On the Sample Complexity for Neural Trees

Michael Schmitt*

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

Abstract. A neural tree is a feedforward neural network with at most one edge outgoing from each node. We investigate the number of examples that a learning algorithm needs when using neural trees as hypothesis class. We give bounds for this sample complexity in terms of the VC dimension. We consider trees consisting of threshold, sigmoidal and linear gates. In particular, we show that the class of threshold trees and the class of sigmoidal trees on n inputs both have VC dimension $\Omega(n \log n)$. This bound is asymptotically tight for the class of threshold trees. We also present an upper bound for this class where the constants involved are considerably smaller than in a previous calculation. Finally, we argue that the VC dimension of threshold or sigmoidal trees cannot become larger by allowing the nodes to compute linear functions. This sheds some light on a recent result that exhibited neural networks with quadratic VC dimension.

1 Introduction

The sample complexity, that is, the number of examples required for a learning algorithm to create hypotheses that generalize well, is a central issue in the theory of machine learning. In this paper we study the sample complexity for hypothesis classes consisting of neural trees. These are feedforward neural networks where each node, except the output node, has exactly one outgoing connection. Since these networks have less degrees of freedom, they are expected to be learnable more efficiently than unrestricted neural networks.

The computational complexity of learning using trees has been extensively studied in the literature. Angluin et al. [1], for instance, investigated the existence of efficient algorithms that use queries to learn Boolean trees, also known as read-once formulas. Research on trees employing neural gates has been initiated by Golea et al. [8]. They designed algorithms for learning so-called μ-Perceptron networks with binary weights. (A μ-Perceptron network is a disjunction of threshold gates where each input node is connected to exactly one threshold gate.) They

* Part of this work was done while the author was with the Institute for Theoretical Computer Science at the Technische Universität Graz, A-8010 Graz, Austria.

also considered tree structures in the form of μ-Perceptron decision lists [9] and nonoverlapping Perceptron networks [10].

We investigate the sample complexity for neural trees in terms of their Vapnik-Chervonenkis (VC) dimension. It is well known that the VC dimension of a function class gives asymptotically tight bounds on the number of training examples needed for probably approximately correct (PAC) learning this class. For detailed definitions we refer the reader to [2, 4, 22]. Moreover, these estimates of the sample complexity in terms of the VC dimension hold even for agnostic PAC learning, that is, in the case when the training examples are generated by some arbitrary probability distribution [11]. Furthermore, the VC dimension is known to yield bounds for the complexity of learning in various on-line learning models [14, 18].

Results on the VC dimension for neural networks abound. See, for instance, the survey by Maass [16]. We briefly mention the most relevant ones for this paper. A feedforward network of threshold gates is known to have VC dimension at most $O(w \log w)$ where w is the number of weights [3]. Networks using piecewise polynomial functions for their gates have VC dimension $O(w^2)$ [7] whereas for sigmoidal networks the bound $O(w^4)$ is known [12]. With respect to lower bounds it has been shown that there are threshold networks with VC dimension $\Omega(w \log w)$ [15]. Furthermore, networks with VC dimension $\Omega(w^2)$ have been exhibited [13]. Among these are networks that consist of both threshold and linear gates, and sigmoidal networks.

Bounds on the VC dimension for neural networks are usually given in terms of the number of programmable parameters, that are, most commonly, the weights, of these networks. In contrast to the majority of the results in the literature, however, we are not looking at the VC dimension of a single network with a fixed underlying graph, but of the entire class of trees employing a specified activation function. This must be taken into account when comparing our results with other ones.

Hancock et al. [10] have shown that the VC dimension of the class of trees having threshold gates as nodes—they called them nonoverlapping Perceptron networks—is $O(n \log n)$ where n is the number of inputs.[1] We take this result as a starting line for our work. The basic definitions are introduced in Section 2. In Section 3 we show that the class of threshold trees has VC dimension $\Omega(n \log n)$, which is asymptotically tight. Moreover, we show that this bound remains valid even when all trees are required to have depth two and the output gate computes a disjunction. This lower bound is then easily transferred to trees with sigmoidal gates. In Section 4 we provide a new calculation for an upper bound that considerably improves the constants of [10]. Section 5 is a short note on how to derive the upper bound $O(n^4)$ for the class of sigmoidal trees. Finally, in Section 6 we show that adding linear gates to threshold or sigmoidal trees cannot increase their VC dimension. Interestingly, it was this use of linear gates that lead to

[1] Since a neural tree on n inputs has $O(n)$ weights, it is convenient to formulate the bounds in terms of the number of inputs. We follow this convention throughout the paper.

a quadratic lower bound for sigmoidal neural networks in the work by Koiran and Sontag [13]. Consequently, if the lower bound $\Omega(n \log n)$ is not tight for sigmoidal trees one has to look for new techniques in search for asymptotically better bounds.

