In: *Neural Computation* **14**(4), pages 919–956, 2002.

# Neural Networks with Local Receptive Fields and Superlinear VC Dimension

Michael Schmitt

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

**Abstract**

Local receptive field neurons comprise such well-known and widely used unit types as radial basis function neurons and neurons with center-surround receptive field. We study the Vapnik-Chervonenkis (VC) dimension of feedforward neural networks with one hidden layer of these units. For several variants of local receptive field neurons we show that the VC dimension of these networks is superlinear. In particular, we establish the bound $\Omega(W \log k)$ for any reasonably sized network with $W$ parameters and $k$ hidden nodes. This bound is shown to hold for discrete center-surround receptive field neurons, which are physiologically relevant models of cells in the mammalian visual system, for neurons computing a difference of Gaussians, which are popular in computational vision, and for standard radial basis function (RBF) neurons, a major alternative to sigmoidal neurons in artificial neural networks. The result for RBF neural networks is of particular interest since it answers a question that has been open for several years. The results also give rise to lower bounds for networks with fixed input dimension. Regarding constants all bounds are larger than those known thus far for similar architectures with sigmoidal neurons. The superlinear lower bounds contrast with linear upper bounds for single local receptive field neurons also derived here.

# 1 Introduction

The receptive field of a neuron is the region of the input domain giving rise to stimuli to which the neuron responds by changing its behavior. Neuron models can be classified according to whether these stimuli are contained in some bounded region or may come from afar, in other words, whether their receptive field is local or not. Prominent examples of local receptive field models are neurons with center-surround receptive field and neurons computing radial basis functions, whereas sigmoidal neurons represent a widely used model type having a non-local receptive field. The impressive computational and learning capabilities of neural networks, being significantly higher than those of single neurons, are well established by experimental findings in biology, by innumerable successful applications in practice, and by substantial formal arguments in the theories of computation, approximation, and learning. An evident question is to what extent these network capabilities depend on the receptive field type of the neurons.

An extensively studied measure for quantifying the computational and learning capabilities of formal systems is the Vapnik-Chervonenkis (VC) dimension. It characterizes the expressiveness of a neural network and is mostly given in terms of the number of network parameters and the network size. A well known fact is that the VC dimension of sigmoidal neural networks is significantly more than linear. This has been shown for networks growing in depth (Koiran and Sontag, 1997; Bartlett et al., 1998), and for constant-depth networks with the number of hidden layers being two (Maass, 1994) and one (Sakurai, 1993). The two latter results for networks of constant depth are of particular significance since they deal with neural architectures as they are used in practice, where one rarely allows the number of hidden layers to grow indefinitely. Moreover, the case of one hidden layer is of even greater interest, because single neurons almost always have a VC dimension that is linear in the input dimension and, hence, linear in the number of model parameters.[1] This has been found for sigmoidal neurons (Cover, 1965; Haussler, 1992) (see also Anthony and Bartlett, 1999) and for several other models such as higher-order sigmoidal neurons with restricted degree (Anthony, 1995), for the neuron computing Boolean monomials (Natschläger and Schmitt, 1996), and for the product unit (Schmitt, 2000). Thus, the fact that networks with the minimal number of one hidden layer have superlinear VC dimension corroborates the enormous computational capabilities arising when sigmoidal neurons cooperate in networks.

In this article we study networks with one hidden layer of local receptive field neurons. We show for several types of receptive fields that the VC dimension

[1] A notable exemplar of a single neuron having superlinear VC dimension is the model of a spiking neuron studied by Maass and Schmitt (1999).

of these networks is superlinear. First, we consider discrete models of cells with center-surround receptive field (CSRF). The first real neurons to be identified as having a receptive field with center-surround organization are the ganglion cells in the visual system of the cat. The recording experiments of Kuffler (1953) from the optic nerve revealed the pure on- and off-type responses within specific areas of ganglion cell receptive fields and their concentric, antagonistic center-surround organization. Also other neurons of the mammalian visual system such as the bipolar cells of the retina and cells in the lateral geniculate nucleus are known to have center-surround receptive fields (see, e.g., Tessier-Lavigne, 1991; Nicholls et al., 1992). CSRF neurons play an important role in algorithmic experiments with self-organizing networks. The question how center-surround receptive fields can emerge in artificial networks by adjusting their parameters has been investigated using unsupervised (Linsker, 1986, 1988; Atick and Redlich, 1993; Schmidhuber et al., 1996) as well as supervised learning mechanisms (Joshi and Lee, 1993; Yasui et al., 1996). It is found that cells similar to those of the first few stages of the mammalian visual system develop when applying simple learning rules and using training data from realistic visual scenes. Neural networks with center-surround receptive fields have also been fabricated in analog VLSI hardware. The silicon retinae constructed by Mead and Mahowald (1988) (see also Mead, 1989), Ward and Syrzycki (1995), and Liu and Boahen (1996) consist of neuromorphic cells performing operations of biological receptive fields.

The second type of receptive field neuron we consider is called the difference-of-Gaussians (DOG) neuron and is a continuous version of the above model. Like this, it also has its origin in neurobiology. Extending the work of Kuffler (1953), Rodieck (1965) has probably been the first to introduce the difference-of-Gaussians as a quantitative model for the functional responses of ganglion cells. The importance and the physiological plausibility of the DOG for satisfactorily fitting functions to experimental data from retinal ganglion cell recordings is also demonstrated by the work of Enroth-Cugell and Robson (1966). The DOG is generally accepted as a mathematical description of the behavior of several cell types in the retino-cortical pathway,[2] such as the above-mentioned bipolar cells, ganglion cells, and cells in the lateral geniculate nucleus (Marr and Hildreth, 1980; Marr, 1982; Glezer, 1995). Models based on DOG functions may even provide better descriptions of experimental data than other common models of visual processing as shown by Hawken and Parker (1987) in their study of the monkey primary visual cortex.

The third and last type of local receptive field neuron in this study is the

[2]In particular, Marr and Hildreth (1980) prove that under certain conditions the difference-of-Gaussian operator closely approximates the Laplacian-of-Gaussians, also known as Marr filter, which they show to be well suited to detect intensity changes and, especially, edges in images.

radial basis function (RBF) neuron, specifically the standard, that is, Gaussian RBF neuron. RBF networks are among the major neural network types used in practice (see, e.g., Bishop, 1995; Ripley, 1996). They are appreciated because of their powerful capabilities in function approximation and learning that are also theoretically well founded. A series of papers specifically deals with showing that under rather mild conditions RBF networks can uniformly approximate continuous functions on compact domains arbitrarily closely (Hartman et al., 1990; Park and Sandberg, 1991, 1993; Mhaskar, 1996). Even before RBF networks were considered as artificial neural networks they were a well established method for multi-variable interpolation and function approximation. A comprehensive account on the approximation theory of radial basis functions up to 1990 is given by Powell (1992). The connections between approximation theory and learning in adaptive networks of RBF neurons were begun to be explored by Broomhead and Lowe (1988) and Poggio and Girosi (1990). Moody and Darken (1989) study learning algorithms for RBF networks that can be implemented as real-time adaptive systems. They show that combined supervised and unsupervised learning methods can be computationally faster in RBF networks than the gradient-based methods devised for sigmoidal networks. The reader may consult the volumes by Howlett and Jain (2001a,b) or Yee and Haykin (2001) for recent developments in RBF neural networks.

There has been previous work on the VC dimension of radial basis function networks. Bartlett and Williamson (1996) show that the VC dimension and the related pseudo dimension of radial basis function networks with discrete inputs is $O(W \log(WD))$, where $W$ is the number of network parameters and the inputs take on values from $\{-D, \dots, D\}$. The best known upper bound for RBF networks with unconstrained inputs is due to Karpinski and Macintyre (1997) and is $O(W^2k^2)$, where $k$ denotes the number of network nodes. Holden and Rayner (1995) address the generalization capabilities of networks having RBF units with fixed parameters and establish a linear upper bound on the VC dimension. Anthony and Holden (1994) consider fully adaptable RBF networks with adjustable hidden and output node parameters. Referring to the lower bound $\Omega(W \log W)$ for sigmoidal networks due to Maass (1994) they write

> " ... we leave as an open question whether it is possible to obtain a lower bound similar to that recently proved by Maass (1993, 1994) for certain feedforward networks ... "

(p. 104). The work of Erlich et al. (1997) together with a result of Lee et al. (1995) gives a linear lower bound (see also Lee et al., 1997). Although there exists already a large collection of VC dimension bounds for neural networks, it has not been known thus far whether the VC dimension of RBF neural networks is superlinear. Major reasons for this might be that previous results establishing superlinear bounds are based on methods geared to sigmoidal[3] neurons or consider networks having an unrestricted number of layers[4] (Sakurai, 1993; Maass, 1994; Koiran and Sontag, 1997; Bartlett et al., 1998).

