# VC Dimension of Neural Networks

Eduardo D. Sontag

Department of Mathematics, Rutgers, The State University of New Jersey

**Abstract.** This paper presents a brief introduction to Vapnik-Chervonenkis (VC) dimension, a quantity which characterizes the difficulty of distribution-independent learning. The paper establishes various elementary results, and discusses how to estimate the VC dimension in several examples of interest in neural network theory.

## 1 Introduction

In this expository paper, we present a brief introduction to the subject of computing and estimating the VC dimension of neural network architectures. We provide precise definitions and prove several basic results, discussing also how one estimates VC dimension in several examples of interest in neural network theory.

We do not address the learning and estimation-theoretic applications of VC dimension. (Roughly, the VC dimension is a number which helps to quantify the difficulty when learning from examples. The sample complexity, that is, the number of "learning instances" that one must be exposed to, in order to be reasonably certain to derive accurate predictions from data, is proportional to this number. This relationship can be made mathematically precise using the formalism of computational learning theory and uniform convergence theorems for empirical probabilities, and it is covered in other papers in this volume, as well as in several good books, notably Vidyasagar (1997).)

The VC dimension is geared towards binary classification. It is possible to generalize the notion of VC dimension in several ways, to deal with the problem of "learning" (approximating from data) real-valued functions. This leads to *pseudodimension* (Haussler 1992, Vidyasagar 1997), "fat-shattering dimension" (Anthony and Bartlett n.d.), and several other notions. Reasons of space preclude covering such topics in this paper; however, many of the tools developed here are also central to the study of these generalizations.

## 2 Concepts and VC Dimension

As a starting point for introducing the necessary concepts, we assume given a set $\mathbb{U}$, to be called the *input space*. Typically, $\mathbb{U}$ will be a subset of $\mathbb{R}^m$, for some $m$; we think of inputs as vectors whose coordinates may represent "features" to be used for classification purposes. Also given is a *concept class* $\mathcal{C}$, which consists of a family of subsets of $\mathbb{U}$. Some examples of input spaces and concept classes are as follows:

1. $\mathbb{U} = \mathbb{R}$, $\mathcal{C}$ = infinite open intervals, or the empty set;
2. $\mathbb{U} = \mathbb{R}$, $\mathcal{C} = \mathbb{R}, \emptyset$, one open interval, or two disjoint infinite open intervals;
3. $\mathbb{U} = \mathbb{R}^2$, $\mathcal{C}$ = open half-spaces; and
4. $\mathbb{U} = \mathbb{R}^2$, $\mathcal{C}$ = all convex sets.

**Definition 1** *A finite subset $S \subseteq \mathbb{U}$ is* shattered *by $\mathcal{C}$ if*

$$(\forall\, S_+ \subseteq S)\ (\exists C \in \mathcal{C})\ \ S_+ = S \bigcap C\,. \tag{1}$$

For the above examples, we have that one may shatter, respectively:

1. every 2-element set, no 3-element set;
2. every 3-element set, no 4-element set;
3. every noncollinear 3-element set, no 4-element set; and
4. any finite subset of unit circle.

The first two of these are quite clear. To see that no 4-element set $S \subseteq \mathbb{R}^2$ can be shattered when $\mathcal{C}$ = all open half-spaces, we can argue as follows (see Fig. 1). If some three points of $S$ lie on a line, then $S$ cannot be shattered, because it is impossible to find a half-space which covers only the midpoint. Similarly if one of the points lies in the convex hull of the remaining three points. On the other hand, if no three points lie on a line, then $S$ determines the corners of a quadrilateral, and a set $S_+$ which consists of one pair of opposing corners cannot be the intersection of $S$ and a half-plane. To see that any finite subset of the unit circle can be shattered, in the last example, we may pick any such subset $S$, and any subset $S_+ \subseteq S$, and take the convex hull of $S_+$ as the desired element of $\mathcal{C}$ (this set is the region inside a polygon whose vertices are the points in $S_+$).

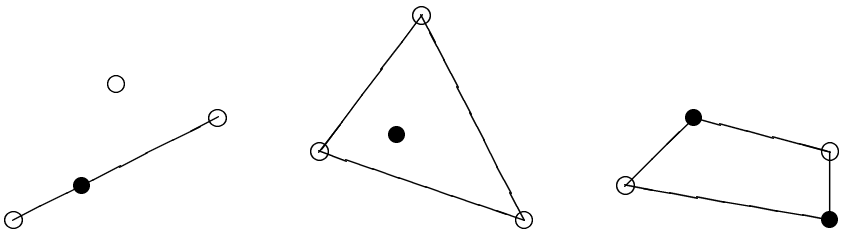

**Fig. 1.** *Four points cannot be shattered by half-spaces*

**Definition 2** *The* Vapnik-Chervonenkis (VC) dimension *of $\mathcal{C}$ is:*

$$\text{VCD}\,(\mathcal{C})\ :=\ \sup\{\,\text{card}\,S \mid S \text{ shattered by } \mathcal{C}\} \tag{2}$$

In the above examples, one obtains, respectively: 2, 3, 3, $\infty$. (Observe that, in the third example, sets of three points that are in a straight line cannot be shattered. Nonetheless, the VC dimension is 3 because *some* 3-element set can be shattered; as a matter of fact, it turns out in this particular example that "almost any" 3-element set can be shattered, but this is not required in the definition.)

### Equivalently, with Functions

The concept-based definition of VC dimension just given originates in combinatorics and computer science. It is useful to provide an equivalent formulation in terms of functions, which is the way in which the subject arises in statistical estimation. Now, instead of $\mathcal{C}$, we assume given a *function class* $\mathcal{F}$, consisting of a set of binary functions $\mathbb{U} \to \{0,1\}$. To each $f \in \mathcal{F}$ we may associate the set

$$C_f = \{\mathbf{u} \in \mathbb{U} \mid f(\mathbf{u}) = 1\} \tag{3}$$

and thus to $\mathcal{F}$ we associate a concept class

$$\mathcal{C}_{\mathcal{F}} := \{C_f, f \in \mathcal{F}\} \ . \tag{4}$$

We define:

$$\mathrm{VCD}\,(\mathcal{F}) \ := \ \mathrm{VCD}\,(\mathcal{C}_{\mathcal{F}})\,. \tag{5}$$

Conversely, to any concept class $\mathcal{C}$ we may associate a function class $\mathcal{F}$ in such a way that $\mathcal{C} = \mathcal{C}_{\mathcal{F}}$ (just take characteristic functions of subsets). We use the two formalisms interchangeably, depending on which is more convenient for any given proof.

Actually, in practice one is often interested in classifiers which arise from neural networks and other devices which produce *real-valued* outputs, and one makes the convention that positive outputs are interpreted as "1" and negative outputs as "0" for purposes of binary classification. Formally, we are now given a set of real-valued functions $\mathcal{F}$, and define, then:

$$\mathrm{VCD}\,(\mathcal{F}) \ := \ \mathrm{VCD}\ (\{\mathcal{H} \circ f, f \in \mathcal{F}\})\,, \tag{6}$$

where $\mathcal{H}(x)$ is the "Heaviside" function which equals 1 if $x > 0$ and equals 0 if $x \leq 0$. (In other words, a "concept" is a set where some possible $f \in \mathcal{F}$ is positive.) We also write $\operatorname{sign} x = \mathcal{H}(x)$.

In this language, saying that a subset $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ is shattered means that, for any possible binary assignment $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{0,1\}^n$, there must exist some function $f = f_\varepsilon \in \mathcal{F}$ which has precisely these signs, i.e., $\mathcal{H}(f(\mathbf{u}_i)) = \varepsilon_i$.

**Parametrized Classes of Functions**

The class $\mathcal{F}$ is more often than not specified by means of neural network architectures with tunable weights, or some other parametric description (splines with nodes and values to be determined, Fourier series with a fixed number of terms but adjustable frequencies and coefficients, etc). In general, we may suppose given a function

$$\beta : \mathbb{W} \times \mathbb{U} \to \mathbb{R}. \tag{7}$$

Typically, $\mathbb{W} = \mathbb{R}^\rho$, where $\rho$ is called the *number of weights* or *parameters*, and $\mathbf{w} = (w_1, \ldots, w_\rho) \in \mathbb{W}$ is a *weight* or *parameter vector*. For each choice of parameters, we obtain a function:

$$\mathcal{F}_\beta := \{\beta(\mathbf{w}, \cdot) \mid \mathbf{w} \in \mathbb{W}\} \tag{8}$$

and we define

$$\text{VCD}\,(\beta) := \text{VCD}\,(\mathcal{F}_\beta)\,. \tag{9}$$

We can express the first three examples given earlier in this language, as follows:

1. the map $\beta : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ given by $\beta\big((a,b),u\big) := a + bu$ leads to the concept class $\mathcal{C}$ which consists of all open infinite intervals (and $\emptyset$);
2. the map $\beta : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ given by $\beta\big((a,b,c),u\big) := a + bu + cu^2$ leads to the second example ($\mathbb{R}$, $\emptyset$, one open interval, or two disjoint infinite open intervals); and
3. the map $\beta : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}$ given by $\beta\big((a,b,c),(u,v)\big) := a + bu + cv$ leads to the third one (half-spaces).

Observe that in all these examples, the parametric description $\beta$ is *linear in the parameters* $a$, $b$, and $c$, and the VC dimension coincides with the respective number of parameters. This is not a coincidence, as we see next.

## 3 Special Case: Linear Parameterizations

Linearly parametrized classes constitute vector spaces; their dimension is the number of independent parameters. Elementary linear algebra gives that the linear dimension of a vector space of functions $\mathcal{F}$ is finite, and equal to $n$, if and only if there is a set of $n$ functions $\{f_1, \ldots, f_n\} \subseteq \mathcal{F}$, and there are some $n$ points $\mathbf{u}_1, \ldots, \mathbf{u}_n$, so that the following matrix:

$$A = \begin{array}{c|ccccc} & f_1 & f_2 & \cdots & f_n \\ \mathbf{u}_1 & f_1(\mathbf{u}_1) & f_2(\mathbf{u}_1) & \cdots & f_n(\mathbf{u}_1) \\ \mathbf{u}_2 & f_1(\mathbf{u}_2) & f_2(\mathbf{u}_2) & \cdots & f_n(\mathbf{u}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{u}_n & f_1(\mathbf{u}_n) & f_2(\mathbf{u}_n) & \cdots & f_n(\mathbf{u}_n) \end{array} \tag{10}$$

has rank $n$, and any $n+1$ by $n+1$ matrix of this type is singular. The following general fact is very useful for linear classes:

**Theorem 1.** *If $\mathcal{F}$ is a vector subspace of $\mathbb{R}^{\mathbb{U}}$, then* $\text{VCD}\,\mathcal{F} = \dim \mathcal{F}$.

