VC Dimension Bounds for Higher-Order Neurons*

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

We investigate the sample complexity for learning using higher-order neurons. We calculate upper and lower bounds on the Vapnik-Chervonenkis dimension and the pseudo dimension for higher-order neurons that allow unrestricted interactions among the input variables. In particular, we show that the degree of interaction is irrelevant for the VC dimension and that the individual degree of the variables plays only a minor role. Further, our results reveal that the crucial parameters that affect the VC dimension of higher-order neurons are the input dimension and the maximum number of occurrences of each variable. The lower bounds that we establish are asymptotically almost tight. In particular, they show that the VC dimension in super-linear in the input dimension. Bounds for higher-order neurons with sigmoidal activation function are also derived.

1 Introduction

Early on in the history of neural network research higher-order neurons have been recognized as natural extensions of the commonly used linearly weighted neuron models such as the threshold gate, or McCulloch-Pitts neuron, and the sigmoidal gate. Networks of higher-order neurons, or sigma-pi units as they are widely called, were not only shown to have larger computational power—a single higher-order neuron is in fact a network with one hidden layer—they also exhibited impressive generalization capabilities and invariance properties when used in applications (see, e.g., [4, 8, 14, 16]). Further, substantial evidence from neurophysiological experiments has accumulated supporting the conclusion that multiplication-like operations play a fundamental role in neural information processing [3, 7, 13]. Such wide-spread interest resulted in a considerable number of learning algorithms that have been taken from linear neural network models, or were even custom-made for higher-order networks (see, e.g., [3, 9, 15]). A problem frequently reported from applications, however, is the fact that training such networks may suffer from combinatorial explosion of higher-order terms. Therefore, some requirement of sparseness has to be imposed on the higher-order neurons to keep learning practical.

In this paper we study the complexity of learning using such sparse higher-order neurons. In particular, we investigate the sample complexity for higher-order neurons in terms of the Vapnik-Chervonenkis (VC) dimension. It is well known that the VC dimension of a function class yields asymptotically tight bounds on the number of training examples needed for probably approximately correct (PAC) learning this class. For detailed definitions and results concerning this model of learnability we refer the reader to [11, 17] and the references therein. Moreover, the bounds for the sample complexity in terms of the VC dimension are even valid for agnostic PAC learning, that is, in the case when the training examples are generated by some arbitrary probability distribution [6, 17]. The functions that we take into consideration as hypotheses for learning are computed by classes of higher-order neurons employing certain activation functions. An activation function may be the standard sigmoid, the identity or the sign function. We characterize such a class

*Work supported in part by the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2, No. 27150.

of higher-order neurons by the three parameters n, k and d, where n is the number of variables (or input dimension), k is the maximum number of occurrences of any variable and d is the largest degree of any individual variable. The bounds established in this paper are given in terms of these parameters. We emphasize that such a class may contain higher-order neurons of total degree n. Thus, there is no restriction on the degree of interaction among the input variables. In contrast, bounds arising from previous results on the VC dimension for polynomial networks depend on the total degree of the neurons [1, 2, 5].

We give a brief outline of the paper and its results. In Section 2 we provide precise definitions of the basic concepts. Section 3 contains derivations of upper bounds on the VC dimension and pseudo dimension, particularly of the main bound which is $O(kn \log(dkn))$. These results show in particular that the total degree of sparse higher-order neurons is not relevant for the VC dimension and further, that the degree of each individual variable is negligible. The upper bounds are contrasted by lower bounds in Section 4. Specifically, we show that the VC and pseudo dimension satisfy the bounds $\Omega(n \log n)$ and $\Omega(k^{1/\log 3} \cdot n)$ (the latter is, roughly, $\Omega(k^{0.63} \cdot n)$). These lower bounds imply that the main upper bound is asymptotically tight with respect to n and almost tight with respect to kn. Interestingly, these lower bounds already hold when the degree of each individual variable is at most one and the input values are from the set $\{-1, 0, 1\}$. Furthermore, the super-linear lower bound in terms of the input dimension n is shown to be valid even for binary inputs $\{-1, 1\}$. Such a super-linear dependency is known for networks of threshold or sigmoidal gates only when the networks have at least two hidden layers [10]. A final section provides some concluding remarks.