In this article we prove that the VC dimension of RBF networks is indeed superlinear, thus answering the question of Anthony and Holden (1994) quoted above. Precisely, we show that every network with $n$ input nodes, $W$ parameters, and one hidden layer of $k$ RBF neurons, where $k \leq 2^{(n+2)/2}$, has VC dimension[5] $\Omega(W \log k)$. Thus, the cooperative network effect observed in sigmoidal networks is also existent in RBF networks. This result also has implications for the complexity of learning with RBF networks, all the more since it entails the same lower bound for the related notions of pseudo dimension and fat-shattering dimension. We do not state these consequences explicitly here but refer the reader to Anthony and Bartlett (1999) instead. Before establishing the lower bound for RBF networks, however, we show that the bound $\Omega(W \log k)$ holds for the VC dimension of DOG networks. From this the bound for RBF networks is then immediately obtained. The result for DOG networks, in turn, is derived from the superlinear lower bound for discrete CSRF networks that we establish first. Thus, this work creates a link between these three neuron models not only by focussing on their the common receptive field property but also by the logical requisite of the successive proofs of the VC dimension bounds.

The article is organized as follows. We introduce definitions and notation in Section 2. The two subsequent sections contain the derivations of the superlinear lower bounds. In Section 3 we consider networks of discrete local receptive field neurons, specifically the binary and ternary CSRF neuron and a discrete variant of the RBF neuron, the binary RBF neuron. In Section 4 we study networks of DOG neurons and standard RBF neural networks. We note that all bounds derived in both these sections have larger constant factors than those known for sigmoidal networks of constant depth thus far. In particular, we obtain the bound $(W/5)\log(k/4)$ for binary CSRF neurons and for DOG neurons, and the bound $(W/12)\log(k/8)$ for ternary CSRF neurons, for binary RBF neurons, and for Gaussian RBF neurons. For comparison, sigmoidal networks are known with one hidden layer and VC dimension at least $(W/32)\log(k/4)$, and with

---

[3]For a quite general definition of a sigmoidal neuron that does not capture radial basis function neurons see, e.g., Koiran and Sontag (1997).

[4]We point out that it might be possible to obtain superlinear lower bounds for local receptive field networks pursuing the approaches of Koiran and Sontag (1997) and Bartlett et al. (1998), but only at the expense of allowing arbitrary depth. In particular, this has no relevance for standard RBF networks.

[5]Note that this result also gives rise to the lower bound $\Omega(W \log W)$ by choosing a network with $k = n$ hidden units.

two hidden layers and VC dimension at least $(W/132)\log(k/16)$ (see Anthony and Bartlett, 1999, Section 6.3). The results in Sections 3 and 4 also give rise to lower bounds for local receptive field networks when the input dimension is fixed. In Section 5 we present upper and lower bounds for single neurons. In particular, we show that the VC dimension of binary RBF and CSRF neurons and of ternary CSRF neurons is linear. Further, we derive such a result for the pseudo dimension of the Gaussian RBF neuron. Finally, in Section 6 we return to networks of discrete local receptive field neurons and establish the upper bound $O(W \log k)$ for all discrete variants of local receptive field neurons considered here. This implies that the lower bounds for these networks are asymptotically optimal. We conclude with Section 7 discussing the results and presenting some open questions. An appendix gives the derivation of a bound for a specific class of functions defined in terms of halfspaces and required for a result in Section 5.1.

# 2 Definitions

We first introduce the types of neurons and networks with local receptive fields that we study. Then we give the definitions of the VC dimension and other basic concepts that are needed in the following.

## 2.1 Networks of Local Receptive Field Neurons

We start with discrete neurons. Let $||u||$ denote the Euclidean norm of vector $u$. A *binary center-surround receptive field (CSRF) neuron* computes the function $g_{\mathrm{bCSRF}} : \mathbb{R}^{2n+2} \to \{0,1\}$ defined as

$$g_{\mathrm{bCSRF}}(c,a,b,x) \quad = \quad \begin{cases} 1 & \text{if } a \leq ||x-c|| \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

with input variables $x_1, \ldots, x_n$, and parameters $c_1, \ldots, c_n, a, b$, where $b > a > 0$. The vector $(c_1, \ldots, c_n)$ is called the *center* of the neuron, and $a, b$ are its *center radius* and *surround radius*, respectively. We also refer to this neuron as *binary off-center on-surround neuron* and call for given parameters $c, a, b$ the set $\{x : g_{\mathrm{bCSRF}}(c,a,b,x) = 1\}$ the *surround region* of the neuron.

A *ternary CSRF neuron* is defined by means of the function $g_{\mathrm{tCSRF}} : \mathbb{R}^{2n+2} \to \{-1,0,1\}$ with

$$g_{\mathrm{tCSRF}}(c,a,b,x) \quad = \quad \begin{cases} 1 & \text{if } a \leq ||x-c|| \leq b, \\ -1 & \text{if } ||x-c|| < a, \\ 0 & \text{otherwise.} \end{cases}$$

This neuron is also called *ternary off-center on-surround neuron.*

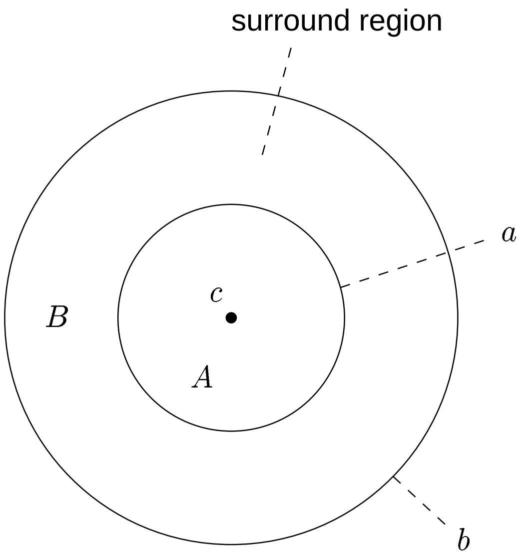


Figure 1: Receptive field of discrete neurons with center $c$, center radius $a$, and surround radius $b$. Output values $A = 0, B = 1$ correspond to a binary CSRF neuron, $A = -1, B = 1$ yields a ternary CSRF neuron, and a binary RBF neuron is given by $A = B = 1$. Outside the regions labelled $A$ or $B$ the output is always 0.

Finally, a *binary radial basis function (RBF) neuron* computes the function $g_{\mathrm{bRBF}} : \mathbb{R}^{2n+1} \to \{0, 1\}$ satisfying

$$g_{\mathrm{bRBF}}(c, b, x) = \begin{cases} 1 & \text{if } \|x - c\| \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 1 shows the receptive field of these neurons for the case $n = 2$. We emphasize that the output values in these definitions are meant to be symbolic and represent discrete levels of neural activity. So, a 1 corresponds to a state where the neuron is highly active, whereas $-1$ indicates low activity. The value 0 signifies that the neuron is silent. Furthermore, the specific assignment of values to the activity levels is not relevant for the results derived in this article. For instance, any two distinct non-zero values $A < B$ instead of $-1, 1$ can be chosen for the ternary CSRF neuron without affecting the validity of the lower bounds. The same holds for the binary CSRF and binary RBF neuron, where the value 1 can be replaced by any other non-zero value.

We further remark that the assignment of output values to points lying on a radius also allows some freedom. For instance, we could alternatively require that for $\|x - c\| = a$ we have $g_{\mathrm{bCSRF}}(c, a, b, x) = 0$ or $g_{\mathrm{tCSRF}}(c, a, b, x) = -1$. Similarly for $\|x - c\| = b$. The VC dimension bounds do not rely on the values for the radii and hence still hold for these and other cases.