*Proof.* Saying that a subset $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ is shattered by $\mathcal{F}$ is the same as saying that there exists some set of $2^n$ functions, let us say $f_1, \ldots f_{2^n}$, for which the columns of the array $f_j(u_i)$ assume all possible $2^n$ sign vectors, that is, we have a sign pattern as follows (after if necessary rearranging the $f_i$'s):

$$B = \begin{array}{c|ccccc} & f_1 & f_2 & f_3 & \cdots & f_{2^n} \\ \mathbf{u}_1 & - & - & - & \cdots & + \\ \mathbf{u}_2 & - & - & - & \cdots & + \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{u}_{n-1} & - & - & + & \cdots & + \\ \mathbf{u}_n & - & + & - & \cdots & + \end{array} \tag{11}$$

(using "$-$" to mean entries $\leq 0$).

Pick any $n$, and suppose that the VC dimension of $\mathcal{F}$ (which could be infinite) is $\geq n$. By definition, there is some subset $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ which is shattered. We claim that this matrix $B$ has rank $n$ (the number of rows). Indeed, if this is not the case, then the rows are linearly independent, i.e., there is some vector $\nu \neq 0$ so that $\nu B = 0$. Assume that there is such a vector, and consider the vector of signs of $\nu$, say $(+, +, -, +, \ldots +)$. Now pick that column $f_i$ of $B$ which has exactly these signs. The inner product $\nu \cdot f_i$ is obviously positive (sum of nonnegative terms, at least one nonzero), contradicting $\nu B = 0$. Since $\text{rank}\, B = n$, there are $n$ linearly independent columns, let us say $f_{i_1}, \ldots, f_{i_n}$. These give rise to a submatrix $A$ as in (10), so $\dim \mathcal{F} \geq n$. As this holds for any $n$, we have that $\dim \mathcal{F} \geq \text{VCD}\,(\mathcal{F})$.

To show the converse inequality, we start with a set of linearly independent elements $f_1, \ldots, f_n$ and a set $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ so that the matrix in Equation (10) has rank $n$. We claim that the set $S$ is shattered. To prove this, we consider any binary assignment $\varepsilon \in \{0,1\}^n$. We need to verify that there is some function $f \in \mathcal{F}$ which has precisely these signs, i.e. $\mathcal{H}(f(\mathbf{u}_i)) = \varepsilon_i$. Since $A$ has rank $n$, there exists some vector $v \in \mathbb{R}^n$ so that $Av = \varepsilon$. Thus $f = (f_1, \ldots, f_n)v$ (which belongs to the subspace $\mathcal{F}$) has signs given by $\varepsilon$ on the respective $\mathbf{u}_i$'s, as desired. As $S$ is shattered, $\text{VCD}\,(\mathcal{F}) \geq n$, which shows $\text{VCD}\,(\mathcal{F}) \geq \dim \mathcal{F}$. ∎

Since in the first three examples there were 2, 3, and 3 independent parameters, appearing linearly, this theorem provides VC dimensions of 2, 3, and 3 respectively, just as we had found directly.

**Margins** For linearly parametrized classes, one can always classify with margins. Precisely: if a finite set $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ is shattered by $\mathcal{F}$, and $\mathcal{F}$ is a linear space, then the following property holds: for any possible binary assignment $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{0,1\}^n$, and for each real $\delta > 0$, there exists some function $f = f_{\varepsilon,\delta} \in \mathcal{F}$ so that $f(\mathbf{u}_i) > \delta$ if $\varepsilon_i = 1$ and $f(\mathbf{u}_i) < -\delta$

if $\varepsilon_i = 0$. Indeed, suppose that $g, h \in \mathcal{F}$ have been found such that $\mathcal{H}(g(\mathbf{u}_i)) = \varepsilon_i$ for all $i$ and $\mathcal{H}(h(\mathbf{u}_i)) = 1 - \varepsilon_i$ for all $i$. Then $(g-h)(\mathbf{u}_i) > 0$ whenever $\varepsilon_i = 1$ and $(g-h)(\mathbf{u}_i) < 0$ whenever $\varepsilon_i = 0$ (strict inequalities). Therefore, for some scalar $\rho$, $f = \rho(g-h)$ has the desired properties.

**Affine Parameterizations** If a class $\mathcal{F}$ is an affine subspace, that is to say, it has the form $\mathcal{G}+f_0 = \{g+f_0, g \in \mathcal{G}\}$, where $\mathcal{G}$ is some vector space of functions and $f_0$ is a fixed function, then $\text{VCD}\,(\mathcal{F}) = \text{VCD}\,(\mathcal{G}) = \dim \mathcal{G}$. To see this, we argue as follows. Suppose that $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ is shattered by $\mathcal{F}$, and pick any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{0,1\}^n$. Then, there exists a function $g \in \mathcal{G}$ so that $\mathcal{H}(g(\mathbf{u}_i) + f_0(\mathbf{u}_i)) = \varepsilon_i$ for all $i$. Similarly, there is a function $h \in \mathcal{G}$ so that $\mathcal{H}(h(\mathbf{u}_i)+f_0(\mathbf{u}_i)) = 1-\varepsilon_i$ for all $i$. Then, $f = g-h = (g+f_0)-(h+f_0) \in \mathcal{G}$ has the property that $\mathcal{H}(f(\mathbf{u}_i)) = \varepsilon_i$ for all $i$. This means that $S$ is also shattered by $\mathcal{G}$. Conversely, suppose that $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subseteq \mathbb{U}$ is shattered by $\mathcal{G}$ and pick any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{0,1\}^n$. The values $f_0(\mathbf{u}_i)$ are all bounded in absolute value by some $\delta > 0$. Let $g \in \mathcal{G}$ classify with margin $\delta$, i.e., $g(\mathbf{u}_i) > \delta$ if $\varepsilon_i = 1$ and $g(\mathbf{u}_i) < -\delta$ if $\varepsilon_i = 0$. It follows that $\mathcal{H}(g(\mathbf{u}_i) + f_0(\mathbf{u}_i)) = \varepsilon_i$ for all $i$. So $S$ is also shattered by $\mathcal{F}$.

Theorem 1 has several immediate applications.

### 3.1 Perceptrons

Perceptrons are just linear discriminators on $\mathbb{R}^m$. Here $\mathcal{F} = \mathcal{P}_m$ consists of all possible affine functions from $\mathbb{U} = \mathbb{R}^m$ into $\mathbb{R}$, i.e., all functions of the form:

$$f(\mathbf{u}) \;=\; f(u_1, \ldots, u_m) \;=\; a_0 + a_1 u_1 + \ldots + a_m u_m \,. \tag{12}$$

These functions are linearly parametrized by vectors $(a_0, \ldots, a_m) \in \mathbb{R}^{m+1}$, so

$$\text{VCD}\,\mathcal{P}_m \;=\; m+1 \,. \tag{13}$$

One may also fix certain of the coefficients at "prewired" values, leaving only a subset of $m' < m$ parameters free. This gives rise to an affine class of dimension $m'$, so that the VC dimension is $m'$. (For instance, take the class consisting of all functions of the form $1 + a_1 u_1 + a_2 u_2 - 3u_3$, where the $a_1$ and $a_2$ parameters are arbitrary real numbers. This is the class of all functions of the form $\mathcal{G} + f_0$, where $f_0(\mathbf{u}) = 1 - 3u_3$ and $\mathcal{G}$ is the $m' = 2$ dimensional space generated by $u_1$ and $u_2$.)

### 3.2 Single Hidden Layer Nets with Fixed Input Weights

Single hidden layer nets are described as follows. We fix an "activation" function $\sigma : \mathbb{R} \to \mathbb{R}$ and positive integers $n$ and $m$, the number of hidden units and the input dimension, respectively.

For each choice of $n$ row $m$-vectors $A_1, \ldots, A_n$ (the *input-layer weights*), $n$ scalars $b_1, \ldots, b_n$ (the *input-layer biases*), and *output-layer weights* $c_0, \ldots, c_m$, the response of such a net is

$$y \;=\; f(\mathbf{u}) \;=\; c_0 \,+\, \sum_{i=1}^{n} c_i \sigma(A_i \mathbf{u} + b_i)\,. \tag{14}$$

Writing each $A_i$ as $a_{i1}, \ldots, a_{im}$, pictorially we have a net as in Fig. 2.

