exp\left(\frac{2\boldsymbol{p}\boldsymbol{c}_i}{\sigma^2}\right) \\
&=\; w_i \cdot \exp\left(-\frac{\rho^2 + \|\boldsymbol{c}_i\|^2}{\sigma^2}\right) \cdot \exp\left(\frac{2p_1 c_{i,1}}{\sigma^2}\right) \cdots \exp\left(\frac{2p_n c_{i,n}}{\sigma^2}\right),
\end{aligned}$$

where we have used the assumption $P \subseteq S(\mathbf{0}, \rho)$ for the last equation and $p_j, c_{i,j}$ to denote the j-th components of the vectors $\boldsymbol{p}, \boldsymbol{c}_i$, respectively. Consider a product unit network with one hidden layer, where unit i has output weight

$$w_i' = w_i \cdot \exp\left(-\frac{\rho^2 + \|\boldsymbol{c}_i\|^2}{\sigma^2}\right)$$

and exponents $2c_{i,j}/\sigma^2$ for $j = 1, \ldots, n$. On inputs from the set

$$P' = \{(e^{p_1}, \ldots, e^{p_n}) : (p_1, \ldots, p_n) \in P\}$$

this product unit network computes the same values as the RBF network on P. Moreover, the exponents of the product units are positive rationals. According to Theorem 3, for some σ_0, any width $0 < \sigma \leq \sigma_0$ can be used. Therefore, we may choose $\sigma^2 = 1/l$ for some natural number l that is sufficiently large and a common multiple of all denominators occurring in any $c_{i,j}$, so that the exponents become integers. With these parameter settings, we have a higher-order neuron with k monomials that computes on P' the same output values as the RBF network on P. As this can be done for every dichotomy of P, it follows that P' is shattered by the higher-order neuron with k monomials.

For the case $n = 1$, we again use the RBF technique and ideas from Schmitt (2002a, 2004). Clearly, the set $M = \{0, \ldots, k\}$ can be shattered by an RBF network with $k + 1$ hidden units and zero bias: For each $i \in M$ we employ an RBF unit with center i; given a dichotomy (M^-, M^+), we let the output weight for unit i be -1 if $i \in M^-$, and 1 if $i \in M^+$. If the width σ is small enough, the output value of the network has the requested sign on every input $i \in M$. Now, let σ be the smallest width sufficient for all dichotomies of M. Then

$$w_0 \exp\left(-\frac{x^2}{\sigma^2}\right) + w_1 \exp\left(-\frac{(x-1)^2}{\sigma^2}\right) + \cdots$$
$$\cdots + w_k \exp\left(-\frac{(x-k)^2}{\sigma^2}\right) \geq 0$$

is, by multiplication with $\exp(x^2/\sigma^2)$, equivalent to

$$w_0 + w_1 \exp\left(\frac{2x-1}{\sigma^2}\right) + \cdots + w_k \exp\left(\frac{2kx - k^2}{\sigma^2}\right) \geq 0.$$

The latter can be written as

$$w_0 + w_1 \exp\left(-\frac{1}{\sigma^2}\right) \exp\left(\frac{2x}{\sigma^2}\right) + \cdots$$
$$\cdots + w_k \exp\left(-\frac{k^2}{\sigma^2}\right) \exp\left(\frac{2kx}{\sigma^2}\right) \geq 0.$$

Substituting $y = \exp(2x/\sigma^2)$, this holds if and only if

$$w_0 + w_1 \exp\left(-\frac{1}{\sigma^2}\right) y + \cdots + w_k \exp\left(-\frac{k^2}{\sigma^2}\right) y^k \quad \geq \quad 0.$$

Thus, for every dichotomy of M we obtain a dichotomy of the set

$$M' \quad = \quad \{e^{2i/\sigma^2} : i = 0, \ldots, k\}$$

induced by a higher-order neuron with k monomials. In other words, M' is shattered by this neuron. $\square$

5 Comparison with k-Term Monotone DNF

A Boolean formula that is a disjunction of up to k monomials without negations can be considered as a polynomial restricted to Boolean inputs. The previously best known lower bound for the VC dimension of higher-order neurons with k monomials was the bound for k-term monotone DNF due to Littlestone (1988). By deriving an upper bound for the latter class and applying Theorem 4, we show that the VC dimension for higher-order neurons with k monomials is strictly larger than for k-term monotone DNF. We use "log" to denote the logarithm of base 2.