We have defined here only off-center on-surround variants of CSRF neurons. In neurobiology models of on-center off-surround cells are equally important (see, e.g., Tessier-Lavigne, 1991; Nicholls et al., 1992). In these neurons the activity in the surround region is lower than in the center. Since we are considering networks that are weighted combinations of neurons, such a definition is redundant here. An on-center off-surround neuron in a network can be replaced by an off-center on-surround neuron, and vice versa, by simply multiplying the weight outgoing from it with a negative number.

The two types of continuous local receptive field neurons considered in this article are defined as follows. A *Gaussian radial basis function neuron* computes the function $g_{\mathrm{RBF}} : \mathbb{R}^{2n+1} \to \mathbb{R}$ defined as

$$g_{\mathrm{RBF}}(c, \sigma, x) \quad = \quad \exp\left(-\frac{\|x - c\|^2}{\sigma^2}\right),$$

with input variables $x_1, \ldots, x_n$, and parameters $c_1, \ldots, c_n$ and $\sigma > 0$. Here $(c_1, \ldots, c_n)$ is the *center* and $\sigma > 0$ the *width*.

A *difference-of-Gaussians (DOG) neuron* is defined as a function $g_{\mathrm{DOG}} : \mathbb{R}^{2n+4} \to \mathbb{R}$ computed by the weighted difference of two RBF neurons with equal centers, that is,

$$g_{\mathrm{DOG}}(c, \sigma, \tau, \alpha, \beta, x) \quad = \quad \alpha g_{\mathrm{RBF}}(c, \sigma, x) - \beta g_{\mathrm{RBF}}(c, \tau, x),$$

where $\sigma, \tau > 0$. Examples of $g_{\mathrm{RBF}}$ and $g_{\mathrm{DOG}}$ for input dimension 2 are shown in Figure 2.

The *neural networks* we are studying are of the *feedforward* type and have one *hidden layer*. They compute functions of the form $f : \mathbb{R}^{W+n} \to \mathbb{R}$, where $W$ is the number of network parameters, $n$ the number of input nodes, and $f$ is defined as

$$f(w, y, x) \quad = \quad w_0 + w_1 h_1(y, x) + \cdots + w_k h_k(y, x).$$

The $k$ *hidden nodes* may compute any of the functions defined above, that is,

$$h_1, \ldots, h_k \in \{g_{\mathrm{bCSRF}}, g_{\mathrm{tCSRF}}, g_{\mathrm{bRBF}}, g_{\mathrm{RBF}}, g_{\mathrm{DOG}}\}.$$

The parameters of the hidden nodes are gathered in $y$ from which each node selects its own parameters. The network has a linear output node with parameters $w_0, \ldots, w_k$ also known as the *output weights*. The parameter $-w_0$ is also called the *output threshold*. For simplicity we sometimes refer to all network parameters as weights. If $h_i = g_{\mathrm{RBF}}$ for $i = 1, \ldots, k$, we have the standard form of a Gaussian radial basis function neural network.

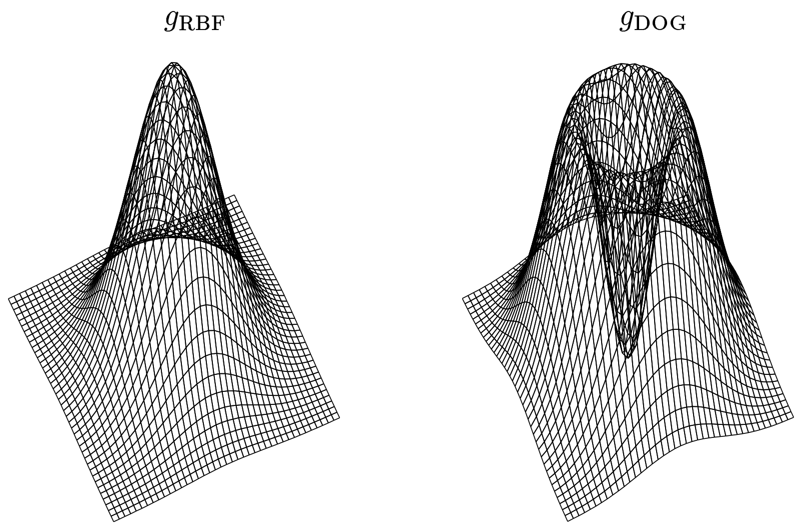


Figure 2: Receptive field functions computed by a Gaussian radial basis function neuron (left) and a difference-of-Gaussians neuron (right).

## 2.2 Vapnik-Chervonenkis Dimension of Neural Networks

A *dichotomy* of a set $S \subseteq \mathbb{R}^n$ is a pair $(S_0, S_1)$ of subsets such that $S_0 \cap S_1 = \emptyset$ and $S_0 \cup S_1 = S$. A class $\mathcal{F}$ of functions mapping $\mathbb{R}^n$ to $\{0,1\}$ is said to *shatter* $S$ if every dichotomy $(S_0, S_1)$ of $S$ is *induced* by some $f \in \mathcal{F}$, in the sense that $f$ satisfies $f(S_0) \subseteq \{0\}$ and $f(S_1) \subseteq \{1\}$. The function $\mathrm{sgn} : \mathbb{R} \to \{0,1\}$ satisfies $\mathrm{sgn}(x) = 1$ if $x \geq 0$, and $\mathrm{sgn}(x) = 0$ otherwise.

**Definition 1.** *Let $\mathcal{N}$ be a neural network and $\mathcal{F}$ be the class of functions computed by $\mathcal{N}$. The* Vapnik-Chervonenkis (VC) dimension *of $\mathcal{N}$ is the cardinality of the largest set shattered by the class $\{\mathrm{sgn} \circ f : f \in \mathcal{F}\}$.*

The pseudo dimension and the fat-shattering dimension are generalizations of the VC dimension that apply in particular to real valued function classes. The lower bounds for local receptive field networks presented in this article are stated for the VC dimension, but they also hold for the pseudo dimension and the fat-shattering dimension. The bound on the pseudo dimension follows from the fact that the VC dimension of a neural network is by definition not larger than its pseudo dimension. The bound on the fat-shattering dimension is implied because the output weights of the neural networks considered here can be scaled arbitrarily. The definition of the pseudo dimension will be given in Section 5.2 where we establish a linear upper bound for the single Gaussian RBF neuron. We refer the reader to Anthony and Bartlett (1999) for a definition of the fat-shattering dimension and results about the relationship between these three notions of dimension.

### 2.3 Further Concepts and Notation

An $(n-1)$-dimensional *hyperplane* in $\mathbb{R}^n$ is represented by a vector $(w_0, \ldots, w_n) \in \mathbb{R}^{n+1}$ and defined as the set

$$\{x \in \mathbb{R}^n : w_0 + w_1 x_1 + \cdots + w_n x_n = 0\}.$$

An $(n-1)$-dimensional *hypersphere* in $\mathbb{R}^n$ is given by a center $c \in \mathbb{R}^n$ and a radius $r > 0$, and defined as the set

$$\{x \in \mathbb{R}^n : \|x - c\| = r\}.$$

We clearly distinguish the hypersphere from a *ball*, which is defined as the set

$$\{x \in \mathbb{R}^n : \|x - c\| \leq r\}.$$

We also consider hyperplanes and hyperspheres in $\mathbb{R}^n$ with a dimension $k < n-1$. In this case, a $k$-dimensional hyperplane is the intersection of two $(k+1)$-dimensional hyperplanes, assuming that the intersection is non-empty. Similarly, the non-empty intersection of two $(k+1)$-dimensional hyperspheres yields a $k$-dimensional hypersphere, provided that the intersection is not a single point.

We use "ln" to denote the natural logarithm and "log" for the logarithm to base 2.

## 3 Superlinear Lower Bounds for Networks of Discrete Neurons

In this section we establish superlinear lower bounds for networks consisting of discrete versions of local receptive field neurons. Crucial is the result for binary CSRF networks, presented in Section 3.2, from which the bounds for ternary CSRF networks and binary RBF networks in Sections 3.3 and 3.4, respectively, follow straightforward. First, however, we introduce a geometric property of certain finite sets of points.