2 Basic Definitions

A *neural tree* is a feedforward neural network where the connectivity, or architecture, of the network is a tree, that is, there is at most one edge outgoing from each node. Further, there is exactly one node, the root of the tree or *output node*, that has no edge outgoing. A *neural tree on n inputs* has n leaves, also called *input nodes*. The *depth* of a tree is the length of the longest path from an input node to the output node. The nodes that are not leaves are also known as *computation nodes*. Each computation node has associated with itself a set of $k+1$ real-valued parameters where k is the in-degree of the node: the weights $w_1, \ldots, w_k$ and the threshold t.

We use trees for computations over the reals by assigning functions to its computation nodes and values to their parameters. We consider three types of functions that the nodes may use. All types can be obtained by applying a so-called activation function to the weighted sum $w_1 x_1 + \cdots + w_k x_k - t$ where $x_1, \ldots, x_k$ are the input values for the node computed by its predecessors. (The values computed by the input nodes are the input values for the tree.) A node becomes a *threshold gate* when it uses the signum function with $\mathrm{sign}(y) = 1$ if $y \geq 0$, and $\mathrm{sign}(y) = 0$ otherwise. A *sigmoidal gate* is a node that uses the sigmoidal function $1/(1+e^{-y})$. Finally, a *linear gate* applies the identity function, that is, it simply outputs the weighted sum.

We say that a neural tree is a *threshold tree* if all its computation nodes are threshold gates. Correspondingly, we speak of *sigmoidal trees* and *linear trees*. If we allow more than one type of activation function for a tree then we shall assume that each of its computation nodes may use all types specified. Since we restrict our investigations to trees that compute $\{0,1\}$-valued functions, we assume the output of a tree to be thresholded at $1/2$ if the output node is a linear or sigmoidal gate. Thus we can associate with each tree on n inputs a set of functions from $\mathbb{R}^n$ to $\{0,1\}$ which are obtained by choosing activation functions for its nodes and varying its parameters over the reals. In a *class of trees* all members have the same number of inputs, denoted by n, and choose gate functions for their nodes from a specified set, which can be of the three types introduced above. The set of functions computed by a class of trees is then defined straightforward as the union of the sets computed by its members.

A *dichotomy* of a set $S \subseteq \mathbb{R}^n$ is a partition of S into two disjoint subsets S_0, S_1 such that $S_0 \cup S_1 = S$. Given a set $\mathcal{F}$ of functions from $\mathbb{R}^n$ to $\{0,1\}$ and a dichotomy S_0, S_1 of S, we say that $\mathcal{F}$ *induces* the dichotomy S_0, S_1 *on* S if there is a function $f \in \mathcal{F}$ such that $f(S_0) \subseteq \{0\}$ and $f(S_1) \subseteq \{1\}$. We say further that $\mathcal{F}$ *shatters* S if $\mathcal{F}$ induces all dichotomies on S. The *Vapnik-Chervonenkis (VC)*

dimension of $\mathcal{F}$, $\mathrm{VCdim}(\mathcal{F})$, is defined as the largest number m such that there is a set of m elements that is shattered by $\mathcal{F}$.

3 Lower Bounds on the VC Dimension for Threshold and Sigmoidal Trees

In this section we consider neural trees consisting of threshold and sigmoidal gates. We first establish a lower bound for the VC dimension for a class of threshold trees with certain restrictions: We assume that each tree has only one layer of hidden nodes and that the output node computes a disjunction.