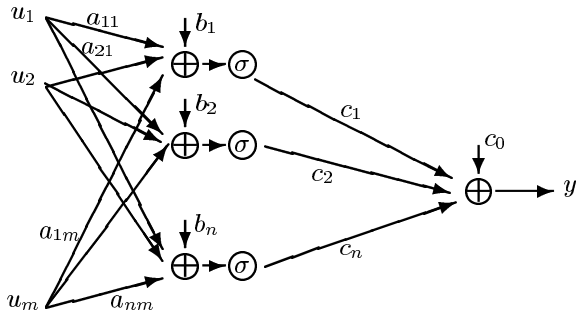


**Fig. 2.** Single-hidden layer net with activation function $\sigma$.

We now fix the input-layer weights $A_1, \ldots, A_n$ and the input-layer biases $b_1, \ldots, b_n$, and leave variable the output-layer weights $c_0, \ldots, c_m$. Let us call the obtained class of functions $\mathcal{F}_{n,\sigma,A,B}$; this is the span of the functions 1 and $\sigma(A_i \mathbf{u} + b_i)$, $i = 1, \ldots, n$, so it has dimension at most $n+1$. Therefore:

$$\text{VCD}\, \mathcal{F}_{n,\sigma,A,B} \;\leq\; n+1\,. \tag{15}$$

**When is** VCD $\boldsymbol{\mathcal{F}_{n,\sigma,A,B} = n+1}$**?** The inequality (15) may be strict, because there is no reason for the functions 1 and $\sigma(A_i \mathbf{u} + b_i)$, $i = 1, \ldots, n$ to span a space of dimension exactly $n+1$. Thus, we are led to the following question: When are the functions 1 and $\sigma(A_i \mathbf{u} + b_i)$, $i = 1, \ldots, n$ linearly independent? That is, we wish to know, for a fixed set of $A$'s and $b$'s, whether the following implication holds:

$$c_0 \,+\, \sum_{i=1}^{n} c_i \sigma(A_i \mathbf{u} + b_i) \equiv 0 \;\Rightarrow\; c_0 = \ldots = c_n = 0\,. \tag{16}$$

This implication is in general false, for example if there are two or more equal sets of weights $(A_i, b_i)$ at the input layer, or weights with opposite signs and $\sigma$ is an odd function ($\sigma(-x) = -\sigma(x)$), or if a weight $A_i$ is zero. But in these cases, one may either collect equal (or opposite) terms into a single first-layer unit, or into $c_0$ (for the $A_i = 0$). The implication is also false even

if these obvious degeneracies do not occur, for activations $\sigma$ such as periodic functions, exponentials, or polynomials:

$$\begin{aligned}\sin(1.u+2\pi)+(-1)\sin(1.u+0)&\equiv 0\\ \exp(1.u+1)-e\exp(1.u+0)&\equiv 0\\ (2.u)^2-4(u)^2&\equiv 0\,.\end{aligned}$$

However, we show next that, except in the trivial cases, independence does hold when a "standard" activation is used.

**Theorem 2.** *Let $\sigma=\tanh$. Assume that $(A_i,b_i)\neq\pm(A_j,b_j)$ for all $i\neq j$ and that $A_i\neq 0$ for all $i$. Then,* $\mathrm{VCD}\,\mathcal{F}_{n,\sigma,A,B}=\ n+1$.

*Proof.* Note that there is some vector $\hat{\mathbf{u}}\in\mathbb{R}^m$ so that, for all $i\neq j$:

1. $A_i\hat{\mathbf{u}}\neq 0$,
2. $b_i=b_j\Rightarrow(A_i-A_j)\hat{\mathbf{u}}\neq 0$, and
3. $b_i=-b_j\Rightarrow(A_i+A_j)\hat{\mathbf{u}}\neq 0$.

(Because the complement of the set of such $u$'s is a finite union of hyperplanes.) Suppose that there would exists nonzero coefficients $c_i$'s so that $f(\mathbf{u})=c_0+\sum_{i=1}^n c_i\sigma(A_i\mathbf{u}+b_i)=0$ for all $\mathbf{u}\in\mathbb{R}^m$. Letting $a_i:=A_i\hat{\mathbf{u}}$, we may reduce the problem to the scalar-input case:

$$g(u)\ :=\ f(u\hat{\mathbf{u}})\ =\ c_0\ +\ \sum_{i=1}^{n} c_i\sigma(a_iu+b_i)\ =\ 0 \tag{17}$$

for all $u\in\mathbb{R}$, and the weights satisfy $a_i\neq 0\ \forall i$ and $(a_i,b_i)\neq\pm(a_j,b_j)\ \forall i\neq j$. Without loss of generality, we may assume that $c_i\neq 0$, $i=1,\ldots n$ (otherwise, we may drop zero terms and have a sum with a smaller $n$). Similarly, we may take all $a_i>0$ (otherwise, we may reverse signs of the necessary $a_i$ and $b_i$, and take $c_i:=-c_i$. Without loss of generality, we suppose $a_1\geq a_i$, $i=2,\ldots,n$. We make a change of variables $v:=a_1u+b_1$, so that now

$$g(v)\ =\ c_0\ +\ c_1\sigma(v)\ +\ \sum_{i=2}^{n} c_i\sigma(a_i'v+b_i')=0 \tag{18}$$

for all real $v$. We have that all $0<a_i'\leq 1$, and $a_i'=1\Rightarrow b_i'\neq 0$. Thus, for all odd integers $k$, $a_i'\frac{\pi}{2}\sqrt{-1}+b_i'\neq k\frac{\pi}{2}\sqrt{-1}$, that is, $a_i'\frac{\pi}{2}\sqrt{-1}+b_i$ is not a pole of $\sigma$ when $\sigma$ is seen as a function of a complex variable. By the principle of analytic continuation, equation (18) must hold for all $v$ on the subset of $\mathbb{C}$ where none of the terms has poles (which is a connected subset containing $\mathbb{R}$). Pick a sequence $\{v_k\}$ of points where $g$ is analytic, $v_k\to\frac{\pi}{2}\sqrt{-1}$. Then, $0\equiv\frac{g(v_k)}{\sigma(v_k)}\to c_1$ implies $c_1=0$, contradiction. ∎

Theorem 2 is from Sussmann (1992) The proof given here is from Albertini, Sontag and Maillot (1993); in addition to simplicity, has the advantage of generalizing, with no changes, to other typical choices of sigmoids, such as $(1+e^{-x})^{-1}$ or $\arctan x$. Other techniques can be used as well, see the discussion in Sontag (1997a). These results can also be interpreted as establishing parameter identifiability of networks ("function determines form", see references in Albertini et al. (1993)).

A special case worth noticing is also that in which there are no lower-level biases, i.e., networks of the form $\sum_{i=1}^{n} c_i\sigma(a_i u)$. Suppose that $\sigma$ is an odd function, and it is infinitely differentiable about zero, with an infinite number of nonzero derivatives $\sigma^{(k)}(0)$ at zero. Then, if all $a_i$ are nonzero and have different absolute values, the functions $\sigma(a_i u)$ are linearly independent, so the corresponding class has VC dimension exactly $n$. To see this, one simply notices that one may take without loss of generality all $a_i > 0$ (using that $\sigma$ is odd, and changing if needed $c_i$ to $-c_i$); now $g^{(k)}(0) = \sum_{i=1}^{n} a_i^k c_i \sigma^{(k)}(a_i u) = 0$ for all $k$ implies, for those $k$ so that $\sigma^{(k)}(0) \neq 0$, that $\sum_{i=1}^{n} a_i^k c_i = 0$; a generalized Vandermonde argument then implies that the $c_i$ must vanish (see Albertini et al. (1993)). For an analytic function $\sigma$, this means that the VC dimension is $n$ for all such nets if and only if $\sigma$ is not a polynomial.

## 4 The Fundamental Fact About VC Dimension

In order to obtain other upper bounds on VC dimension, we need to review what is perhaps the single most important property of VC dimension. This result, frequently called "Sauer's Lemma," is from Vapnik and Cervonenkis (1968) (see also Vapnik (1992)), and was discovered independently by Sauer (1972) and Shelah (1972); interestingly, Sauer credits Erdös with posing it as a conjecture.

Assume given a set $\mathcal{F}$ of functions $\mathbb{U} \to \{0,1\}$. For each $m$, and each sequence $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) \in \mathbb{U}^m$, we count the number of classifications possible on the inputs in this sequence:

$$\gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) := \text{card}\,\{(f(\mathbf{u}_1), f(\mathbf{u}_2), \ldots, f(\mathbf{u}_m)) \in \{0,1\}^m \mid f \in \mathcal{F}\}\,.$$

Observe that the $m$-element set $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ is shattered if and only if $\gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$ attains its maximal possible value, namely $2^m$. In general, $\gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$ measures the number of elements of the set of functions $\mathcal{F}|_S$ consisting of the restrictions of the functions in $\mathcal{F}$ to $S$.

The important fact is that $\gamma$ grows only polynomially, instead of exponentially, on the sample length $m$, provided that the VC dimension be finite. Moreover, the degree of the polynomial is the VC dimension. For each two nonnegative integers with $m \geq d$, we define $\Phi(m,d)$ as the number of possible subsets of an $m$-element set with at most $d$ elements, that is,

$$\Phi(m,d) := \sum_{i=0}^{d} \binom{m}{i} \leq 2\frac{m^d}{d!} \leq \left(\frac{em}{d}\right)^d \tag{19}$$

(the two shown upper bounds are not difficult to establish).

**Theorem 3.** *(Vapnik-Chervonenkis-Sauer-Shelah.) Suppose that* $\text{VCD}\,(\mathcal{F}) = d < \infty$. *Then, for each* $m \geq d$ *and all sequences* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$,

$$\gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) \;\leq\; \Phi(m, d)\,. \tag{20}$$

Observe that the bound is best possible in general, since the concept class consisting of all $d$-element subsets of $\{1, \ldots, m\}$ achieves the estimate. The key for the proof is the following lemma about binary matrices.

**Lemma 1.** *Let* $m \geq 1$ *and* $0 \leq d \leq m$, *and suppose that the matrix* $C \in \{0,1\}^{m \times r}$ *is so that all its columns are distinct, where* $r$ *is an integer satisfying* $r > \Phi(m, d)$. *Then, there is some* $d+1$ *by* $2^{d+1}$ *submatrix of* $C$ *whose columns are distinct.*

Let us first see why this implies Theorem 3. Suppose that $\text{VCD}\,(\mathcal{F}) = d$, pick any sequence $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$ with $m \geq d$, and let $r = \gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$. We list all possible different classifications as columns:

$$\begin{array}{cccc} f_1(\mathbf{u}_1) & f_2(\mathbf{u}_1) & \ldots & f_r(\mathbf{u}_1) \\ f_1(\mathbf{u}_2) & f_2(\mathbf{u}_2) & \ldots & f_r(\mathbf{u}_2) \\ \vdots & \vdots & \ldots & \vdots \\ f_1(\mathbf{u}_m) & f_2(\mathbf{u}_m) & \ldots & f_r(\mathbf{u}_m) \end{array} \;. \tag{21}$$

If it were the case that $r > \Phi(m, d)$, then the Lemma says that there is a submatrix with $d+1$ rows and all possible $2^{d+1}$ columns. The corresponding $u_i$'s would then provide a subset of cardinality $d+1$ which is shattered, contradicting the fact that the maximal shattered set has size $d$. So $r \leq \Phi(m, d)$, as wanted.