2 Basic Definitions

A *higher-order neuron on n inputs* is a sum

$$w_1 m_1 + \ldots + w_r m_r \qquad (1)$$

where $m_1, \ldots, m_r$ are monomials (or product units) over the input variables $x_1, \ldots, x_n$, and $w_1, \ldots, w_r$ are the weights (or parameters) of the neuron. The set $\{m_1, \ldots, m_r\}$ is called the *structure* of the neuron. It may contain the *constant monomial* of value 1. In a higher-order *read-k* neuron each variable occurs ("is read") in at most k monomials. If $k = 1$ we also call it a higher-order *read-once* neuron. The *degree* of a variable x_i in a given higher-order neuron is defined to be the largest degree that x_i has in any of the neuron's monomials. We say that a higher-order neuron has *individual degree d* if each of its variables has degree at most d.

Given a set $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_m\} \subseteq \mathbb{R}^n$, a *dichotomy* of S is a partition of S into disjoint subsets S_0, S_1. A set $\mathcal{F}$ of functions from $\mathbb{R}^n$ to $\{0, 1\}$ is said to *induce* the dichotomy S_0, S_1 on S if there is some $f \in \mathcal{F}$ such that $f(S_0) \subseteq \{0\}$ and $f(S_1) \subseteq \{1\}$. We say that S is *shattered* by a set $\mathcal{F}$ of $\{0,1\}$-valued functions if $\mathcal{F}$ induces all dichotomies on S. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{F}$ is defined as the largest number m such that there is a set of m elements that is shattered by $\mathcal{F}$. Given a set $\mathcal{F}$ of real-valued functions, S is said to be *P-shattered* by $\mathcal{F}$ if there exist real numbers $y_1, \ldots, y_m$ such that every dichotomy of $\{(\mathbf{s}_i, y_i) : i = 1, \ldots, m\}$ is induced by some function of the form $(\mathbf{s}, y) \mapsto \mathrm{sign}(f(\mathbf{s}) - y)$ for some $f \in \mathcal{F}$ (where $\mathrm{sign}(x) = 1$ if $x \geq 0$, and $\mathrm{sign}(x) = 0$ otherwise). The *pseudo dimension* of $\mathcal{F}$ is the largest number m such that there is a set of m elements that is P-shattered by $\mathcal{F}$.

Choosing a higher-order neuron in terms of its structure and assigning real numbers to its weights uniquely specifies a function $f : \mathbb{R}^n \to \mathbb{R}$ where n is the number of the neuron's inputs. The pseudo dimension of a class of neurons is defined to be the pseudo dimension of the set of functions computed by the neurons in the class with all possible assignments of values to their weights. When speaking of the VC dimension of a class of neurons we assume that, in order to make the output binary, the neurons use the sign function as activation function, that is, the output of each neuron is thresholded at 0 by applying the sign function to the value computed by the sum (1). We also consider neurons using real-valued activation functions: A higher-order neuron with *sigmoidal activation function* is one that computes its output value by applying the standard sigmoid $\sigma(y) = 1/(1 + e^{-y})$ to the sum (1).

3 Upper Bounds

In this section we establish upper bounds on the VC dimension and pseudo dimension for classes of higher-order neurons in terms of the number n of inputs, the individual degree d, and the maximum number k of times that any variable occurs. The main result is as follows.

Theorem 1 *The class of higher-order read-k neurons on n inputs with individual degree d has VC dimension $2kn\log(2edkn)$.*

For the proof of this statement we require the following result which generalizes a well-known bound for threshold gates to higher-order neurons with fixed structure. Note that this result does not impose any restrictions on the individual degree of the input variables.