Theorem 1. *For each $m, k \geq 1$ there exists a set $S \subseteq \{0,1\}^{m+k\cdot 2^{k-1}}$ of cardinality $|S| = m \cdot k$ that is shattered by the class of depth-two threshold trees with disjunctions as output gates.*

Proof. Let the set $S \subseteq \{0,1\}^{m+k\cdot 2^{k-1}}$ be defined as

$$S = \{e_i : i = 1, \ldots, m\} \times \{d_j : j = 1, \ldots, k\}$$

where $e_i \in \{0,1\}^m$ is the i-th unit vector and $d_j \in \{0,1\}^{k\cdot 2^{k-1}}$ is specified as follows: Let $A_1, \ldots, A_{2^k}$ be an enumeration of all subsets of $\{1, \ldots, k\}$. We arrange the $k \cdot 2^{k-1}$ components of d_j in 2^k blocks $b_1, \ldots, b_{2^k}$ such that block b_l has length $|A_l|$ for $l = 1, \ldots, 2^k$. (Hence the block corresponding to the empty set has length 0.) Thus all blocks together comprise exactly $k \cdot 2^{k-1}$ components.

Fix $j \in \{1, \ldots, k\}$, $l \in \{1, \ldots, 2^k\}$ and consider the components of block b_l in d_j. Each of these components is assumed to represent one specific element of A_l. For such an element $a \in A_l$ let $d_j(b_l, a)$ denote the value of this component. We define

$$d_j(b_l, a) = \begin{cases} 1 & \text{if } a = j \\ 0 & \text{otherwise .} \end{cases}$$

Thus we proceed for $j = 1, \ldots, k$ and $l = 1, \ldots, 2^k$. Observe that due to this construction, each block b_l contains at most one 1. (Furthermore, the number of 1s in d_j is equal to 2^{k-1} since j occurs in exactly half of all subsets A_l.)

Obviously, S consists of $m \cdot k$ elements. We claim that S is shattered by the class of depth-two threshold trees having a disjunction as output gate. In order to prove this we show that for each $S' \subseteq S$ there are weights and thresholds for such a tree such that this network outputs 1 for elements in S', and 0 for elements in $S \backslash S'$. Fix $S' \subseteq S$. For $i = 1, \ldots, m$ let $\alpha(i)$ be the (unique) element in $\{1, \ldots, 2^k\}$ such that

$$A_{\alpha(i)} = \{j : e_i d_j \in S'\} .$$

For convenience we call the input nodes $1, \ldots, m$ the e-inputs, and the input nodes $m+1, \ldots, m+k\cdot 2^{k-1}$ the d-inputs. We employ for each element in the range of α a threshold gate $G_{\alpha(i)}$ that has connections from all d-inputs in block $b_{\alpha(i)}$ and from none of the other blocks b_l where $l \neq \alpha(i)$. Further, we

connect e-input i to gate $G_{\alpha(i)}$ for $i = 1, \ldots, m$. (Notice that this may result in gates that have connections from more than one e-input.) The weights of all connections are fixed to 1 and the thresholds are set to 2. Obviously, there is at most one connection outgoing from each input node, so that the disjunction of these threshold gates is a tree.

Finally, we verify that the network computes the desired function on S. Suppose that $x \in S'$ where $x = e_i d_j$. The definition of α implies that $j \in A_{\alpha(i)}$. Hence $d_j(b_{\alpha(i)}, j)$ is defined and has value $d_j(b_{\alpha(i)}, j) = 1$. Since gate $G_{\alpha(i)}$ receives two 1s—one from e-input i and one from block $b_{\alpha(i)}$—the output of the network for $e_i d_j$ is 1.

Assume on the other hand that $e_i d_j \in S \backslash S'$. Then $j \notin A_{\alpha(i)}$ and gate $G_{\alpha(i)}$, which is the only gate that receives a 1 from an e-input, receives only 0s from block $b_{\alpha(i)}$. All other gates G_l, where $l \neq \alpha(i)$, receive at most one 1, which is then from block b_l only. Hence, the output of the network for $e_i d_j$ is 0. □

Choosing $m = n/2$ and $k = (\log(n/2))/2 + 1$ in Theorem 1, we have $m + k \cdot 2^{k-1} \leq n/2 + 2^{2(k-1)} = n$. Hence there is a set $S \subseteq \{0,1\}^n$ of cardinality $m \cdot k = \Omega(n \log n)$ that is shattered by the class of trees considered.