*Proof of Lemma 1.* We proceed by induction on $m$. If $m = 1$, then $d = 0$ or $d = 1$. Since, $2^m \geq r > \Phi(m, d)$ and $\Phi(1, 1) = 2$, necessarily $d = 0$. So $r > \Phi(1, 0) = 1$, that is, $r = 2$. This means that $C$ is itself of size $d+1$ by $2^{d+1}$.

We next assume the result true for $m-1$, and prove it for $m$. The columns of $C$ can be rearranged, in several manners, so that $C$ has the following form:

$$\begin{array}{cccccc} 1 & \{ & 0 \cdots 0 & 1 \cdots 1 & * \ldots * \\ m-1 & \{ & A & A & B \end{array}$$

where $A$ is an $m-1$ by $r_1$ submatrix and $B$ has size $r - 2r_1$. For instance, we may take $r_1 = 0$ and no "$A$" submatrix.

Among these rearrangements, there are some for which $r_1$ is as large as possible; from now on, we assume that we have picked one such. Note that all columns of the matrix $(A\ B)$ must be distinct. (Indeed, if $A$ would have two equal columns, or if $A$ and $B$ had a column in common, then there would

be some two equal columns in $C$; if $B$ had two equal columns, then, the first entry in $C$ for the corresponding columns would have to be distinct, and thus these columns could be moved to the $A$ blocks, contradicting the maximal choice of $r_1$.)

We now claim that either:

1. $(A\ B)$ has some $d+1$ rows with $2^{d+1}$ distinct columns, or
2. $A$ has some $d$ rows with $2^d$ distinct columns.

The lemma will follow from this claim: in the first case the last $m-1$ rows already give an appropriate submatrix of $C$, and in the second case we use that

$$\begin{pmatrix} 0\cdots 0 & 1\cdots 1 \\ A & A \end{pmatrix} \tag{22}$$

has some $d+1$ rows with $2^{d+1}$ distinct columns.

So we show the claim, using the induction hypothesis. Since $2^m \geq r > \Phi(m,d)$ and $\Phi(m,m) = 2^m$, necessarily $m-1 \geq d$. We let $r_2 = r - r_1$, the number of columns of $(A\ B)$. There are two cases:

1. $r_2 > \Phi(m-1,d)$: in this case, the inductive assumption applies to the data $m-1$, $r_2$, $d$, so we have the first case of the claim; or
2. $r_2 \leq \Phi(m-1,d)$: now $r_1 = r - r_2 > \Phi(m,d) - \Phi(m-1,d) = \Phi(m-1,d-1)$, so the result applied to $m-1, r_1, d-1$ gives that $A$ has $d$ rows as wanted.

This completes the proof of the Lemma. ∎

## 5 Basic Techniques

In this section, we cover several basic techniques which are used in estimating upper bounds on VC dimension. They all use Theorem 3 as a tool.

### 5.1 Boolean Closures

We start showing how to establish upper bounds on the VC dimensions of those concept classes which arise as unions, intersections, or other Boolean operations, starting from classes whose VC dimensions have already been estimated.

Given $k$ classes of functions $\mathbb{U} \to \{0,1\}$, $\mathcal{F}_1, \ldots, \mathcal{F}_k$, and a fixed Boolean function $b : \{0,1\}^k \to \{0,1\}$. we define

$$b(\mathcal{F}_1, \ldots, \mathcal{F}_k) := \{b(f_1(\cdot), \ldots, f_k(\cdot)) \mid f_i \in \mathcal{F}_i, i = 1, \ldots, k\} . \tag{23}$$

**Lemma 2.** *With $c_k = 2k \log ek$, a constant which does not depend on the classes $\mathcal{F}_i$ nor on the Boolean function $b$,*

$$\text{VCD}\,(b(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \leq c_k \max_{i=1,\ldots,k} \{\text{VCD}\,(\mathcal{F}_i)\} . \tag{24}$$

*Proof.* We assume that $S \subseteq \mathbb{U}$ is shattered, and $\operatorname{card} S = n$. Restricting all functions to $S$, we think of each $\mathcal{F}_i$ as a set of functions from $S$ to $\{0,1\}$, and $\mathcal{F} := \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ as a set of functions $S \to \{0,1\}^k$. Since the mapping

$$\mathcal{F} \to b(\mathcal{F}_1, \ldots, \mathcal{F}_k) \ : \ f_1, \ldots, f_k \mapsto b \circ (f_1, \ldots, f_k) \tag{25}$$

is onto,

$$\operatorname{card} b(\mathcal{F}_1, \ldots, \mathcal{F}_k) \ \leq \ \operatorname{card} \mathcal{F} \ = \ \prod_i \operatorname{card} \mathcal{F}_i \, . \tag{26}$$

We assume that all the $d_i = \text{VCD}\,(\mathcal{F}_i) < \infty$ (otherwise, $d = \infty$ and there is nothing to prove). By Theorem 3,

$$\operatorname{card} \mathcal{F}_i \ \leq \ \left( \frac{en}{d_i} \right)^{d_i} \tag{27}$$

for each $i$. With $d := \max_{i=1,\ldots,k} d_i$, this gives

$$\operatorname{card} b(\mathcal{F}_1, \ldots, \mathcal{F}_k) \ \leq \ \left( \frac{en}{d} \right)^{dk} \, . \tag{28}$$

As $S$ is shattered by $\mathcal{F}$, $2^n \leq \left(\frac{en}{d}\right)^{dk}$. An easy calculus argument then gives $n < (2k \log ek)\, d$. ∎

See Dudley (1984), Pollard (1990), and Vidyasagar (1997) for more results along these lines.

**Single Hidden Layer Nets with Fixed Output Weights.** As an application of Lemma 2, we consider single hidden layer networks as in (14), but now ask what is the VC dimension when the output weights $c_i$ are constant and we vary the input weights instead. We can only apply the result to Boolean functions, so we take $\sigma = \mathcal{H}$.

The function classes $\mathcal{F}_i$ are all the same, and consist of perceptrons $\mathcal{H}(a\mathbf{u} + b)$, so we know that $d = m + 1$. So, $\text{VCD}\,(\mathcal{F}) \leq c_n(m+1)$. The same upper bound obtains if the second-level operation is a more general Boolean operation than a linear threshold, of course.

The argument just given uses that $\sigma = \mathcal{H}$ in an essential manner. When $\sigma$ is not Boolean, not only is the estimate $\text{VCD}\,(\mathcal{F}) \leq c_n(m+1)$ false, but $\text{VCD}\,(\mathcal{F})$ *may be infinite even if* $n = 2$, $m = 1$, *and* $\mathcal{F}_1 = \mathcal{F}_2$ *has VC dimension one.* To see an example of this phenomenon, consider the following activation function (cf. Sontag (1992)):

$$\sigma(u) \ := \ \frac{1}{\pi} \arctan u \, + \, \frac{1}{2} \, + \, \frac{\cos u}{\alpha(1+u^2)} \, , \tag{29}$$

where $\alpha$ is any fixed constant $> 2\pi$. The graph of $\sigma$ has a "sigmoidal" shape, with range $(0,1)$, and strictly positive derivative everywhere. The functions

of the form $f(u) = \sigma(au+b)$ are all monotonic, so the classes $\mathcal{F}_1$ and $\mathcal{F}_2$ each have VC dimension 1. However, if we take the 1-2 architecture with fixed weights $-c_0 = c_1 = c_2 = 1$, that is, the set of functions $\mathcal{F}$ of the form:

$$f(u) = -1 + \sigma(a_1 u + b_1) + \sigma(a_2 u + b_2) \tag{30}$$

(with different possible $a_i$'s and $b_i$'s), we obtain $\text{VCD}(\mathcal{F}) = \infty$. In fact, this is true even if we also restrict attention to $b_1 = b_2 = 0$ and $a_1 = -a_2 = \lambda$, where $\lambda$ is a scalar parameter. In that case, we write the corresponding function as $f_\lambda$ and observe that:

$$f_\lambda(u) = -1 + \sigma(\lambda u) + \sigma(-\lambda u) = \frac{2\cos\lambda u}{\alpha(1+\lambda^2 u^2)}\,. \tag{31}$$

To show that some set of cardinality $n$ can be shattered, for arbitrary given $n$, we pick any $n$ and any $n$ real numbers with the property that $\pi, u_1, \ldots, u_n$ are rationally independent (i.e., $r_0 + \sum r_i u_i = 0$, with the $r_i$'s integers, implies $r_0 = \cdots = r_n = 0$). It is easy to prove that a "random" choice of $(u_1, \ldots, u_n)$ has this property, with probability one, and that the set of vectors

$$\{(\lambda u_1, \ldots, \lambda u_n) \mid \lambda \in \mathbb{N}\} \tag{32}$$

modulo $2\pi$ is dense in $[0, 2\pi]^n$. Hence, also the set of vectors

$$\{(\cos(\lambda u_1), \ldots, \cos(\lambda u_n)) \mid \lambda \in \mathbb{N}\} \tag{33}$$

is dense in $[-1, 1]^n$. This implies, in particular, that $(f_\lambda(u_1), \ldots, f_\lambda(u_n))$ can be made to have any sign sequence, by choice of appropriate parameters $\lambda$, so the set $\{u_1, \ldots, u_n\}$ can indeed be shattered.

### 5.2 Compositions (Cascades)

As another application of the fundamental Theorem 3, we now estimate the VC dimension of classes of functions obtained as compositions (cascades, or series connections) of basic function classes. The basic setup is as follows. We suppose given a set $\mathcal{F}$ of functions $\mathbb{U} \to \mathbb{V}$ and a set $\mathcal{G}$ of functions $\mathbb{V} \to \mathbb{W}$, and define

$$\mathcal{G} \circ \mathcal{F} := \{g \circ f \mid g \in \mathcal{G}, f \in \mathcal{F}\} \tag{34}$$

(a set of functions $\mathbb{U} \to \mathbb{W}$). We assume given "growth functions" for each class, which bound the numbers $\gamma$ of possible classifications, that is, two functions $p$ and $q$ so that:

1. for each $S \subseteq \mathbb{U}$ with $\operatorname{card} S \leq n$, $\operatorname{card} \mathcal{F}|_S \leq p(n)$, and
2. for each $R \subseteq \mathbb{V}$ with $\operatorname{card} R \leq n$, $\operatorname{card} \mathcal{G}|_R \leq q(n)$.