Lemma 2 *Consider a higher-order read-k neuron on n inputs with fixed structure and sign activation function. The number of dichotomies it induces on a set of cardinality m is at most $2(em/(kn))^{kn}$.*

Proof. Assume that the structure of the higher-order neuron consists of r monomials with corresponding weights $w_1, \ldots, w_r$. Let a set $S \subseteq \mathbb{R}^n$ of cardinality m be given. We construct from this S a new set $S' \subseteq \mathbb{R}^r$ of the same cardinality by applying to each element of S the set of monomials such that the i-th monomial gives rise to the i-th component of an element in S'. (In other words, S' is the set of activation vectors of the product units when applied to S.) Now it is obvious that the number of dichotomies that the higher-order neuron induces on S cannot be larger than the number of dichotomies that a threshold gate with weights $w_1, \ldots, w_r$ induces on S'. The latter number is known to be bounded from above by $2(em/r)^r$. (See, e.g., [12] for a derivation.) Since the neuron satisfies the read-k condition we have that $r \leq kn$. Hence, the higher-order neuron induces on S at most $2(em/(kn))^{kn}$ dichotomies. $\square$

Proof of Theorem 1. We first estimate the number of dichotomies induced by the specified class of higher-order neurons on an arbitrary set S of cardinality m. Then, assuming that S is shattered we derive the claimed result.

For a higher-order read-k neuron on n inputs with individual degree d there are at most $(dkn)^{kn}$ different structures. (The structure can have at most kn monomials and each occurrence of a variable must have a degree from $\{1, \ldots, d\}$.) Now let a set $S \subseteq \mathbb{R}^n$ of cardinality m be given. From Lemma 2 we know that a higher-order read-k neuron with fixed structure and variable weights can induce at most $2(em/(kn))^{kn}$ dichotomies on S. Varying over all structures and weight assignments yields a number of at most

$$(dkn)^{kn} \cdot 2(em/(kn))^{kn}$$

different dichotomies that are induced on S by the class of neurons considered.

Assume now that S is shattered by this class, that is, all 2^m dichotomies of S are induced. Then we must have $2^m \leq 2(edm)^{kn}$ which implies

$$m \leq kn\log(edm) + 1.$$

From this we obtain that $m \leq 2kn\log(2edkn)$ by a calculation which is omitted here. This proves the statement of the theorem. $\square$

The following corollary summarizes the main result of this section and its implications for the pseudo dimension. The statement concerning the pseudo dimension of higher-order read-k neurons immediately follows from the definitions. For higher-order neurons with sigmoidal activation function it is implied by the fact that the standard sigmoid is nondecreasing (see, e.g., [6]).

Corollary 3 *The class of higher-order read-k neurons on n inputs with individual degree d has VC dimension $O(kn\log(dkn))$. This bound also holds for the pseudo dimension of this class. Furthermore, the bound is valid for the pseudo dimension of the class of higher-order read-k neurons on n inputs with individual degree d and sigmoidal activation function.*

4 Lower Bounds

We now calculate lower bounds for classes of higher-order neurons. The first result shows that the upper bound from the previous section is asymptotically tight with respect to the number of inputs. We derive this bound from the following more general statement.

Theorem 4 *For each $m, l \geq 1$ there exists a set $S \subseteq \{-1,1\}^{m+2^l}$ of cardinality $|S| = m \cdot l$ that is shattered by the class of higher-order read-once neurons with individual degree one.*

Proof. For $a = 1, \ldots, m$ let $\mathbf{u}_a \in \{-1,1\}^m$ and denote by $\mathbf{u}_a(i)$ the i-th component of $\mathbf{u}_a$. We define

$$\mathbf{u}_a(i) = \begin{cases} -1 & \text{if } i = a \text{ ,} \\ 1 & \text{otherwise ,} \end{cases} \tag{2}$$

for $i = 1, \ldots, m$. Further, let $L_1, \ldots, L_{2^l}$ be an enumeration of the subsets of $\{1, \ldots, l\}$. For $b \in 1, \ldots, 2^l$ we define $\mathbf{v}_b \in \{-1,1\}^{2^l}$ by

$$\mathbf{v}_b(j) = \begin{cases} -1 & \text{if } j \in L_b \text{ ,} \\ 1 & \text{otherwise ,} \end{cases}$$

for $j = 1, \ldots, 2^l$. Then the set $S \subseteq \{-1,1\}^{m+2^l}$ is

$$S = \{\mathbf{u}_a : a = 1, \ldots, m\} \times \{\mathbf{v}_b : b = 1, \ldots, l\}.$$