### 3.1 Geometric Preliminaries

**Definition 2.** *A set $S$ of $m$ points in $\mathbb{R}^n$ is said to be* in spherically general position *if the following two conditions are satisfied:*

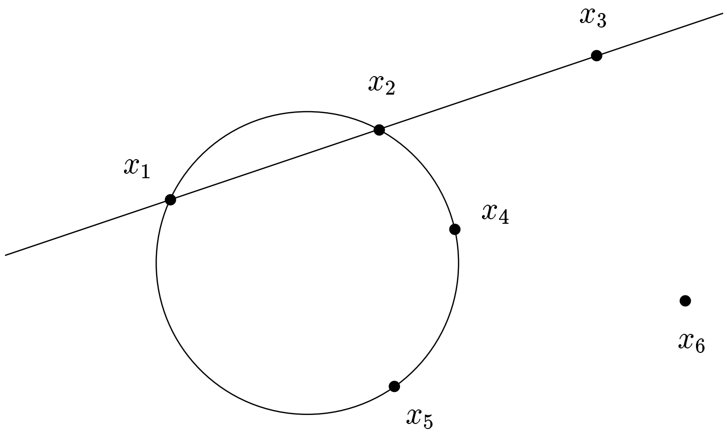


Figure 3: Positive and negative examples for sets in spherically general position (Definition 2). The set $\{x_1, x_2, x_3\}$ has a line passing through its points, and the set $\{x_1, x_2, x_4, x_5\}$ lies on a cirlce. Hence, any set that includes one of these sets (or both) is not in spherically general position since they violate condition (1) and (2), respectively. A positive example is the set $\{x_2, x_3, x_4, x_5, x_6\}$.

*(1) For every $k \leq \min(n, m-1)$ and every $(k+1)$-element subset $P \subseteq S$, there is no $(k-1)$-dimensional hyperplane containing all points in $P$.*

*(2) For every $l \leq \min(n, m-2)$ and every $(l+2)$-element subset $Q \subseteq S$, there is no $(l-1)$-dimensional hypersphere containing all points in $Q$.*

The definition is illustrated by Figure 3 showing six points in $\mathbb{R}^2$. The entire set is not in spherically general position, as witnessed by the line and the circle. It is easy, but may take a while, to verify that the set $\{x_2, x_3, x_4, x_5, x_6\}$ indeed is in spherically general position.

Sets satisfying only condition (1) are commonly referred to as being "in general position" (see, e.g., Cover, 1965; Nilsson, 1990). Thus, a set in spherically general position is particularly in general position. (The converse does not always hold, as can bee seen from Figure 3: The set $\{x_1, x_2, x_4, x_5\}$ meets condition (1) but not condition (2).) For establishing the superlinear lower bounds on the VC dimension we require sets in spherically general position with sufficiently many elements. It is easy to show that for any dimension $n$ there exist arbitrarily large such sets. The proof of the following proposition provides a method for constructing them.

**Proposition 1.** *For every $n, m \geq 1$ there exists a set $S \subseteq \mathbb{R}^n$ of $m$ points in spherically general position.*

*Proof.* We perform induction on $m$. Clearly, every single point trivially satisfies conditions (1) and (2). Assume that some set $S \subseteq \mathbb{R}^n$ of cardinality $m$ has been constructed. Then by the induction hypothesis, for every $k \leq \min(n,m)$, every $k$-element subset $P \subseteq S$ does not lie on a hyperplane of dimension less than $k-1$. Hence, every $P \subseteq S$, $|P| = k \leq \min(n,m)$, uniquely specifies a $(k-1)$-dimensional hyperplane $H_P$ that includes $P$. The induction hypothesis implies further that no point in $S \setminus P$ lies on $H_P$. Analogously, for every $l \leq \min(n, m-1)$, every $(l+1)$-element subset $Q \subseteq S$ does not lie on a hypersphere of dimension less than $l-1$. Thus, every $Q \subseteq S$, $|Q| = l+1 \leq \min(n, m-1)+1$, uniquely determines an $(l-1)$-dimensional hypersphere $B_Q$ containing all points in $Q$ and none of the points in $S \setminus Q$.

To obtain a set of cardinality $m+1$ in spherically general position we observe that the union of all hyperplanes and hyperspheres considered above, that is, the union of all $H_P$ and all $B_Q$ for all subsets $P$ and $Q$, has Lebesgue measure 0. Hence, there is some point $s \in \mathbb{R}^n$ not contained in any hyperplane $H_P$ and not contained in any hypersphere $B_Q$. By adding $s$ to $S$ we then obtain a set of cardinality $m+1$ in spherically general position. $\square$

## 3.2 Networks of Binary CSRF Neurons

The following theorem is the main step in establishing the superlinear lower bound.

**Theorem 2.** *Let $h, q, m \geq 1$ be arbitrary natural numbers. Suppose $\mathcal{N}$ is a network with one hidden layer consisting of binary CSRF neurons, where the number of hidden nodes is $h + 2^q$ and the number of input nodes is $m+q$. Assume further that the output node is linear. Then there exists a set of cardinality $hq(m+1)$ shattered by $\mathcal{N}$. This even holds if the output weights of $\mathcal{N}$ are fixed to* 1.

*Proof.* Before starting with the details we give a brief outline. The main idea is to imagine the set we want to shatter as being composed of groups of vectors, where the groups are distinguished by means of the first $m$ components and the remaining $q$ components identify the group members. We catch these groups by hyperspheres such that each hypersphere is responsible for up to $m+1$ groups. The condition of spherically general position will ensure that this operation works. The hyperspheres are then expanded to become surround regions of off-center on-surround neurons. To induce a dichotomy of the given set, we split the groups. We do this for each group using the $q$ last components in such a way that the points with designated output 1 stay within the surround region of the respective neuron and the points with designated output 0 are expelled from it. In order for this to succeed, we have to make sure that the displaced points do not fall into the surround region of some other neuron. The verification of the split operation will constitute the major part of the proof.

Let us first choose the vectors. By means of Proposition 1 we select a set $\{s_1, \ldots, s_{h(m+1)}\} \subseteq \mathbb{R}^m$ in spherically general position. Let $e_1, \ldots, e_q$ denote the unit vectors in $\mathbb{R}^q$, that is, those with a 1 in exactly one component and 0 elsewhere. We define the set $S$ by

$$S \quad = \quad \{s_i : i = 1, \ldots, h(m+1)\} \times \{e_j : j = 1, \ldots, q\}.$$

Clearly, $S$ is a subset of $\mathbb{R}^{m+q}$ and has cardinality $hq(m+1)$. It remains to show that $S$ is shattered by $\mathcal{N}$.

Let $(S_0, S_1)$ be some arbitrary dichotomy of $S$. Consider an enumeration $M_1, \ldots, M_{2^q}$ of all subsets of the set $\{1, \ldots, q\}$. Let the function $f : \{1, \ldots, h(m+1)\} \to \{1, \ldots, 2^q\}$ be defined by

$$M_{f(i)} \quad = \quad \{j : s_i e_j \in S_1\},$$

where $s_i e_j$ denotes the vector resulting from the concatenation of $s_i$ and $e_j$. We use $f$ to define a partition of $\{s_1, \ldots, s_{h(m+1)}\}$ into sets $T_k$ for $k = 1, \ldots, 2^q$ by

$$T_k \quad = \quad \{s_i : f(i) = k\}.$$

We further partition each set $T_k$ into subsets $T_{k,p}$ for $p = 1, \ldots, \lceil |T_k|/(m+1) \rceil$, where each subset $T_{k,p}$ has cardinality $m+1$, except if $m+1$ does not divide $|T_k|$, in which case there is exactly one subset of cardinality less than $m+1$. Since there are at most $h(m+1)$ elements $s_i$, the partitioning of all $T_k$ results in no more than $h$ subsets of cardinality $m+1$. Further, the fact $k \leq 2^q$ permits at most $2^q$ subsets of cardinality less than $m+1$. Thus, there are no more than $h + 2^q$ subsets $T_{k,p}$.

We employ one hidden node $H_{k,p}$ for each subset $T_{k,p}$. Thus we get by with $h + 2^q$ hidden nodes in $\mathcal{N}$ as claimed. Since $\{s_1, \ldots, s_{h(m+1)}\}$ is in spherically general position, there exists for each $T_{k,p}$ an $(m-1)$-dimensional hypersphere containing all points in $T_{k,p}$ and no other point. If $|T_{k,p}| = m+1$, this hypersphere is unique; if $|T_{k,p}| < m+1$, there is a unique $(|T_{k,p}|-2)$-dimensional hypersphere which can be extended to an $(m-1)$-dimensional hypersphere that does not contain any further point. (Note that we require condition (1) of Definition 2, otherwise no hypersphere of dimension $|T_{k,p}|-2$ including all points of $T_{k,p}$ might exist.) Clearly, if $|T_{k,p}| = 1$, we can also extend this single point to an $(m-1)$-dimensional hypersphere not including any further point.