The following totally elementary remark plays a central role.

**Lemma 3.** *For each* $S \subseteq \mathbb{U}$ *with* $\operatorname{card} S \leq n$, $\operatorname{card} (\mathcal{G} \circ \mathcal{F})|_S \leq p(n)q(n)$.

*Proof.* Let $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Pick a subclass $\mathcal{F}_0 = \{f_1, \ldots, f_{p(n)}\}$ of $\mathcal{F}$ such that $\mathcal{F}_0|_S = \mathcal{F}|_S$. For each $i \in \{1, \ldots, p(n)\}$, we consider the following subset of $\mathbb{V}$:

$$R_i \ := \ \{f_i(\mathbf{u}_1), \ldots, f_i(\mathbf{u}_n)\}\,. \tag{35}$$

For each such subset $R_i$, there exists, in turn, a subclass of functions $\mathcal{G}_i = \{g_1^i, \ldots, g_{q(n)}^i\}$ of $\mathcal{G}$ so that $\mathcal{G}_i|_{R_i} = \mathcal{G}|_{R_i}$. Now pick any element $g \circ f \in \mathcal{G} \circ \mathcal{F}$. Choose $i$ so that $f|_S = f_i|_S$ and $j$ so that $g|_{R_i} = g_j^i|_{R_i}$. thus $(g \circ f)|_S = (g_j^i \circ f_i)|_S$. It follows that

$$(\mathcal{G} \circ \mathcal{F})|_S \ = \ \left\{g_j^i \circ f_i \mid i = 1, \ldots, p(n)\,,\ j = 1, \ldots, q\right\}|_S\,, \tag{36}$$

and this proves the Lemma. ∎

By induction, $\operatorname{card} (\mathcal{F}_1 \circ \ldots \circ \mathcal{F}_\ell)|_S \ \leq \ p_1(n) \ldots p_\ell(n)$ if we have growth functions $p_i$ for each $i$.

**Multilayer Nets with Binary Activations.** As an application of Lemma 3, we consider multilayer networks with $\mathcal{H}$ activations. The *functions computed by* $(k-1)$*-hidden layer* nets are, by definition, those functions of the form

$$f_k \circ \ldots \circ f_1\,, \tag{37}$$

where, for each $i$,

$$f_i = (f_i^1, \ldots, f_i^{n_i})\,;\ \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i} \tag{38}$$

is a binary-vector valued perceptron: $f_i^j(\mathbf{u}) = \mathcal{H}(A_i^j \mathbf{u} + b_i^j)$. The integers $n_i$ for $i = 1, \ldots, k-1$ are the numbers of units at the $i$th level, while $n_0$ is the number of inputs $m$ and we take scalar outputs: $n_k = 1$. Note that the intermediate values are all Boolean. We may allow some of the weights (entries of $A_i^j$'s, and $b_i^j$'s) to be fixed and others to be variable, and let $\rho_{ij}$ be the number of variable weights for $f_i^j$. (If all weights are variable, $\rho_{ij} = n_{i-1} + 1$ for all $i, j$.) The total number of parameters is $\rho = \sum_{ij} \rho_{ij}$. As a consequence of Theorem 3, the number of possible functions $f_i^j$ (perceptrons) on any set of cardinality $n$ is bounded by $(\frac{en}{\rho_{ij}})^{\rho_{ij}}$, because the VC dimension of perceptrons with $\rho_{ij}$ parameters was found to be $\rho_{ij}$. Thus, letting $\mathcal{F}_i$ be the set of possible functions $f_i$, there are $\leq \prod_j (\frac{en}{\rho_{ij}})^{\rho_{ij}}$ such functions on each set of cardinality $n$. We conclude from Lemma 3 that the total number of functions that $\mathcal{F}$ can compute on a set of cardinality $n$ is bounded by:

$$\prod_i \prod_j \left(\frac{en}{\rho_{ij}}\right)^{\rho_{ij}} \ \leq \ \prod_{ij} (en)^{\rho_{ij}} \ \leq \ (en)^{\rho}\,. \tag{39}$$

Now, if there is any set $S$ of cardinality $n$ which is shattered, this would imply that $2^n \leq (en)^\rho$, from which, by an elementary argument, one concludes $n \leq 2\rho \log e\rho$, and thus:

**Theorem 4.** *The class of functions computed by multilayer neural networks with binary activations and $\rho$ weights has VC dimension $O(\rho \log \rho)$.*

This upper bound on VC dimension of multilayer nets is from Cover (1968), and was also obtained in Baum and Haussler (1989).

The bound is tight, in the sense that Maass (1994) and Sakurai (1993) showed how to obtain $\rho$-parameter classes of multilayer nets whose VC dimension is proportional to $\rho \log \rho$. (Maass gave a construction using three layers and binary inputs, while Sakurai used two layers but arbitrary real inputs.) Thus, for $\mathcal{H}$-activation feedforward nets, VCD $\approx c\rho \log \rho$.

### 5.3 Counting Weights

The arguments given for multilayer nets with binary activations cannot be generalized to real-valued activations. The reason for this is that the number of functions on a set of size $n$ is no more finite, in general.

This fact is most strikingly illustrated, perhaps, by simply adding *linear* activations to the threshold model: the VC dimension of a $\rho$-parameter class jumps from being proportional to $\rho \log \rho$ to the far larger number $\rho^2$. We discuss such nets with linear and threshold activations next. This serves the purpose of introducing a very different technique for upper bounding VC dimensions. This technique is not based on counting functions on inputs, but instead relies upon the "dual" idea of counting functions on weights.

**Multilayer Nets with Both $\mathcal{H}$ and Linear Activations.** The basic technical fact needed arose in 19th century work by L. Schläfli, see also Cover (1965). It states that the number of regions into which Euclidean space can be partitioned, using $n$ hyperplanes, does not grow like $2^n$, but only as a polynomial, $n^d$, whose degree is the dimension $d$ of the space.

Precisely, let $\Psi(n, d)$ be the largest number of regions into which $n$ hyperplanes can partition $\mathbb{R}^d$. That is, given hyperplanes $H_1, \ldots, H_n$ in $\mathbb{R}^d$, $\Psi(n, d)$ is the best possible bound on the number of connected components of the set $N = \mathbb{R}^d \setminus (H_1 \bigcup \ldots \bigcup H_n)$. (For example, a little thought shows that $\Psi(1, d) = 2$, $\Psi(n, 1) = n + 1$, $\Psi(2, 2) = 4$, and (see Fig. 3) $\Psi(3, 2) = 7$.)

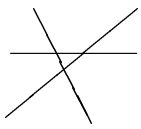

**Fig. 3.** Seven regions formed by three lines in the plane.

**Lemma 4.** *For $n \geq d$, $\Psi(n, d) \leq \Phi(n, d)$.*

*Proof.* Suppose that some $n$ hyperplanes in $\mathbb{R}^d$ define $q$ regions, and now add a new hyperplane $H$. Take any of these $q$ regions, and pick one which $H$ intersects; since the region is a convex set, it can be divided into at most two subregions. Thus, one extra region may be created due to each such intersection. On the other hand, the total number of regions intersected is at most $\Psi(n, d-1)$, the number of regions into which $\mathbb{R}^{d-1}$ can be decomposed by $n$ hyperplanes (because if two regions were different, they would make different regions as intersections with $H$, which has dimension $d-1$). In conclusion, after adding one more hyperplane, the number of regions is at most $q + \Psi(n, d-1)$. As $q \leq \Psi(n, d)$, we have, therefore:

$$\Psi(n+1, d) \ \leq \ \Psi(n, d) + \Psi(n, d-1)\,. \tag{40}$$

This inequality, valid for all $d, n$, together with the boundary conditions $\Psi(1, d) = 2$ and $\Psi(n, 1) = n + 1$ remarked earlier, and properties of combinatorial coefficients give us

$$\Psi(n, d) \ \leq \ \sum_{i=0}^{d} \binom{n}{i}\,, \tag{41}$$

as desired. ∎

Actually, the inequality in the Lemma is an equality; in fact, any $n$ hyperplanes in "general position" achieve the bound, but the inequality is enough for our purposes.

Consider now any feedforward first-order architecture with gates which are either linear or Heaviside activations (that is, a feedforward threshold network in which "skip" or "direct" connections are allowed between levels.) Such an architecture computes a parametric class of functions $\beta : \mathbb{W} \times \mathbb{U} \to \mathbb{R}$, in the sense discussed earlier, where $\beta$ can be written as a composition of linear combinations and activations $\mathcal{H}$. We omit the formal definition of such an object, which could be given in fairly obvious graph-theoretic terms (for which see, for instance, the paper Koiran and Sontag (1997)), and instead illustrate with an example.