Corollary 5. *Let $n \geq 1$ and $3 \leq k \leq 2^n$. The VC dimension of the higher-order neuron with k monomials in n variables exceeds the VC dimension of the class of k-term monotone DNF in n variables by more than $k \log(k/e) + 1$.*

Proof. A k-term monotone DNF formula corresponds to a collection of up to k subsets of the set of variables. For n variables, there are no more than $\sum_{i=0}^{k} \binom{2^n}{i}$ such collections. The known inequality $\sum_{i=0}^{d} \binom{m}{i} < (em/d)^d$, where $1 \leq d \leq m$, (see, e.g., Anthony and Bartlett, 1999, Theorem 3.7) yields

$$\sum_{i=0}^{k} \binom{2^n}{i} \quad < \quad \left(\frac{e}{k}\right)^k 2^{nk}.$$

By definition, the VC dimension of a finite function class $\mathcal{F}$ cannot be larger than $\log |\mathcal{F}|$. Hence, the VC dimension for k-term monotone DNF is less than $nk - k\log(k/e)$. Theorem 4 implies that this bound falls short of the VC dimension for higher-order neurons with k monomials in n variables by at least $k \log(k/e) + 1$. $\square$

It is easy to see that in the cases $k = 1, 2$, which are not covered by Corollary 5, the VC dimension of higher-order neurons is larger as well. First, as there are no more than 2^n Boolean monotone monomials, the VC dimension of monotone monomials is at most n. Second, the number of monotone DNF formulas with at most two terms is not larger than $2^{2n} + 1$, and $\log(2^{2n} + 1)$ is less than $2n + 1$.

6 Conclusion

A new lower bound for the VC dimension of higher-order neurons with a given number of monomials has been derived. The bound is stronger and more general than the previous bound established via Boolean formulas in monotone DNF. Moreover, the new bound implies that the VC dimension of higher-order neurons with k monomials exceeds the VC dimension of the class of k-term monotone DNF formulas. Therefore, the techniques that use DNF formulas for deriving lower bounds on the VC dimension of higher-order neurons seem to have reached their limits.

We have introduced a method that via Gaussian RBF networks accomplishes to shatter sets by higher-order neurons. This seems to be paradoxical as, with regard to the domain of the parameters, the Gaussian RBF network appears to be more powerful than the higher-order neuron: Each parameter of a Gaussian RBF network may assume any real number, whereas the higher-order neuron must have exponents that are nonnegative and integers. Nevertheless, we have shown here that RBF networks can be used to establish lower bounds on the computational capabilities of higher-order neurons. While the previous lower bound via monotone DNF formulas gives rise to monomials with exponents not larger than 1, the approach that uses RBF networks shows that and how large exponents can be employed to shatter sets of a cardinality that is larger than known before. Moreover, the constructions give reason to a completely new interpretation of the exponent vectors of the monomials when higher-order neurons are used for classification tasks: They have been chosen as centers of balls. This perspective might open new ways to the design of learning algorithms for higher-order neurons.

With the result presented in this article we have narrowed the gap between lower and upper bound for the VC dimension of higher-order neurons. As the bounds are not yet tight it is to be hoped that the method introduced here may lead to further insights that eventually yield additional improvements.

Acknowledgment

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

Bartlett, P. L. and Maass, W. (2003). Vapnik-Chervonenkis dimension of neural nets. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 1188–1192. MIT Press, Cambridge, MA, second edition.

Bartlett, P. L., Maiorov, V., and Meir, R. (1998). Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10:2159–2173.

Ben-David, S. and Lindenbaum, M. (1998). Localization vs. identification of semi-algebraic sets. *Machine Learning*, 32:207–224.

Durbin, R. and Rumelhart, D. (1989). Product units: A computationally powerful and biologically plausible extension to backpropagation networks. *Neural Computation*, 1:133–142.

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

Erlich, Y., Chazan, D., Petrack, S., and Levy, A. (1997). Lower bound on VC-dimension by local shattering. *Neural Computation*, 9:771–776.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, NJ, second edition.

Karpinski, M. and Macintyre, A. (1997). Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176.

Karpinski, M. and Werther, T. (1993). VC dimension and uniform learnability of sparse polynomials and rational functions. *SIAM Journal on Computing*, 22:1276–1285.

Koiran, P. and Sontag, E. D. (1997). Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54:190–198.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1995). Lower bounds on the VC dimension of smoothly parameterized function classes. *Neural Computation*, 7:1040–1053.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, 3:246–257.

Schmitt, M. (2002a). Descartes' rule of signs for radial basis function neural networks. *Neural Computation*, 14:2997–3011.

Schmitt, M. (2002b). Neural networks with local receptive fields and superlinear VC dimension. *Neural Computation*, 14:919–956.

Schmitt, M. (2002c). On the complexity of computing and learning with multiplicative neural networks. *Neural Computation*, 14:241–301.

Schmitt, M. (2004). New designs for the Descartes rule of signs. *American Mathematical Monthly*, 111:159–164.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.