Suppose that $(c_{k,p}, r_{k,p})$ with center $c_{k,p}$ and radius $r_{k,p}$ represents the hypersphere associated with subset $T_{k,p}$. It is obvious from the construction above that all radii satisfy $r_{k,p} > 0$. Further, since the subsets $T_{k,p}$ are pairwise disjoint, there is some $\varepsilon > 0$ such that every point $s_i \in \{s_1, \ldots, s_{h(m+1)}\}$ and every just defined hypersphere $(c_{k,p}, r_{k,p})$ satisfy

$$\text{if} \;\; s_i \notin T_{k,p} \;\; \text{then} \;\; \big| \|s_i - c_{k,p}\| - r_{k,p} \big| > \varepsilon. \tag{1}$$

In other words, $\varepsilon$ is smaller than the distance between any $s_i$ and any hypersphere $(c_{k,p}, r_{k,p})$ that does not contain $s_i$. Without loss of generality we assume that $\varepsilon$

is sufficiently small such that

$$\varepsilon \quad \leq \quad \min_{k,p} r_{k,p}. \tag{2}$$

The parameters of the hidden nodes are adjusted as follows: We define the center $\widehat{c}_{k,p} = (\widehat{c}_{k,p,1}, \ldots, \widehat{c}_{k,p,m+q})$ of hidden node $H_{k,p}$ by assigning the vector $c_{k,p}$ to the first $m$ components and specifying the remaining ones by

$$\widehat{c}_{k,p,m+j} \quad = \quad \begin{cases} 0 & \text{if } j \in M_k, \\ -\varepsilon^2/4 & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, q$. We further define new radii $\widehat{r}_{k,p}$ by

$$\widehat{r}_{k,p} \quad = \quad \sqrt{r_{k,p}^2 + (q - |M_k|)\left(\frac{\varepsilon}{2}\right)^4 + 1}$$

and choose some $\gamma > 0$ satisfying

$$\gamma \quad \leq \quad \min_{k,p} \frac{\varepsilon^2}{8\widehat{r}_{k,p}}. \tag{3}$$

The center and surround radii $\widehat{a}_{k,p}, \widehat{b}_{k,p}$ of the hidden nodes are then specified as

$$\begin{aligned} \widehat{a}_{k,p} &= \widehat{r}_{k,p} - \gamma, \\ \widehat{b}_{k,p} &= \widehat{r}_{k,p} + \gamma. \end{aligned}$$

Note that $\widehat{a}_{k,p} > 0$ holds, because $\varepsilon^2 < \widehat{r}_{k,p}^2$ implies $\gamma < \widehat{r}_{k,p}$.

This completes the assignment of parameters to the hidden nodes $H_{k,p}$. We now derive two inequalities concerning the relationship between $\varepsilon$ and $\gamma$ that we need in the following. First, we estimate $\varepsilon^2/2$ from below by

$$\begin{aligned} \frac{\varepsilon^2}{2} &> \frac{\varepsilon^2}{4} + \frac{\varepsilon^2}{64} \\ &> \frac{\varepsilon^2}{4} + \frac{\varepsilon^4}{(8\widehat{r}_{k,p})^2} \qquad \text{for all } k, p, \end{aligned}$$

where the last inequality is obtained from $\varepsilon^2 < \widehat{r}_{k,p}^2$. Using (3) for both terms on the right-hand side, we get

$$\frac{\varepsilon^2}{2} \quad > \quad 2\widehat{r}_{k,p}\gamma + \gamma^2 \qquad \text{for all } k, p. \tag{4}$$

Second, from (2) we get

$$-r_{k,p}\varepsilon + \frac{\varepsilon^2}{2} \quad < \quad -\frac{\varepsilon^2}{4} \qquad \text{for all } k, p,$$

and (3) yields

$$-\frac{\varepsilon^2}{4} \quad < \quad -2\widehat{r}_{k,p}\gamma \qquad \text{for all } k,p.$$

Putting the last two inequalities together and adding $\gamma^2$ to the right-hand side, we obtain

$$-r_{k,p}\varepsilon + \frac{\varepsilon^2}{2} \quad < \quad -2\widehat{r}_{k,p}\gamma + \gamma^2 \qquad \text{for all } k,p. \tag{5}$$

We next establish three facts about the hidden nodes.

**Claim (i).** *Let $s_ie_j$ be some point and $T_{k,p}$ some subset where $s_i \in T_{k,p}$ and $j \in M_k$. Then hidden node $H_{k,p}$ outputs 1 on $s_ie_j$.*

According to the definition of $\widehat{c}_{k,p}$, if $j \in M_k$, we have

$$\|s_ie_j - \widehat{c}_{k,p}\|^2 \quad = \quad \|s_i - c_{k,p}\|^2 + (q - |M_k|)\left(\frac{\varepsilon}{2}\right)^4 + 1.$$

The condition $s_i \in T_{k,p}$ implies $\|s_i - c_{k,p}\|^2 = r_{k,p}^2$, and thus

$$\begin{aligned}\|s_ie_j - \widehat{c}_{k,p}\|^2 \quad &= \quad r_{k,p}^2 + (q - |M_k|)\left(\frac{\varepsilon}{2}\right)^4 + 1 \\ &= \quad \widehat{r}_{k,p}^2.\end{aligned}$$

It follows that $\|s_ie_j - \widehat{c}_{k,p}\| = \widehat{r}_{k,p}$, and since $\widehat{a}_{k,p} < \widehat{r}_{k,p} < \widehat{b}_{k,p}$, point $s_ie_j$ lies within the surround region of node $H_{k,p}$. Hence, Claim (i) is shown.

**Claim (ii).** *Let $s_ie_j$ and $T_{k,p}$ satisfy $s_i \in T_{k,p}$ and $j \notin M_k$. Then hidden node $H_{k,p}$ outputs 0 on $s_ie_j$.*

From the assumptions we get here

$$\begin{aligned}\|s_ie_j - \widehat{c}_{k,p}\|^2 \quad &= \quad \|s_i - c_{k,p}\|^2 + (q - |M_k| - 1)\left(\frac{\varepsilon}{2}\right)^4 + \left(1 + \frac{\varepsilon^2}{4}\right)^2 \\ &= \quad r_{k,p}^2 + (q - |M_k|)\left(\frac{\varepsilon}{2}\right)^4 + 1 + \frac{\varepsilon^2}{2} \\ &= \quad \widehat{r}_{k,p}^2 + \frac{\varepsilon^2}{2}.\end{aligned}$$

Employing (4) on the right-hand side results in

$$\|s_ie_j - \widehat{c}_{k,p}\|^2 \quad > \quad \widehat{r}_{k,p}^2 + 2\widehat{r}_{k,p}\gamma + \gamma^2.$$

Hence, taking square roots we have $\|s_ie_j - \widehat{c}_{k,p}\| > \widehat{r}_{k,p} + \gamma$, implying that $s_ie_j$ lies outside the surround region of $H_{k,p}$. Thus, Claim (ii) follows.

**Claim (iii).** *Let $s_i e_j$ be some point and $T_{k,p}$ some subset such that $s_i \in T_{k,p}$. Then every hidden node $H_{k',p'}$ with $(k',p') \neq (k,p)$ outputs 0 on $s_i e_j$.*

Since $s_i \in T_{k,p}$ and $s_i$ is not contained in any other subset $T_{k',p'}$, condition (1) implies

$$\|s_i - c_{k',p'}\|^2 > (r_{k',p'} + \varepsilon)^2 \quad \text{or} \quad \|s_i - c_{k',p'}\|^2 < (r_{k',p'} - \varepsilon)^2. \tag{6}$$

We distinguish between two cases: whether $j \in M_{k'}$ or not.