Take the architecture that has two $\mathcal{H}$ activations and one linear function at a first level, and a linear function at the top level. This means that

$$\begin{aligned} \beta(\mathbf{w}, \mathbf{u}) = c_0 &+ c_1 \mathcal{H}(a_{11}u_1 + a_{12}u_2 + b_1) \\ &+ c_2 \mathcal{H}(a_{21}u_1 + a_{22}u_2 + b_2) + c_3 u_1 + c_4 u_2 \end{aligned} \tag{42}$$

where the parameters are given by the list

$$\mathbf{w} = (b_1, b_2, a_{11}, a_{12}, a_{21}, a_{22}, c_0, c_1, c_2, c_3, c_4)\,. \tag{43}$$

The critical observation, for this example, but also leading to a general fact, is that *if* these 6 functions:

$$\mathcal{H}(a_{11}u_1 + a_{12}u_2 + b_1)$$

$$\begin{aligned}
&\mathcal{H}(a_{21}u_1 + a_{22}u_2 + b_2)\\
&\mathcal{H}(c_0 + c_1 + c_2 + c_3u_1 + c_4u_2)\\
&\mathcal{H}(c_0 + c_1 \quad\quad + c_3u_1 + c_4u_2)\\
&\mathcal{H}(c_0 \quad\quad + c_2 + c_3u_1 + c_4u_2)\\
&\mathcal{H}(c_0 \quad\quad\quad\quad + c_3u_1 + c_4u_2)
\end{aligned}$$

happen to coincide on two given pairs $(\mathbf{w}, \mathbf{u})$ and $(\mathbf{w}', \mathbf{u}')$, *then* it must be the case that $\mathcal{H}(\beta(\mathbf{w}, \mathbf{u})) = \mathcal{H}(\beta(\mathbf{w}', \mathbf{u}'))$. Observe that the last 4 functions are the sign-functions of the input $\mathbf{u} = (u_1, u_2)$ which could potentially be computed by the "top" level function $c_0 + c_1h_1 + c_2h_2 + c_3u_1 + c_4u_1$, when taking into account all the possible combinations $(h_1, h_2)$ of binary outputs produced by the binary gates at the lower level. This observation may be summarized by the following property: $\mathcal{H}(\beta(\mathbf{w}, \mathbf{u}))$ is a Boolean function of the six Boolean functions on the above list.

The general fact, which can be proved in exactly the same manner, by introducing a Boolean function attached to each gate and each possible combination of binary outputs from the (at most $2^{g-1}$) gates in lower levels, is as follows:

**Proposition 1.** *Consider an architecture as described above, and suppose that there are a total of $g$ Heaviside gates (including one at the top level). Then, there exist $r \leq g2^{g-1}$ Boolean functions of the form*

$$Q_i(\mathbf{w}, \mathbf{u}) = \mathcal{H}(L_i(\mathbf{w}, \mathbf{u}))\,, \tag{44}$$

*where each $L_i$ is an affine function of $\mathbf{u}$ with parameters $\mathbf{w}$, and a Boolean function $b$ of $r$ arguments, such that*

$$\mathcal{H}(\beta(\mathbf{w}, \mathbf{u})) \;=\; b(Q_1(\mathbf{w}, \mathbf{u}), \ldots, Q_r(\mathbf{w}, \mathbf{u})) \tag{45}$$

*for all $(\mathbf{w}, \mathbf{u})$.*

Note that the expression for $r$ is an overly conservative estimate. For the example that we discussed above it gives $r = 12$, but 6 was enough, because gates at the first level do not get inputs from other gates. This rough estimate is all we need, however, to allow us to prove the following counterpart of Theorem 4:

**Theorem 5.** *The class of functions computed by multilayer neural networks with binary as well as linear activations and $\rho$ weights has VC dimension $O(\rho^2)$.*

*Proof.* Take any sequence of inputs $\mathbf{u}_1, \ldots, \mathbf{u}_n$, and let $s := \gamma(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. Pick parameters $\mathbf{w}_1, \ldots, \mathbf{w}_s$ so that the $s$ distinct classifications

$$\begin{array}{c}
\mathcal{H}(\beta(\mathbf{w}_1, \mathbf{u}_1)), \ldots, \mathcal{H}(\beta(\mathbf{w}_1, \mathbf{u}_n))\\
\vdots\\
\mathcal{H}(\beta(\mathbf{w}_s, \mathbf{u}_1)), \ldots, \mathcal{H}(\beta(\mathbf{w}_s, \mathbf{u}_n))
\end{array}$$

result.

The fact that these classifications are all distinct means that, for each pair $i \neq j$, there must be some $k \in \{1,\ldots,n\}$ so that $\mathcal{H}(\beta(\mathbf{w}_i,\mathbf{u}_k)) \neq \mathcal{H}(\beta(\mathbf{w}_j,\mathbf{u}_k))$. It follows therefore from Proposition 1 that there must be some index $\ell \in \{1,\ldots,r\}$ so that

$$\mathcal{H}(L_\ell(\mathbf{w}_i,\mathbf{u}_k)) \neq \mathcal{H}(L_\ell(\mathbf{w}_j,\mathbf{u}_k))\,. \tag{46}$$

Consider the set of $rn$ hyperplanes

$$H_{k\ell} \;:=\; \{\mathbf{w} \mid L_\ell(\mathbf{w},\mathbf{u}_k) = 0\} \subset \mathbb{R}^\rho\,. \tag{47}$$

The inequality (46) implies that, for each $i \neq j$ there is some $k$ and some $\ell$ such that $\mathbf{w}_i$ and $\mathbf{w}_j$, as points in the Euclidean space $\mathbb{R}^\rho$, belong to different half-spaces determined by $H_{k\ell}$ (possibly one of them being in $H_{k\ell}$ itself, but not both of them). Let us assign, to each index $i \in \{1,\ldots,s\}$, a symbol $\theta(i)$ which denotes the connected component of

$$\mathbb{R}^\rho \setminus \left(\bigcup_{k,l} H_{k\ell}\right) \tag{48}$$

in which $\mathbf{w}_i$ lies, or the hyperplane $H_{k\ell}$ in which $\mathbf{w}_i$ lies if it happens to fall in one such hyperplane.

We have just proved that the mapping that sends $i$ into $\theta(i)$ is one-to-one. Therefore, $s$ is upper bounded by the sum of $rn$ (number of hyperplanes) and $\Psi(rn,\rho)$, the number of possible connected components determined by the $rn$ hyperplanes in $\mathbb{R}^\rho$. We know, from Lemma 4 and Equation (19), that $\Psi(rn,\rho) \leq (\frac{crn}{\rho})^\rho$ for some constant $c$, so $s \leq (\frac{crn}{\rho})^\rho$. On the other hand, it also holds that $r \leq g2^{g-1}$, where $g$ is the number of gates, and an upper bound on the number of gates is the number of parameters $\rho$. We conclude that there can be at most

$$s \;\leq\; (c2^\rho n)^\rho \;\leq\; (c'n)^\rho\, 2^{\rho^2} \tag{49}$$

distinct classifications on a set of cardinality $n$, where $c'$ is some constant that does not depend on $n$ nor $\rho$. If such a set is shattered, then $s = 2^n \leq (c'n)^\rho\, 2^{\rho^2}$. It follows using a little calculus that $n = O(\rho^2)$, as desired. ∎

An upper bound of order $\rho^2$ is best possible, in the following sense: one can find a family of maps $\beta_\rho$, each realizable by a network architecture having $c\rho$ linear and threshold units (where $c$ is some constant), so that $\text{VCD}\,(\beta_\rho) = \rho^2$ for each $\rho$. We next sketch this construction, which was given in Koiran and Sontag (1997). For each $\rho \in \mathbb{N}$, consider the set of real numbers in the interval $[0,1]$ which have a binary expansion with $\rho$ digits:

$$\mathcal{B}_\rho \;:=\; \left\{ w \in \mathbb{R} \;\middle|\; w = \sum_{i=1}^{\rho} \frac{b_i}{2^i}\ ,\ b_1,\ldots,b_\rho \in \{0,1\} \right\}, \tag{50}$$

let

$$S_\rho := \{1, \ldots, \rho\}^2 , \tag{51}$$

and define $\beta_\rho : \mathbb{R}^\rho \times \mathbb{R}^2 \to \mathbb{R}$ so that, for each $w \in \mathcal{B}_\rho$ and each $(i,j) \in S$,

$$\beta_\rho ((w_1, \ldots, w_\rho), (i,j)) \; = \; i\text{th bit of } w_j . \tag{52}$$

**Claim:** $S_\rho$ is shattered by $\mathcal{F}_{\beta_\rho}$. Indeed, suppose that we write the desired binary classifications, for each element of $S_\rho$, as a matrix $M \in \{0,1\}^{\rho \times \rho}$ (the $(i,j)$th entry indicates how $(i,j)$ is to be labeled). Pick $\mathbf{w} := (w_1, \ldots, w_\rho)$, where $w_j$ is that dyadic rational whose expansion gives the bits in the $j$th column of $M$. Then, $\beta(\mathbf{w}, (i,j))$ is the $i$th bit of $w_j$, that is, the $i$th element of the $j$th column of $M$, as wanted.

It remains to show that $\beta_\rho$ can be seen as the response of network made up of linear and threshold activations with $O(\rho)$ parameters. To achieve this, we construct such a net as a cascade of 3 subnets, implementing, respectively, the following subfunctions:

1. $j \mapsto w_j$
2. $w_j \mapsto$ first $\rho$ binary fractional bits $(b_1, \ldots, b_\rho)$
3. $((b_1, \ldots, b_\rho), i) \mapsto b_i$.

These subnets are, in turn, computed as follows. The map $j \mapsto w_j$ is obtained from

$$w_1 \; + \sum_{k=2}^{\rho} (w_k - w_{k-1}) \; \mathcal{H}(j - k + 0.5) . \tag{53}$$

The second one, $w_j \mapsto$ bits $(b_1, \ldots, b_\rho)$, is obtained recursively using:

$$b_k \; = \; \mathcal{H} \left[ 2^{k-1} \left( w_j - \sum_{\ell=1}^{k-1} 2^{-\ell} b_\ell \right) - 0.5 + 2^{-(\rho+1)} \right] . \tag{54}$$

Finally, $((b_1, \ldots, b_\rho), i) \mapsto b_i$ is easy if multiplications are allowed:

$$b_1 \; + \sum_{\ell=2}^{\rho} (b_\ell - b_{\ell-1}) \; \mathcal{H}(i - \ell + 0.5) . \tag{55}$$

We cannot multiply directly, but can emulate binary multiplication via:

$$uv = \mathcal{H}(u + v - 1.5) , \tag{56}$$

so this step can also be implemented by nets using linear and threshold activations. For more details, see Koiran and Sontag (1997). (The construction in that paper was motivated by a related one, given in Goldberg and Jerrum (1995), which showed that real-number programs, in the Blum-Shub-Smale model of computation, with running time $T$ have VC dimension $\Omega(T^2)$.)