Corollary 2. *The VC dimension of the class of threshold trees on n inputs is $\Omega(n \log n)$. This even holds if all input values are binary and the class is restricted to trees of depth two with a disjunction as output gate.*

It is well known that in a network that computes a Boolean function, a threshold gate can be replaced by a sigmoidal gate without changing the function of the network. (If necessary, the weights have to be scaled appropriately. See, for instance, Maass et al. [17] for a treatment in the context of circuit complexity). Thus, the lower bound $\Omega(n \log n)$ also holds for depth-two trees that may consist of threshold or sigmoidal gates.

Corollary 3. *The class of depth-two trees on n inputs with threshold or sigmoidal gates has VC dimension $\Omega(n \log n)$. This even holds if the inputs are restricted to binary values.*

We note that depth two is minimal for this lower bound since a threshold gate and a sigmoidal gate both have VC dimension $n + 1$: This follows for the threshold gate from a bound on the number of regions generated by a certain number of hyperplanes in $\mathbb{R}^n$ which is due to Schläfli [20] (see also [5]). For the sigmoidal gate this follows from the fact that its pseudo dimension is $n + 1$ [11].

Together with the upper bound $O(n \log n)$ due to Hancock et al. [10] we get asymptotically tight bounds for the class of threshold trees.

Corollary 4. *The VC dimension of the class of threshold trees on n inputs is $\Theta(n \log n)$. This holds also for the class of depth-two threshold trees. Moreover, this bound is valid for both classes if the inputs are restricted to binary values.*

4 Improved Upper Bounds for Threshold Trees

In this section we establish upper bounds for the VC dimension of the class of threshold trees. Regarding the constants they are better than a previous bound derived by Hancock et al. [10] which is $13n\log(2en) + 4n\log\log(4n)$.

Theorem 5. *The class of threshold trees on n inputs, where $n \geq 16e$, has VC dimension at most $6n\log(\sqrt{3}n)$.*

Proof. We estimate the number of dichotomies that are induced by the class of threshold trees on an arbitrary set of cardinality m. First we bound the number of these trees, then we calculate an upper bound on the number of dichotomies that a single tree induces when all its weights and thresholds are varied.

For the number of trees on n inputs we use the upper bound $(4n)^{n-1}$, which was derived in [10, Lemma 3]. We assume without loss of generality that the computation nodes at the lowest level (i.e., those nodes that have input nodes as predecessors) have in-degree 1 and that all other computation nodes have in-degree at least 2. Each of the computation nodes at the lowest level induces at most $2m$ dichotomies on a set of cardinality m. The whole level induces therefore at most $(2m)^n$ different functions. The computation nodes with in-degree at least 2 form a tree that consists of at most $n-1$ nodes and has at most $2n-2$ edges leading to one of these nodes.

According to a result by Shawe-Taylor[2] [21] the number of dichotomies that a threshold network with N computation nodes, partitioned into ν equivalence classes, and W edges induces on a set of cardinality m is at most

$$2^{\nu}\left(\frac{emN}{W-\nu}\right)^{W-\nu} .$$

Using $N = n-1$, $\nu = n-1$, and $W = 2n-2$ we get that the number of dichotomies induced by a threshold tree consisting of $n-1$ computation nodes and $2n-2$ edges is at most $2^{n-1}(em)^{n-1}$.

Putting the bounds together, the total number of dichotomies induced on a set of cardinality m by the class of threshold trees on n inputs is at most

$$(4n)^{n-1}\cdot(2m)^n\cdot 2^{n-1}(em)^{n-1} .$$

Assume now that a set of cardinality m is shattered. Then

$$\begin{aligned} 2^m &\leq (4n)^{n-1}\cdot(2m)^n\cdot 2^{n-1}(em)^{n-1} \\ &= 2(16en)^{n-1}\cdot m^{2n-1} \\ &\leq 2(mn)^{2n-1} . \end{aligned}$$

For the last inequality we have used the assumption $n \geq 16e$. Taking logarithms on both sides we obtain

$$m \leq (2n-1)\log(mn) + 1 .$$

[2] We do not make use of the equivalence relations involved in this result but of the improvement that it achieves compared to [3].