Obviously, the cardinality of S is $m \cdot l$. In order to show that S can be shattered by the class of higher-order neurons as claimed suppose that an arbitrary dichotomy S_0, S_1 of S is given. Denote the input variables by $x_1, \ldots, x_m, y_1, \ldots, y_{2^l}$ such that the $\mathbf{u}$ components of S are assigned to the x variables and the $\mathbf{v}$ components to the y variables. We construct 2^l monomials $m_1, \ldots, m_{2^l}$ over these variables as follows: Assume that the monomials are empty at the beginning of the construction. First, for $j = 1, \ldots, 2^l$ we put variable y_j into the corresponding monomial m_j. Let the function $\beta : \{1, \ldots, m\} \rightarrow \{1, \ldots, 2^l\}$ satisfy

$$L_{\beta(a)} = \{b : \mathbf{u}_a \mathbf{v}_b \in S_1\}$$

where $\mathbf{u}_a \mathbf{v}_b$ is the concatenation of the vectors $\mathbf{u}_a$ and $\mathbf{v}_b$. Clearly, the function β is well-defined. Now, for $a = 1, \ldots, m$, whenever $L_{\beta(a)} \neq \emptyset$ we put variable x_a into monomial $m_{\beta(a)}$. The weights for the monomials are all set to 1. Finally, we add the constant monomial with weight -1. This completes the construction of the monomials. Since each variable occurs in at most one monomial and if so, with degree one, the neuron satisfies the conditions as claimed. The final step is to verify that the function computed by this neuron indeed induces the dichotomy S_0, S_1.

Suppose that $\mathbf{s} \in S_0$ where $\mathbf{s} = \mathbf{u}_a \mathbf{v}_b$. Since the definition of β implies that $b \notin L_{\beta(a)}$, the variable x_a does not occur in any of the monomials. Therefore, all x variables occurring have value 1 and the output value of the neuron does not depend on $\mathbf{u}_a$. On $\mathbf{v}_b$ exactly half of the monomials yield -1, the other half yield 1. This arises from the definition of $\mathbf{v}_b(j)$ since each j is a member of exactly half of the sets $L_1 \ldots, L_{2^l}$. Thus the values of the non-constant monomials add up to 0 and the output value is -1.

On the other hand suppose that $\mathbf{u}_a \mathbf{v}_b \in S_1$. Then $b \in L_{\beta(a)}$ and x_a occurs in monomial $m_{\beta(a)}$. Further, x_a is the only x variable that has value -1. Therefore, compared to the previous case, exactly one monomial changes its value: monomial $m_{\beta(a)}$ from -1 to 1. Thus the sum of the monomials yields 1. Recalling that the output is thresholded at 0 both cases imply that the neuron induces the dichotomy as claimed. $\square$

If we set $m = n/2$ and $l = \lfloor \log(n/2) \rfloor$ in Theorem 4, we get $m + 2^l \leq n$. Hence there exists a set $S \subseteq \{-1,1\}^n$ of cardinality $m \cdot l = \Omega(n \log n)$ that is shattered by the class of higher-order neurons considered in this theorem. By definition, this follows for the pseudo dimension of this class, and using the fact that sigmoidal gates can approximate threshold gates (by scaling the weights if necessary) it also follows for higher-order neurons with sigmoidal activation function.

Corollary 5 *The VC dimension of the class of higher-order read-once neurons on n inputs is $\Omega(n \log n)$. This even holds if the individual degree is one and all input values are from the set $\{-1,1\}$. Moreover, this result is valid for the pseudo dimension of this class and the class of higher-order read-once neurons with sigmoidal activation function.*

The next result will be used to establish a lower bound also in terms of k, the maximum number of occurrences of any variable.