## 6 Algebraic Techniques

Recall that in general one may have infinite VC dimension, even for fairly simple function classes such as those arising from networks with two hidden units, each of which computes a simple-looking increasing function. The formula for the activation (29) used in that example shows that there is a "hidden oscillation" given by the trigonometric function which is used in its definition. If the activations that are used in network can be expressed in terms of "purely algebraic" operations, or even using exponentiation, this pathological behavior cannot arise. The mathematical techniques required in order to discuss these facts are a bit less elementary than the simple combinatorial and linear algebra tools used so far in this exposition. We briefly sketch some of them in this section.

The first general finiteness result was obtained by Stengle and Yukich (1989). It states that if $\beta$ can be defined purely in terms of polynomials, then $\text{VCD}\,(\beta) < \infty$. To formulate a precise statement, we need to employ some logic formalism.

We say that $\beta$ is *algebraic* if the inequality $\beta(\mathbf{w}, \mathbf{u}) > 0$ can be expressed entirely in terms of multiplications, additions, real constants, logical connectives, equalities and inequalities, and quantifiers, that is, if there exists a first-order formula $F$ in the theory of real numbers $\text{Th}(\mathbb{R}, +, \cdot)$ such that

$$\beta(\mathbf{w}, \mathbf{u}) > 0 \iff F(\mathbf{w}, \mathbf{u}) \text{ is true .} \tag{57}$$

As an illustration, suppose that $\sigma$ is the saturated-linear activation

$$\sigma(x) \;=\; \begin{cases} x & \text{if } |x| \leq 1 \\ \operatorname{sign} x & \text{otherwise,} \end{cases} \tag{58}$$

and

$$\beta((a,b,c,d),u) \;=\; c\,\sigma(au+b)+d\,. \tag{59}$$

This is algebraic, because the value is positive if and only if

$$(\exists\, z)\,[cz+d>0\;\&\;z=\sigma(au+b)] \tag{60}$$

where we may in turn replace "$z = \sigma(au+b)$" by:

$$\begin{aligned}[(z=1)\,\&\,(au+b>1)] \;\;&\text{or}\;\; [(-1\leq z\leq 1)\,\&\,(z=au+b)] \\ &\text{or}\;\; [(z=-1)\,\&\,(au+b<-1)]\,.\end{aligned}$$

(Note how the quantified variable $z$ appears as a hidden-unit activation.) Similarly, using instead of this $\sigma$ a rational activation such as $\sigma(x) = \frac{x}{1+|x|}$ is also allowed, since we can write "$z = \sigma(au+b)$" as $(1+|au+b|)z-(au+b)=0$ and in turn express the absolute values in terms of basic inequalities. In general, any network with polynomial, rational, or even piecewise rational activations will give rise to algebraic $\beta$.

The first main observation in this context is that it is possible to apply the Tarski-Seidenberg elimination of quantifiers theorem in order to show that one may always rewrite $F$ as a Boolean function of polynomials: there are polynomials $P_i$, $i = 1, \ldots, k$ and a Boolean function $b$ so that

$$F(\mathbf{w}, \mathbf{u}) \equiv b[\mathcal{H}(P_1(\mathbf{w}, \mathbf{u})), \ldots, \mathcal{H}(P_k(\mathbf{w}, \mathbf{u}))] \,. \tag{61}$$

(A typical example of such an elimination step for quantified real formulas is the usual discriminant for quadratic equations: "$(\exists\, z)(z^2 + wz + u = 0)$" is equivalent to "not $(4u - w^2 > 0)$".) Thus, $\mathcal{H}(\beta(\mathbf{w}, \mathbf{u})) = b(f_1^{\mathbf{w}}(\mathbf{u}), \ldots, f_k^{\mathbf{w}}(\mathbf{u}))$, where, for each $i$, $f_i^{\mathbf{w}} = \mathcal{H}(P_i(\mathbf{w}, \cdot))$. If we let $\mathcal{F}_i = \{f_i^{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^\rho\}$, we have from Lemma 2 that $\text{VCD}\,(b(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \leq c_k \max_{i=1,\ldots,k}\{\text{VCD}\,(\mathcal{F}_i)\}$, where $c_k$ is a constant which depends only on $k$.

On the other hand, and this is the second important observation, $\text{VCD}\,(\mathcal{F}_i)$ is finite, for each $i$. This is because

$$\mathcal{F}_i \subseteq \mathcal{G}_i := \{\mathcal{H} \circ P \mid P = \text{ poly of degree } \leq d_i\} \,,$$

where $d_i :=$ degree of $P_i$ on $\mathbf{u} \in \mathbb{R}^m$, and $\mathcal{G}_i$ is a vector space of finite dimension (namely, $\binom{d_i+m}{m}$). In general, it is obvious from the definition of VC dimension that if $\mathcal{F} \subseteq \mathcal{G}$ then $\text{VCD}\,(\mathcal{F}) \leq \text{VCD}\,(\mathcal{G})$. We then conclude:

**Theorem 6.** *For algebraic $\beta$,* $\text{VCD}\,(\beta) < \infty$.

All this assumes that $\beta$ is algebraic. Unfortunately, most continuous activations used in neural network applications are *not* algebraic. It would appear at first sight that the situation is then hopeless, as illustrated by the 1-2 net with infinite VC dimension given earlier. However, it turns out that it is possible to *add exponentials and still preserve finiteness*, and thus one may include the standard saturations built in terms of rational operations and $e^x$, such as tanh or the close variation $\frac{1}{1+e^{-x}}$. Even certain inverse-trig functions such as $\arctan x$ can be used as activations and still $\text{VCD}\,(\beta) < \infty$ holds. This is proved in the paper Macintyre and Sontag (1993). as an easy application of deep work in logic carried out by van den Dries, Wilkie, Khovanskii, Shelah, Laskowski, and others.

Back to the algebraic case, it turns out that one may actually obtain explicit bounds for VC dimension. Specifically, suppose that activations are piecewise polynomial (or rational), with definitions involving polynomials of degree at most $D$ in each of at most $p$ pieces. Then, Goldberg and Jerrum (1995) shows that the VC dimension of the class determined by $\beta$ is

$$O(\,\rho + (\log p + \log D)\rho^2\,) \,. \tag{62}$$

Note that this elegantly generalizes the cases of linear and $\mathcal{H}$ activations ($D = 1$ and $p = 2$) and perceptrons ($D = p = 1$). The proof, like in the "linear and $\mathcal{H}$" case, relies upon counting connected components of a set

analogous to $N = \mathbb{R}^d \setminus (\bigcup H_{k\ell})$, where now each "$H_{k\ell}$" is an algebraic set (set of zeroes of polynomial) instead of a hyperplane. Note that

$$\mathbb{R}^\rho \setminus \Big( \{P_1 = 0\} \bigcup \ldots \bigcup \{P_s = 0\} \Big) \;=\; \{\mathbf{w} \mid P_1(\mathbf{w}) P_2(\mathbf{w}) \cdots P_s(\mathbf{w}) \neq 0\} \tag{63}$$

and that the last set is a projection of

$$\{(\mathbf{w}, z) \mid \; 1 \;-\; z\, P_1(\mathbf{w}) \cdots P_s(\mathbf{w}) = 0\}\,. \tag{64}$$

As projections do not increase the number of connected components, it is in principle only required to count components of algebraic sets. Such counts can be found in work by Milnor and Thom in the early 1960s (there are upper bounds of the type $(csD)^\rho$ for sets defined by $s$ polynomials of degree $D$ in $\rho$ variables), as well as later more precise estimates due to Warren and discussed in the above reference.

For a fixed algebraic sigmoid, the Goldberg and Jerrum bound gives us a VC dimension $O(\rho^2)$. This bound is best possible, as shown in Koiran and Sontag (1997). This builds upon the quadratic lower bound explained earlier for networks made up of linear and $\mathcal{H}$ activations. The trick is to use the fact that any sigmoidal activation is – in the very appropriate words of an anonymous reviewer of Koiran and Sontag (1997) – "locally linear and globally threshold". Precisely, let us say that $\sigma$ is *sigmoidal* if there is at least one point $x_0$ where the derivative $\sigma'(x_0)$ exists and is nonzero, and also the following two limits exist and are different:

$$\lim_{x \to \infty} \sigma(x) \;\neq\; \lim_{x \to -\infty} \sigma(x) \tag{65}$$

(without loss of generality, we may take these limits as 1 and 0 respectively). Then

$$\sigma_\varepsilon(x) \;=\; \frac{\sigma(x_0 + \varepsilon x) - \sigma(x_0)}{\varepsilon \sigma'(x_0)} \;\approx\; x \tag{66}$$

and

$$\sigma(x/\varepsilon) \;\approx\; \mathcal{H}(x) \tag{67}$$

for small $\varepsilon$, which allow us to replace linear and $\mathcal{H}$ activations by instances of $\sigma$. This provides lower bounds of order $\rho^2$ for architectures that use algebraic sigmoids (such as saturation). Note that the derivative property excludes $\sigma = \mathcal{H}$, as should be the case since for threshold nets one gets a smaller VC dimension, $O(\rho \log \rho)$.

For the non-algebraic activation $\sigma = \tanh$ (and some other related ones), Karpinski and Macintyre (1997) proved, using similar counting arguments for sets defined by exponential formulas, that the VC dimension is $O(\rho^4)$. (See also Sakurai (1995).) Whether for sigmoidal nets this fourth-order bound can be decreased is still open; the only known (to the author) lower bound is of order $\rho^2$, and follows by the constructions for general sigmoidal activations discussed in the previous paragraph.

Finally, we should mention a recent result obtained by Bartlett, Maiorov and Meir (1997). It states that *if the number of layers is fixed* then the VC dimension of feedforward networks which use piecewise polynomial activation functions grows as $\rho \log \rho$.

## 7 Some Further Remarks

We close with remarks concerning other notions of shattering as well some facts about recurrent networks.

### 7.1 How Special are Shattered Sets?

We defined the VC dimension as the largest $k$ such that *some* $k$-element set is shattered. In general, one cannot shatter *all* sets of that size; for instance, half-spaces in $\mathbb{R}^2$ shatter all sets of three points which do not lie in a segment, and more generally, for perceptrons (half-spaces), one may shatter all affinely independent $(\rho+1)$-subsets, but not dependent sets. This leads however to the question: when can one shatter "generic" sets of size equal to the VC dimension?