*Case 1.* If $j \in M_{k'}$ then by the definition of $\widehat{c}_{k',p'}$ we have

$$\|s_i e_j - \widehat{c}_{k',p'}\|^2 \;=\; \|s_i - c_{k',p'}\|^2 + (q - |M_{k'}|)\left(\frac{\varepsilon}{2}\right)^4 + 1.$$

From this, using (6) and the definition of $\widehat{r}_{k',p'}$ we obtain

$$\begin{array}{rcl} \|s_i e_j - \widehat{c}_{k',p'}\|^2 & > & \widehat{r}^2_{k',p'} + 2r_{k',p'}\varepsilon + \varepsilon^2 \\ & \text{or} & \\ \|s_i e_j - \widehat{c}_{k',p'}\|^2 & < & \widehat{r}^2_{k',p'} - 2r_{k',p'}\varepsilon + \varepsilon^2. \end{array} \tag{7}$$

We derive bounds for the right-hand sides of these inequalities as follows. From (4) we have

$$\varepsilon^2 \;>\; 4\widehat{r}_{k',p'}\gamma + 2\gamma^2,$$

which, after adding $2r_{k',p'}\varepsilon$ to the left-hand side and halving the right-hand side, gives

$$2r_{k',p'}\varepsilon + \varepsilon^2 \;>\; 2\widehat{r}_{k',p'}\gamma + \gamma^2. \tag{8}$$

From (2) we get $\varepsilon^2/2 < r_{k',p'}\varepsilon$, that is, the left-hand side of (5) is negative. Hence, we may double it to obtain from (5)

$$-2r_{k',p'}\varepsilon + \varepsilon^2 \;<\; -2\widehat{r}_{k',p'}\gamma + \gamma^2.$$

Using this and (8) in (7) leads to

$$\|s_i e_j - \widehat{c}_{k',p'}\|^2 > (\widehat{r}_{k',p'} + \gamma)^2 \quad \text{or} \quad \|s_i e_j - \widehat{c}_{k',p'}\|^2 < (\widehat{r}_{k',p'} - \gamma)^2.$$

And this is equivalent to

$$\|s_i e_j - \widehat{c}_{k',p'}\| > \widehat{b}_{k',p'} \quad \text{or} \quad \|s_i e_j - \widehat{c}_{k',p'}\| < \widehat{a}_{k',p'},$$

meaning that $H_{k',p'}$ outputs 0.

*Case 2.* If $j \notin M_{k'}$ then

$$\|s_i e_j - \widehat{c}_{k',p'}\|^2 \;=\; \|s_i - c_{k',p'}\|^2 + (q - |M_{k'}|)\left(\frac{\varepsilon}{2}\right)^4 + 1 + \frac{\varepsilon^2}{2}.$$

As a consequence of this together with (6) and the definition of $\widehat{r}_{k',p'}$ we get

$$\begin{array}{rcl} \|s_i e_j - \widehat{c}_{k',p'}\|^2 & > & \widehat{r}^2_{k',p'} + 2r_{k',p'}\varepsilon + \varepsilon^2 + \dfrac{\varepsilon^2}{2} \\ & \text{or} & \\ \|s_i e_j - \widehat{c}_{k',p'}\|^2 & < & \widehat{r}^2_{k',p'} - 2r_{k',p'}\varepsilon + \varepsilon^2 + \dfrac{\varepsilon^2}{2}, \end{array}$$

from which we derive, using for the second inequality $\varepsilon \leq r_{k',p'}$ from (2),

$$\begin{array}{rcl} \|s_i e_j - \widehat{c}_{k',p'}\|^2 & > & \widehat{r}^2_{k',p'} + r_{k',p'}\varepsilon + \dfrac{\varepsilon^2}{2} \\ & \text{or} & \\ \|s_i e_j - \widehat{c}_{k',p'}\|^2 & < & \widehat{r}^2_{k',p'} - r_{k',p'}\varepsilon + \dfrac{\varepsilon^2}{2}. \end{array} \tag{9}$$

Finally, from (4) we have

$$r_{k',p'}\varepsilon + \frac{\varepsilon^2}{2} \quad > \quad 2\widehat{r}_{k',p'} + \gamma^2,$$

and, employing this together with (5), we obtain from (9)

$$\|s_i e_j - \widehat{c}_{k',p'}\|^2 > (\widehat{r}_{k',p'} + \gamma)^2 \quad \text{or} \quad \|s_i e_j - \widehat{c}_{k',p'}\|^2 < (\widehat{r}_{k',p'} - \gamma)^2,$$

which holds if and only if

$$\|s_i e_j - \widehat{c}_{k',p'}\| > \widehat{b}_{k',p'} \quad \text{or} \quad \|s_i e_j - \widehat{c}_{k',p'}\| < \widehat{a}_{k',p'}.$$

This shows that $H_{k',p'}$ outputs 0 also in this case. Thus, Claim (iii) is established.

We complete the network $\mathcal{N}$ by connecting every hidden node with weight 1 to the output node, which then computes the sum of the hidden node output values.

We finally show that we have indeed obtained a network that induces the dichotomy $(S_0, S_1)$. Assume that $s_i e_j \in S_1$. Claims (i), (ii), and (iii) imply that there is exactly one hidden node $H_{k,p}$, namely one satisfying $k = f(i)$ by the definition of $f$, that outputs 1 on $s_i e_j$. Hence, the network outputs 1 as well. On the other hand, if $s_i e_j \in S_0$, it follows from Claims (ii) and (iii) that none of the hidden nodes outputs 1. Therefore, the network output is 0. Thus, $\mathcal{N}$ shatters $S$ with output threshold 1/2 and the proof is completed. $\square$

The construction in the previous proof was based on the assumption that the difference between center radius and surround radius, given by the value $2\gamma$, can be made sufficiently small. This may require constraints on the precision of computation that are not available in natural or artificial systems. It is possible, however, to obtain the same result even if there is a lower bound on the difference of the radii. One simply has to scale the elements of the shattered set by a sufficiently large factor.

We apply the result now to obtain a superlinear lower bound for the VC dimension of networks with center-surround receptive field neurons. By $\lfloor x \rfloor$ we denote the largest integer less or equal to $x$.

**Corollary 3.** *Suppose $\mathcal{N}$ is a network with one hidden layer of $k$ binary CSRF neurons and input dimension $n \geq 2$, where $k \leq 2^n$, and assume that the output node is linear. Then $\mathcal{N}$ has VC dimension at least*

$$\left\lfloor \frac{k}{2} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor + 1\right).$$

*This even holds if the weights of the output node are not adjustable.*

*Proof.* We use Theorem 2 with $h = \lfloor k/2 \rfloor$, $q = \lfloor \log(k/2) \rfloor$, and $m = n - \lfloor \log(k/2) \rfloor$. The condition $k \leq 2^n$ guarantees that $m \geq 1$. Then there is a set of cardinality

$$hq(m+1) \;=\; \left\lfloor \frac{k}{2} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor + 1\right),$$

that is shattered by the network specified in Theorem 2. Since the number of hidden nodes is $h + 2^q \leq k$ and the input dimension is $m + q = n$, the network satisfies the required conditions. Furthermore, it was shown in the proof of Theorem 2 that all weights of the output node can be fixed to 1. Hence, they need not be adjustable. □

VC dimension bounds for neural networks are often expressed in terms of the number of weights and the network size. In the following we give a lower bound of this kind.

**Corollary 4.** *Consider a network $\mathcal{N}$ with input dimension $n \geq 2$, one hidden layer of $k$ binary CSRF neurons, where $k \leq 2^{n/2}$, and a linear output node. Let $W = k(n+2) + k + 1$ denote the number of weights. Then $\mathcal{N}$ has VC dimension at least*

$$\frac{W}{5} \log\left(\frac{k}{4}\right).$$

*This holds even in the case when the weights of the output node are fixed.*

*Proof.* According to Corollary 3, $\mathcal{N}$ has VC dimension at least $\lfloor k/2 \rfloor \cdot \lfloor \log(k/2) \rfloor \cdot (n - \lfloor \log(k/2) \rfloor + 1)$. The condition $k \leq 2^{n/2}$ implies

$$n - \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor + 1 \;\geq\; \frac{n+4}{2}.$$

We may assume that $k \geq 5$. (The statement is trivial for $k \leq 4$.) It follows, using $\lfloor k/2 \rfloor \geq (k-1)/2$ and $k/10 \geq 1/2$, that

$$\left\lfloor \frac{k}{2} \right\rfloor \;\geq\; \frac{2k}{5}.$$

Finally, we have

$$\left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor \;\geq\; \log\left(\frac{k}{2}\right) - 1 \;=\; \log\left(\frac{k}{4}\right).$$

Hence, $\mathcal{N}$ has VC dimension at least $(n+4)(k/5)\log(k/4)$, which is at least as large as the claimed bound $(W/5)\log(k/4)$. $\square$

In the networks considered thus far the input dimension was assumed to be variable. It is an easy consequence of Theorem 2 that even when $n$ is constant, the VC dimension grows still linearly in terms of the network size.