Theorem 6 *For each $m, l \geq 1$ there exists a set $S \subseteq \{-1,0,1\}^{m+\lfloor \log l \rfloor + 1}$ of cardinality $|S| = m \cdot l$ that is shattered by the class of higher-order read-$(3l^{\log 3})$ neurons with individual degree one.*

Proof. Let $S \subseteq \{-1,0,1\}^{m+\lfloor \log l \rfloor+1}$ be the set

$$S = \{\mathbf{u}_a : a = 1,\ldots,m\} \times \{\mathbf{v}_b : b = 0,\ldots,l-1\}$$

where $\mathbf{u}_a \in \{-1,1\}^m$ is as in Eq. (2) of Theorem 4 and $\mathbf{v}_b \in \{0,1\}^{\lfloor \log l \rfloor+1}$ is the binary representation of b. Obviously, S has cardinality $m \cdot l$.

Let S_0, S_1 be an arbitrary dichotomy of S. We show that it can be induced by a higher-order neuron as claimed. Denote the input variables by $x_1,\ldots,x_m$ and $y_1,\ldots,y_{\lfloor \log l \rfloor+1}$ for the $\mathbf{u}$ and $\mathbf{v}$ components respectively. Assume that we are allowed to use monomials that contain—besides factors x_i and y_j—also factors of the type $(1-y_j)$. We call such a monomial a y-incorrect monomial. We construct l such monomials m_b, where $b = 0,\ldots,l-1$, as follows: For $j = 1,\ldots,\lfloor \log l \rfloor+1$, monomial m_b contains the factor y_j if $\mathbf{v}_b(j) = 1$, otherwise it contains the factor $(1-y_j)$. Thus, m_b reflects the binary representation of b. Furthermore, for each $\mathbf{u}_a\mathbf{v}_b \in S_1$ we put variable x_a into monomial m_b. This completes the construction of the monomials. We assign to each of them the value -1 as weight. Note that a y-incorrect monomial with γ factors of the type $(1-y_j)$ can be written as a sum of 2^γ correct monomials. This implies that the sum of the l y-incorrect monomials can be equivalently transformed into a sum of at most

$$\sum_{\gamma=0}^{\lfloor \log l \rfloor+1} \binom{\lfloor \log l \rfloor+1}{\gamma} 2^\gamma \leq 3l^{\log 3}$$

correct monomials. Further, this transformation does not increase the degree of each individual variable. Thus we obtain a sum of correct monomials where each variable occurs at most $3l^{\log 3}$ times and has individual degree one as required.

In order to see that this higher-order neuron induces the dichotomy S_0, S_1 it suffices to consider its equivalent formulation in terms of the y-incorrect monomials. For each $\mathbf{u}_a\mathbf{v}_b \in S$ there is exactly one of these monomials that does not yield 0: This is monomial m_b. Now, if $\mathbf{u}_a\mathbf{v}_b \in S_0$ then all x variables in m_b receive value 1. Since the monomial's weight has value -1, the output of the neuron is -1. On the other hand, if $\mathbf{u}_a\mathbf{v}_b \in S_1$ then x_a, which receives value -1, occurs in m_b yielding output 1. Hence, the dichotomy is induced as claimed. □

Choosing $m = n - \lfloor \log(k^{1/\log 3}/3) \rfloor - 1$ and $l = k^{1/\log 3}/3$ such that $k \leq (3 \cdot 2^{\varepsilon n})^{\log 3}$ for some fixed ε with $0 < \varepsilon < 1$, we obtain $m + \lfloor \log l \rfloor + 1 = n$ and $m \cdot l \geq (n - \varepsilon n - 1)k^{1/\log 3}/3$. Hence, by virtue of Theorem 6 the class of higher-order read-k neurons with individual degree one shatters some set $S \subseteq \{-1,0,1\}^n$ of cardinality $\Omega(k^{1/\log 3} \cdot n)$, which is roughly $\Omega(k^{0.63} \cdot n)$. Thus, we have the following lower bounds on the VC dimension and pseudo dimension in terms of n and k.