We weaken this to

$$m \leq 2n \log(mn) \ . \tag{1}$$

Assume without loss of generality that $m \geq \log n$. Then it is easy to see that for each such m there is a real number $r \geq 1$ such that m can be written as $m = r \log(rn)$. Substituting this in (1) yields

$$\begin{aligned} r \log(rn) &\leq 2n(\log(rn \log(rn))) \\ &= 2n(\log(rn) + \log\log(rn)) \qquad (2) \\ &\leq 3n \log(rn) \ . \qquad (3) \end{aligned}$$

The last inequality follows from $\log(rn) \leq \sqrt{rn}$ which holds since $rn \geq 16e$. We divide both sides by $\log(rn)$ and get

$$r \leq 3n \ . \tag{4}$$

This implies

$$r \log(rn) \leq 3n \log(3n^2) \ .$$

Resubstituting $m = r \log(rn)$ for the left hand side and rearranging the right hand side yields

$$m \leq 6n \log(\sqrt{3}n)$$

as claimed. □

In the statement of Theorem 5 the number n of inputs is required to satisfy $n \geq 16e$. We shall show now that we can get the upper bound as close to $4n \log(\sqrt{2}n)$ as we want provided that n is large enough.

Theorem 6. *For each $\varepsilon > 0$, the class of threshold trees on n inputs has VC dimension at most $4(1+\varepsilon)n \log(\sqrt{2(1+\varepsilon)}n)$ for all sufficiently large n.*

Proof. (Sketch) Fix $\varepsilon > 0$. For n sufficiently large we have $\log(rn) \leq (rn)^{\varepsilon}$. Using this in the inequality from (2) to (3) we can infer $r \leq 2(1+\varepsilon)n$ in place of (4). This leads then to the claimed result. □

5 A Note on the Upper Bound for Sigmoidal Trees

Using known results on the VC dimension it is straightforward to derive the upper bound $O(n^4)$ for sigmoidal trees. We give a brief account.

Proposition 7. *The class of sigmoidal trees on n inputs has VC dimension $O(n^4)$.*

Proof. The VC dimension of a sigmoidal neural network with w weights is $O(w^4)$. This has been shown by Karpinski and Macintyre [12]. By Sauer's Lemma (see e.g. [2]) the number of dichotomies induced by a class of functions with VC dimension $d \geq 2$ on a set of cardinality $m \geq 2$ can be bounded by m^d. Thus a sigmoidal tree on n inputs induces at most $m^{O(n^4)}$ dichotomies on such a set. Combining this with the bound $(4n)^{n-1}$ employed in the proof of Theorem 5 and using similar arguments we obtain the bound as claimed. □

6 Trees with Linear Gates

By the work of Goldberg and Jerrum [7] it has been known that neural networks employing piecewise polynomial activation functions have VC dimension $O(w^2)$, where w is the number of weights. The question whether this bound is tight for such networks has been settled by Koiran and Sontag [13]. They have shown that networks consisting of threshold and linear gates can have VC dimension $\Omega(w^2)$. This result was somewhat unexpected since networks consisting of linear gates only compute linear functions and have therefore VC dimension $O(w)$. On the other hand, networks consisting of threshold gates only have VC dimension $O(w \log w)$. This follows from work by Cover [6] and has also been shown by Baum and Haussler [3]. Results that this bound is tight for threshold networks are due to Sakurai [19] and Maass [15].

Therefore, the question arises whether a similar increase of the VC dimension is possible for threshold or sigmoidal trees by allowing some of the nodes to compute linear functions. We show now that this cannot happen.

Theorem 8. *Let $\mathcal{T}$ be a class of neural trees consisting of threshold or sigmoidal gates and let $\mathcal{T}^{\mathrm{lin}}$ be a class of trees obtained from $\mathcal{T}$ by replacing some of the gates by linear gates. Then* $\mathrm{VCdim}(\mathcal{T}) \geq \mathrm{VCdim}(\mathcal{T}^{\mathrm{lin}})$.

Proof. We show that nodes computing linear functions can be replaced or eliminated without changing the function of the tree. Assume T is a tree and y is a node in T computing a linear function. If y is the output node then y can be replaced by a threshold gate or a sigmoidal gate, where weights and threshold are modified if necessary. (Note that the output of the tree is thresholded at $1/2$ for linear and sigmoidal output gates.)

If y is a hidden node then there is a unique edge e outgoing from y to its successor z. Denote the weight of e by w. Assume that y computes the function $u_1x_1 + \cdots + u_kx_k - t$ where $x_1, \ldots, x_k$ are the predecessors of y, and $u_1, \ldots, u_k, t$ are its weights and threshold. We delete node y and edge e, and introduce k edges from $x_1, \ldots, x_k$ respectively to z. We assign weight wu_i to the edge from x_i for $i = 1, \ldots, n$ and decrease the threshold of z by wt. It can readily be seen that the resulting network is still a tree and computes the same function as T. □

Combining Theorems 5 and 6 with Theorem 8 we obtain an upper bound for the class of trees with threshold or linear gates.