**Corollary 5.** *Assume that the input dimension is fixed and consider a network $\mathcal{N}$ with one hidden layer of binary CSRF neurons and a linear output node. Then the VC dimension of $\mathcal{N}$ is $\Omega(k)$ and $\Omega(W)$, where $k$ is the number of hidden nodes and $W$ the number of weights. This even holds in the case of fixed output weights.*

*Proof.* Choose $m, q \geq 1$ such that $m + q \leq n$, and let $h = k - 2^q$. Since $n$ is constant, $hq(m+1)$ is $\Omega(k)$. Thus, according to Theorem 2, there is a set of cardinality $\Omega(k)$ shattered by $\mathcal{N}$. Since the number of weights is $k(n+3)+1$, which is $O(k)$, the lower bound $\Omega(W)$ also follows. $\square$

### 3.3 Networks of Ternary CSRF Neurons

The results from the previous section now easily allow to derive similar bounds for ternary CSRF neurons. The following statement is the counterpart of Theorem 2.

**Theorem 6.** *Suppose $\mathcal{N}$ is a network with one hidden layer consisting of ternary CSRF neurons and a linear output node. Let $2(h+2^q)$ be the number of hidden nodes and $m+q$ the number of input nodes, where $h, m, q \geq 1$. Then there exists a set of cardinality $hq(m+1)$ shattered by $\mathcal{N}$. This even holds for fixed output weights.*

*Proof.* The idea is to use the same set $S$ as in the proof of Theorem 2 and to simulate the behavior of a binary neuron by two ternary neurons. Let $\widehat{\mathcal{N}}$ be the network constructed in the proof of Theorem 2. Assume first that we have off-center on-surround neurons available for the construction of $\mathcal{N}$. For every hidden node $\widehat{H}$ of $\widehat{\mathcal{N}}$ we introduce two hidden nodes $H, H'$ for $\mathcal{N}$ defining their parameters as follows: Node $H$ gets the same center and radii as $\widehat{H}$. Node $H'$

also gets the same center, but for the center radius we choose 0 and the surround radius is defined to be the center radius of $\widehat{H}$. Formally, $H'$ can be regarded as an off-center on-surround neuron without center region.

It is easy to see that on any input vector from $S$ the sum of the output values of $H$ and $H'$ is equal to the output value of $\widehat{H}$. Here we use the fact that no point in $S$ lies on the radii of any hidden node of $\widehat{\mathcal{N}}$. Hence, by what was shown in Theorem 2, a dichotomy $(S_0, S_1)$ is accomplished by the sum of the output values of the hidden nodes, being 0 for elements of $S_0$ and 1 for elements of $S_1$.

In case that we are dealing with on-center off-surround neurons we use the property that they are negatives of off-center on-surround neurons. Thus, defining the corresponding output weight to be $-1$ instead of 1 we obtain the same result. $\square$

In analogy to the previous section, we are now able to infer three lower bounds: A superlinear bound in terms of input dimension and network size, a superlinear bound in terms of weight number and network size, and a linear bound for fixed input dimension.

**Corollary 7.** *Suppose $\mathcal{N}$ is a network with input dimension $n \geq 2$, one hidden layer of $k \leq 2^{n+1}$ ternary CSRF neurons, and a linear output node. Then $\mathcal{N}$ has VC dimension at least*

$$\left\lfloor \frac{k}{4} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor + 1\right)$$

*even for fixed output weights.*

*Proof.* From $k \leq 2^{n+1}$ follows that $\lfloor \log(k/4) \rfloor \leq n - 1$. Hence, applying Theorem 6 with $h = \lfloor k/4 \rfloor$, $q = \lfloor \log(k/4) \rfloor$, $m = n - \lfloor \log(k/4) \rfloor$, and observing that $2(h + 2^q) \leq k$ we immediately obtain the claimed result. $\square$

**Corollary 8.** *Consider a network $\mathcal{N}$ with input dimension $n \geq 2$, one hidden layer of $k$ ternary CSRF neurons, where $k \leq 2^{(n+2)/2}$, and a linear output node. Let $W = k(n+3)+1$ denote the number of weights. Then $\mathcal{N}$ has VC dimension at least*

$$\frac{W}{12} \log\left(\frac{k}{8}\right)$$

*even for fixed output weights.*

*Proof.* Since $k \leq 2^{(n+2)/2}$, we have $(n - \lfloor \log(k/4) \rfloor + 1) \geq (n+4)/2$. Further, $\lfloor k/4 \rfloor \geq (k-3)/4$ implies for $k \geq 9$ that $\lfloor k/4 \rfloor \geq k/6$. (The statement is trivial for $k \leq 8$.) Using these estimates in the bound of Corollary 7 together with $\lfloor \log(k/4) \rfloor \geq \log(k/8)$ gives the result. $\square$

**Corollary 9.** *Assume that the input dimension $n \geq 2$ is fixed and consider a network $\mathcal{N}$ with one hidden layer of ternary CSRF neurons and a linear output node. Then the VC dimension of $\mathcal{N}$ is $\Omega(k)$ and $\Omega(W)$, where $k$ is the number of hidden nodes and $W$ the number of weights. This holds even for fixed output weights.*

*Proof.* The result can be deduced from Theorem 6 by analogy with Corollary 5. $\square$

### 3.4 Networks of Binary RBF Neurons

Finally, we consider the third variant of a discrete local receptive field neuron and show that networks of binary RBF neurons also respect the bounds established above for ternary CSRF neurons.

**Theorem 10.** *Suppose $\mathcal{N}$ is a network with one hidden layer consisting of binary RBF neurons and a linear output node. Let $n \geq 2$ be the input dimension, $k$ the number of hidden nodes, and assume that $k \leq 2^{n+1}$. Then $\mathcal{N}$ has VC dimension at least*

$$\left\lfloor \frac{k}{4} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor + 1\right).$$

*Let $W$ denote the number of weights and assume that $k \leq 2^{(n+2)/2}$. Then the VC dimension of $\mathcal{N}$ is at least*

$$\frac{W}{12} \log\left(\frac{k}{8}\right).$$

*For fixed input dimension $n \geq 2$ the VC dimension of $\mathcal{N}$ satisfies the bounds $\Omega(k)$ and $\Omega(W)$. All these bounds are valid even when the output weights are fixed.*

*Proof.* The main idea is to employ two binary RBF neurons for the simulation of one binary CSRF neuron. This is easy to achieve. Given a neuron of the latter type we provide the two RBF neurons with its center and assign its center radius to the first and its surround radius to the second neuron. If we give output weights $-1$ and $1$ to the the first and second neuron, respectively, then it is clear that on points not lying on the radii the summed ouptut of the weighted RBF neurons is equivalent to the output of the CSRF neuron.

Thus, we can do a similar construction as in the proof of Theorem 6 obtaining a network of size twice the original network such that both networks shatter the set from Theorem 2. We recall that the networks have the property that the parameters can be chosen such that no point of this set lies on any radius. The consequences stated in Corollaries 7 to 9 for ternary CSRF neurons then follow immediately for binary RBF neurons. $\square$

# 4 Superlinear Lower Bounds for Networks of Continuous Neurons

We now turn toward networks of continuous local receptive field neurons. In this section we first establish lower bounds for networks of DOG neurons. Their derivation mainly builds on constructions and results from the previous section. The bounds for RBF networks are then easily obtained.

## 4.1 Networks of DOG Neurons

We begin by deriving a result in analogy with Theorem 2.

**Theorem 11.** *Let $h, q, m \geq 1$ be arbitrary natural numbers. Suppose $\mathcal{N}$ is a network with $m + q$ input nodes, one hidden layer of $h + 2^q$ DOG neurons, and a linear output node. Then there is a set of cardinality $hq(m + 1)$ shattered by $\mathcal{N}$.*

*Proof.* We use ideas and results from the proof of Theorem 2. In particular, we show that the set constructed there can be shattered by a network of new model neurons, the so-called extended Gaussian neurons which we introduce below. Then we demonstrate that a network of these extended Gaussian neurons can be simulated by a network of DOG neurons, which establishes the statement of the theorem.