Corollary 7 *The VC dimension of the class of higher-order read-k neurons on n inputs is $\Omega(k^{1/\log 3} \cdot n)$. This even holds if the individual degree is one and all input values are from the set $\{-1,0,1\}$. Moreover, this result is valid for the pseudo dimension of this class and the class of higher-order read-k neurons with sigmoidal activation function.*

5 Conclusions

We have established almost tight bounds on the VC dimension and pseudo dimension for higher-order neurons. These bounds can be used now to estimate the number of examples that a learning algorithm needs to train higher-order neurons that generalize well. Moreover, the results can be applied even when there is no a priori restriction on the degree of interaction among the input variables of a higher order neuron generated by such a learning algorithm. Instead, only the individual degree and the maximum number of occurrences of each variable appear in the bounds.

In particular, the results imply that the sample complexity for sparse higher-order neurons is not affected by the total degree and only marginally affected by the individual degree of the variables. For learning applications this justifies the use of higher-order neurons of unlimited degree without having to cope with a much higher sample complexity. On the other hand, the superlinear lower bound that we have established in terms of the input dimension shows that the variety of the class of functions that can be computed by single sparse higher-order neurons comes very close to the richness of functions computed by some hidden-layer networks of threshold or sigmoidal gates. This might suggest the use of single higher-order neurons for tasks that require at least networks of threshold or sigmoidal gates.

Finally, a further lower bound established here implies that the VC dimension is small only when each variable does not occur too often. It is therefore an interesting task to develop learning algorithms that are particularly capable of keeping these numbers low. We believe that by using such general and well-known techniques as pruning or weight elimination many learning algorithms can be adapted such that these requirements are met.

References

[1] M. Anthony. Classification by polynomial surfaces. *Discrete Applied Mathematics*, 61:91–103, 1993.

[2] P. L. Bartlett, V. Maiorov, and R. Meir. Almost linear VC dimension bounds for piecewise polynomial networks. *Neural Computation*, 10(8):2159–2173, 1998.

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

[4] C. L. Giles and T. Maxwell. Learning, invariance, and generalization in high-order neural networks. *Applied Optics*, 26:4972–4978, 1987.

[5] P. W. Goldberg and M. R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

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

[7] C. Koch and T. Poggio. Multiplying with synapses and neurons. In T. McKenna, J. Davis, and S. Zornetzer, editors, *Single Neuron Computation*, chapter 12, pages 315–345. Academic Press, Boston, Mass., 1992.

[8] Y. C. Lee, G. Dolen, H. H. Chen, G. Z. Sun, T. Maxwell, H. Lee, and C. L. Giles. Machine learning using a higher order correlation network. *Physica D*, 22:276–306, 1986.

[9] L. R. Leerink, C. L. Giles, B. G. Horne, and M. A. Jabri. Learning with product units. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 537–544, 1995.

[10] W. Maass. Neural nets with superlinear VC-dimension. *Neural Computation*, 6:877–884, 1994.

[11] W. Maass. Vapnik-Chervonenkis dimension of neural nets. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1000–1003. MIT Press, Cambridge, Mass., 1995.

[12] W. Maass and M. Schmitt. On the complexity of learning for a spiking neuron. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 54–61, ACM, New York, 1997.

[13] B. W. Mel. Information processing in dendritic trees. *Neural Computation*, 6:1031–1085, 1994.

[14] S. J. Perantonis and P. J. G. Lisboa. Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers. *IEEE Transactions on Neural Networks*, 3:241–251, 1992.

[15] N. J. Redding, A. Kowalczyk, and T. Downs. Constructive higher-order network algorithm that is polynomial time. *Neural Networks*, 6:997–1010, 1993.

[16] L. Spirkovska and M. B. Reid. Higher-order neural networks applied to 2D and 3D object recognition. *Machine Learning*, 15:169–199, 1994.

[17] M. Vidyasagar. *A Theory of Learning and Generalization.* Communications and Control Engineering. Springer, London, 1997.