To make this precise, we fix a concept class and let, for each $k$, $S_k$ be the subset of $\mathbb{U}^k$, possibly empty, consisting of the ordered sets of $k$ inputs that can be shattered. Note that $S_k \neq \emptyset$ if and only if $k \leq \text{VCD}$. To talk about genericity, we need to have more structure on the space of inputs, so let us assume from now on that inputs are $m$-vectors, so $S \subseteq (\mathbb{R}^m)^k = \mathbb{R}^{mk}$. We define then, a new notion of dimension:

$$\mu := \sup \left\{ k \geq 1 \mid S_k \text{ is a dense subset of } \mathbb{R}^{mk} \right\} .$$

(One could substitute the words "open dense" instead of just "dense" in this definition, with no change, as long as the class of concepts is defined in terms of continuous mappings, because in that case if a set is shattered then a small perturbation of the set also is.) This measure of classifying power was introduced in Sontag (1992), where basic properties and bounds were provided. As an example, for perceptrons we have $\mu = \text{VCD} = m+1$. Generally, any time that one has a *vector space* of analytically parametrized class of functions, if the VC dimension is $k$ then generic sets of that size can be shattered, as simply a determinant must be nonzero, so for such linear classes one gets equality of VC dimension and $\mu$. However, $\mu$ may be strictly less that the VC dimension, as shown by the class of concepts "intersections of two halfspaces in $\mathbb{R}^2$" (which we may think of as 2-1 "and" nets): there $\mu = 3$ (we cannot shatter a four-point set if one point is in the convex hull of the other three, and the set of such configurations is open), but the VC dimension is 4 (any 4 vertices of a quadrilateral will be shattered).

It was shown in Sontag (1997b) that, for parametric classes $\beta : \mathbb{R}^\rho \times \mathbb{U} \to \mathbb{R}$ defined in terms of rational functions and exponentials, and in particular for any architecture using the standard activation $\tanh(x)$ (or $1/(1+e^{-x})$):

$$\mu \leq 2\rho + 1\,. \tag{68}$$

In conclusion, while the VC dimension for neural networks may be very high (order $\rho^4$ being the best known bound), whenever $k > 2\rho+1$, shattered $k$-sets are "special".

### 7.2 VC Dimension for Dynamical Systems

For processing by dynamical systems, for which inputs are presented sequentially as opposed to in parallel, the VC dimension scales not only with the number of parameters (weights) but also with the size $k$ of the "input window" being processed. This leads to a very different set of estimates, analogous to the work in learning theory in which one studies the learning of strings of length $k$ by $n$-state automata (work by Gold, Angluin, Rivest-Shapire, and others). In order to illustrate these, we briefly discuss recurrent networks.

Feedback, or recurrent, networks, are specified by recursions such as

$$x(t+1) \;=\; \boldsymbol{\sigma}\left(Ax(t) + Bu(t)\right)\,, \quad t = 0, 1, 2, \ldots$$

where $x(t)$ is the $n$-vector of states at time $t$, and $u(t)$ is the scalar input at time $t$, $\boldsymbol{\sigma}(z_1, \ldots, z_n) := (\sigma(z_1), \ldots, \sigma(z_n))$ represents the componentwise application of a scalar nonlinearity $\sigma$ (the activation), and the matrices $A$ and $B$ encode the weights or parameters defining the system. One may also consider the initial state $x(0)$ as a parameter. Thus there are $\rho = n^2 + 2n$ weights, in the $n \times n$ matrix $A$, the $n$ vector $B$, and the initial state vector $x(0)$. We think as one coordinate of $x(t)$, let us say the first one, as indicating the output. It is also possible to consider continuous-time (differential equation) models, of course. See Sontag (1997a) for an introduction to the subject and many more details. For each fixed dynamic order (dimension) $n$ and fixed activation $\sigma$, and we may introduce a *family* of concepts: for each input length $k$, there is the class $\mathcal{P}_{n,k}$ of input sequences which lead to a positive output at time $k+1$.

The results in Dasgupta and Sontag (1996) and Koiran and Sontag (1998) show that, in rough terms (see the papers for the precise statements):

1. when $\sigma = \mathcal{H}$, $\text{VCD}\,(\mathcal{P}_{n,k}) \approx p(n) \log k$, for some polynomial $p$;
2. when $\sigma$ is the identity, $\text{VCD}\,(\mathcal{P}_{n,k}) \approx p(n) \log k$;
3. for sigmoidal piecewise polynomial activations, $\text{VCD}\,(\mathcal{P}_{n,k}) \approx p(n)k$.

For the sigmoid $1/(1+e^{-z})$, the bounds known are between orders $k$ and $k^2$ on the input length $k$. See also Sontag (1998) for a continuous-time result.

**Acknowledgements** The author wishes to acknowledge Akito Sakurai and Pirkko Kuusela for a careful reading of the manuscript, and for many useful comments. The research supported here was supported in part by US Air Force Grant F49620-97-1-0159.

*The following web site:*

**http://www.math.rutgers.edu/~sontag/**

*contains several of the preprints by the author cited here, as well as much related material.*

## References

Albertini, F., Sontag, E. and Maillot, V.: 1993, Uniqueness of weights for neural networks, *in* R. Mammone (ed.), *Artificial Neural Networks for Speech and Vision*, Chapman and Hall, London, pp. 115–125.

Anthony, M. and Bartlett, P.: n.d., A theory of learning in artificial neural networks, to appear.

Bartlett, P., Maiorov, V. and Meir, R.: 1997, Almost linear vc dimension bounds for piecewise polynomial networks, preprint, Technion Department of Electrical Engineering.

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

Cover, T.: 1965, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Computers* **EC-14**, 326–334. Reprinted in *Artificial Neural Networks: Concepts and Theory*, IEEE Computer Society Press, Los Alamitos, Calif., 1992, P. Mehra and B. Wah, eds.

Cover, T.: 1968, Capacity problems for linear machines, *in* L. Kanal (ed.), *Pattern Recognition*, Thompson, pp. 283–289.

Dasgupta, B. and Sontag, E.: 1996, Sample complexity for learning recurrent perceptron mappings, *IEEE Trans. Inform. Theory* **42**, 1479–1487. Summary in *Advances in Neural Information Processing Systems 8 (NIPS95)* (D.S. Touretzky, M.C. Moser, and M.E. Hasselmo, eds.), MIT Press, Cambridge, MA, 1996, pp204-210.

Dudley, R.: 1984, *A Course on Empirical Processes,* École d'été de probabilités de Saint-Flour, XII—1982, 1–142, Vol. 1097 of *Lecture Notes in Math.*, Springer, Berlin-New York.

Goldberg, P. and Jerrum, M.: 1995, Bounding the vapnik-chervonenkis dimension of concept classes parametrized by real numbers, *Machine Learning* **18**, 131–148.

Haussler, D.: 1992, Decision theoretic generalizations of the pac model for neural nets and other learning applications, *Information and Computation* **100**, 78–150.

Karpinski, M. and Macintyre, A.: 1997, Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks, *J. Computer Sys. Sci.* **54**, 169–176. Summary in "Polynomial bounds for VC dimension of sigmoidal neural networks," in *Proc. 27th ACM Symposium on Theory of Computing, 1995*, pp200-208.

Koiran, P. and Sontag, E.: 1997, Neural networks with quadratic vc dimension, *J. Computer Sys. Sci.* **54**, 190–198. Summary in *Advances in Neural Information Processing Systems 8 (NIPS95)* (D.S. Touretzky, M.C. Moser, and M.E. Hasselmo, eds.), MIT Press, Cambridge, MA, 1996, pp197-203.

Koiran, P. and Sontag, E.: 1998, Vapnik-chervonenkis dimension of recurrent neural networks, *Discrete Applied Math.* to appear.

Maass, W.: 1994, Perspectives of current research about the complexity of learning in neural nets, *in* V. Roychowdhury, K. Siu and A. Orlitsky (eds), *Theoretical Advances in Neural Computation and Learning*, Kluwer Acedemic Publishers, pp. 295–336.

Macintyre, A. and Sontag, E.: 1993, Finiteness results for sigmoidal 'neural' networks, *Proc. 25th Annual Symp. Theory Computing*, San Diego, pp. 325–334.

Pollard, D.: 1990, *Empirical Processes: Theory and Applications*, Vol. 2 of *NSF-CBMS Regional Conf. Series in Probability and Statistics*, American Statistical Association, Alexandria, VA.

Sakurai, A.: 1993, Tighter bounds of the vc-dimension of three-layer networks, *Proc. World Congress on Neural Networks*, pp. 540–543.

Sakurai, A.: 1995, Polynomial bounds for the vc-dimension of sigmoidal, radial-basis function, and sigma-pi networks, *Proc. World Congress on Neural Networks*, pp. 58–63.

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

Shelah, S.: 1972, A combinatorial problem: Stability and order for models and theories in infinitary languages, *Pacific Journal of Math.* **41**, 247–261.

Sontag, E.: 1992, Feedforward nets for interpolation and classification, *J. Comp. Syst. Sci.* **45**, 20–48.

Sontag, E.: 1997a, Recurrent neural networks: Some systems-theoretic aspects, *in* K. W. M. Karny and V. Kurkova (eds), *Dealing with Complexity: a Neural Network Approach*, Springer-Verlag, London, pp. 1–12.

Sontag, E.: 1997b, Shattering all sets of $k$ points in 'general position' requires $(k-1)/2$ parameters, *Neural Computation* **9**, 337–348.

Sontag, E.: 1998, A learning result for continuous-time recurrent neural networks, *Systems and Control Letters* **34**, 151-158.

Stengle, G. and Yukich, J.: 1989, Some new vapnik-chervonenkis classes, *The Annals of Statistics* **14**, 1441–1446.

Sussmann, H.: 1992, Uniqueness of the weights for minimal feedforward nets with a given input-output map, *Neural Networks* **5**, 589–593.

Vapnik, V.: 1992, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, Berlin.

Vapnik, V. and Cervonenkis, A. J.: 1968, The uniform convergence of frequencies of the appearance of events to their probabilities, *Dokl. Akad. Nauk SSSR.* in Russian.

Vidyasagar, M.: 1997, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer, London.