Corollary 9. *The class of neural trees on n inputs with threshold or linear gates has VC dimension at most $6n\log(\sqrt{3}n)$ for $n \geq 16e$. Furthermore, for each $\varepsilon > 0$, this class has VC dimension at most $4(1+\varepsilon)n\log(\sqrt{2(1+\varepsilon)}n)$ for all sufficiently large n.*

The technique used in the proof can also be applied to neural trees employing a much wider class of gates. If the function computed by a gate can be decomposed into a linear and a non-linear part then the method of deleting a hidden

linear node works the same way. Only if the node to be treated is the output node there have to be some further demands on its function. For instance, if the non-linear part of this function is monotonous then a linear output node can be replaced by such a gate without decreasing the VC dimension of the tree.

7 Conclusions

Finding methods that incorporate prior knowledge into learning algorithms is an active research area in theoretical and applied machine learning. In the case of neural learning algorithms such knowledge might be reflected in a restricted connectivity of the network generated by the algorithm. We have studied the impact of a particular kind of such a restriction on the sample complexity for neural networks. Results were given in terms of bounds for the VC dimension.

We have established the asymptotically tight bound $\Omega(n \log n)$ for the class of threshold trees. We have also derived an improved upper bound for this class. Due to our result demonstrating that the use of linear gates in threshold trees cannot increase their VC dimension, a known technique to construct networks with quadratic VC dimension does not work for trees. As a consequence of this, the gap between the currently best known lower and upper bounds for the class of sigmoidal trees, which are $\Omega(n \log n)$ and $O(n^4)$, is larger than it is for sigmoidal networks. To reduce this gap and to extend these investigations to other frequently used types of gates are challenging open problems for future research.

Acknowledgement. I thank an anonymous referee for helpful comments leading to a clarification in the proof of Theorem 1.

References

1. D. Angluin, L. Hellerstein, and M. Karpinski. Learning read-once formulas with queries. *Journal of the Association for Computing Machinery*, 40:185–210, 1993.
2. M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 1992.
3. E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.
4. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36:929–965, 1989.
5. T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.
6. T. M. Cover. Capacity problems for linear machines. In L. N. Kanal, editor, *Pattern Recognition*, pages 283–289, Thompson Book Co., Washington, 1968.
7. 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.

8. M. Golea, M. Marchand, and T. R. Hancock. On learning μ-Perceptron networks with binary weights. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 591–598. Morgan Kaufmann, San Mateo, CA, 1993.
9. M. Golea, M. Marchand, and T. R. Hancock. On learning μ-Perceptron networks on the uniform distribution. *Neural Networks*, 9:67–82, 1996.
10. T. R. Hancock, M. Golea, and M. Marchand. Learning nonoverlapping Perceptron networks from examples and membership queries. *Machine Learning*, 16:161–183, 1994.
11. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
12. M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176, 1997.
13. P. Koiran and E. D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54:190–198, 1997.
14. N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
15. W. Maass. Neural nets with superlinear VC-dimension. *Neural Computation*, 6:877–884, 1994.
16. 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.
17. W. Maass, G. Schnitger, and E. D. Sontag. A comparison of the computational power of sigmoid and Boolean threshold circuits. In V. Roychowdhury, K.-Y. Siu, and A. Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, pages 127–151. Kluwer, Boston, 1994.
18. W. Maass and G. Turán. Lower bound methods and separation results for on-line learning models. *Machine Learning*, 9:107–145, 1992.
19. A. Sakurai. Tighter bounds of the VC-dimension of three-layer networks. In *Proceedings of the World Congress on Neural Networks WCNN'93*, volume 3, pages 540–543, 1993.
20. L. Schläfli. *Theorie der vielfachen Kontinuität.* Zürcher & Furrer, Zürich, 1901. Reprinted in: L. Schläfli, Gesammelte Mathematische Abhandlungen, Band I, Birkhäuser, Basel, 1950.
21. J. Shawe-Taylor. Sample sizes for threshold networks with equivalences. *Information and Computation*, 118:65–72, 1995.
22. L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.