We define an extended Gaussian neuron with $n$ inputs to compute the function $\tilde{g} : \mathbb{R}^{2n+2} \to \mathbb{R}$ with

$$\tilde{g}(c, \sigma, \alpha, x) \;=\; 1 - \left( \alpha \exp\left( -\frac{\|x - c\|^2}{\sigma^2} \right) - 1 \right)^2 ,$$

where $x_1, \ldots, x_n$ are the input variables and $c_1, \ldots, c_n$, $\alpha$, and $\sigma > 0$ are real-valued parameters. Thus, the computation of an extended Gaussian neuron is performed by scaling the output of a Gaussian RBF neuron with $\alpha$, squaring the difference to 1, and comparing this value with 1.

Let $S \subseteq \mathbb{R}^{m+q}$ be the set of cardinality $hq(m + 1)$ constructed in the proof of Theorem 2. In particular, $S$ has the form

$$S \;=\; \{s_i e_j : i = 1, \ldots, h(m + 1); j = 1, \ldots, q\}.$$

We have also defined in that proof binary CSRF neurons $H_{k,p}$ as hidden nodes in terms of parameters $\widehat{c}_{k,p} \in \mathbb{R}^{m+q}$, which became the centers of the neurons, and $\widehat{r}_{k,p} \in \mathbb{R}$, which gave the center radii $\widehat{a}_{k,p} = \widehat{r}_{k,p} - \gamma$ and the surround radii $\widehat{b}_{k,p} = \widehat{r}_{k,p} + \gamma$ using some $\gamma > 0$. The number of hidden nodes was not larger than $h + 2^q$. We replace the CSRF neurons by extended Gaussian neurons $G_{k,p}$

with parameters $c_{k,p}, \sigma_{k,p}, \alpha_{k,p}$ defined as follows. Assume some $\sigma > 0$ that will be specified later. Then we let

$$
\begin{aligned}
c_{k,p} &= \widehat{c}_{k,p}, \\
\sigma_{k,p} &= \sigma, \\
\alpha_{k,p} &= \exp\left(\frac{\widehat{r}_{k,p}^2}{\sigma^2}\right).
\end{aligned}
$$

These hidden nodes are connected to the output node with all weights being 1. We call this network $\mathcal{N}'$ and claim that it shatters $S$.

Consider some arbitrary dichotomy $(S_0, S_1)$ of $S$ and some $s_i e_j \in S$. Then node $G_{k,p}$ computes

$$
\begin{aligned}
\tilde{g}(c_{k,p}, \sigma_{k,p}, \alpha_{k,p}, s_i e_j) &= 1 - \left(\alpha_{k,p} \exp\left(-\frac{\|s_i e_j - c_{k,p}\|^2}{\sigma_{k,p}^2}\right) - 1\right)^2 \\
&= 1 - \left(\exp\left(\frac{\widehat{r}_{k,p}^2}{\sigma^2}\right) \cdot \exp\left(-\frac{\|s_i e_j - \widehat{c}_{k,p}\|^2}{\sigma^2}\right) - 1\right)^2 \\
&= 1 - \left(\exp\left(-\frac{\|s_i e_j - \widehat{c}_{k,p}\|^2 - \widehat{r}_{k,p}^2}{\sigma^2}\right) - 1\right)^2. \qquad (10)
\end{aligned}
$$

Suppose first that $s_i e_j \in S_1$. It was shown by Claims (i), (ii), and (iii) in the proof of Theorem 2 that there is exactly one hidden node $H_{k,p}$ that outputs 1 on $s_i e_j$. In particular, Claim (i) established that this node satisfies

$$
\|s_i e_j - \widehat{c}_{k,p}\|^2 = \widehat{r}_{k,p}^2.
$$

Hence, according to (10) node $G_{k,p}$ outputs 1. We note that this holds for all values of $\sigma$. Further, the derivations of Claims (ii) and (iii) yielded that those nodes $H_{k,p}$ that output 0 on $s_i e_j$ satisfy

$$
\|s_i e_j - \widehat{c}_{k,p}\|^2 > (\widehat{r}_{k,p} + \gamma)^2 \quad \text{or} \quad \|s_i e_j - \widehat{c}_{k,p}\|^2 < (\widehat{r}_{k,p} - \gamma)^2. \qquad (11)
$$

This implies for the computation of $G_{k,p}$ that in (10) we can make the expression

$$
\exp\left(-\frac{\|s_i e_j - \widehat{c}_{k,p}\|^2 - \widehat{r}_{k,p}^2}{\sigma^2}\right)
$$

as close to 0 as necessary by choosing $\sigma$ sufficiently small. Since this does not affect the node that outputs 1, network $\mathcal{N}'$ computes a value close to 1 on $s_i e_j$.

On the other hand, for the case $s_i e_j \in S_0$ it was shown in Theorem 2 that all nodes $H_{k,p}$ output 0. Thus, each of them satisfies condition (11), implying that if $\sigma$ is sufficiently small each node $G_{k,p}$, and hence $\mathcal{N}'$, outputs a value close to 0. Altogether, $S$ is shattered by thresholding the output of $\mathcal{N}'$ at 1/2.

Finally, we show that $S$ can be shattered by a network $\mathcal{N}$ of the same size with DOG neurons as hidden nodes. The computation of an extended Gaussian neuron can be rewritten as

$$\begin{aligned}
\tilde{g}(c,\sigma,\alpha,x) &= 1-\left(\alpha \exp\left(-\frac{\|x-c\|^2}{\sigma^2}\right)-1\right)^2 \\
&= 1-\left(\alpha^2 \exp\left(-\frac{2\|x-c\|^2}{\sigma^2}\right)-2\alpha \exp\left(-\frac{\|x-c\|^2}{\sigma^2}\right)+1\right) \\
&= 2\alpha \exp\left(-\frac{\|x-c\|^2}{\sigma^2}\right)-\alpha^2 \exp\left(-\frac{2\|x-c\|^2}{\sigma^2}\right) \\
&= g_{\mathrm{DOG}}(c,\sigma,\sigma/\sqrt{2},2\alpha,\alpha^2,x).
\end{aligned}$$

Hence, the extended Gaussian neuron is equivalent to a weighted difference of two Gaussian neurons with center $c$, widths $\sigma, \sigma/\sqrt{2}$ and weights $2\alpha, \alpha^2$, respectively. Thus, the extended Gaussian neurons can be replaced by DOG neurons, which completes the proof. □

We note that the network of extended Gaussian neurons constructed in the previous proof has all output weights fixed, whereas the output weights of the DOG neurons, that is, the parameters $\alpha$ and $\beta$ in the notation of Section 2.1, are calculated from the parameters of the extended Gaussian neurons and, therefore, depend on the particular dichotomy to be implemented. (It is trivial for a DOG network to have an output node with fixed weights since the DOG neurons have built-in output weights.)

We are now able to deduce a superlinear lower bound on the VC dimension of DOG networks.

**Corollary 12.** *Suppose $\mathcal{N}$ is a network with one hidden layer of DOG neurons and a linear output node. Let $\mathcal{N}$ have $k$ hidden nodes and input dimension $n \geq 2$, where $k \leq 2^n$. Then $\mathcal{N}$ has VC dimension at least*

$$\left\lfloor \frac{k}{2} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{2}\right) \right\rfloor + 1\right).$$

*Let $W$ denote the number of weights and assume that $k \leq 2^{n/2}$. Then the VC dimension of $\mathcal{N}$ is at least*

$$\frac{W}{5}\log\left(\frac{k}{4}\right).$$

*For fixed input dimension the VC dimension of $\mathcal{N}$ is bounded by $\Omega(k)$ and $\Omega(W)$.*

*Proof.* The results are implied by Theorem 11 in the same way as Corollaries 3, 4, and 5 follow from Theorem 2. □

## 4.2 Networks of Gaussian RBF Neurons

We can now give the answer to the question of Anthony and Holden (1994) quoted in the introduction.

**Theorem 13.** *Suppose $\mathcal{N}$ is a network with one hidden layer of Gaussian RBF neurons and a linear output node. Let $k$ be the number of hidden nodes and $n$ the input dimension, where $n \geq 2$ and $k \leq 2^{n+1}$. Then $\mathcal{N}$ has VC dimension at least*

$$\left\lfloor \frac{k}{4} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor \cdot \left(n - \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor + 1\right).$$

*Let $W$ denote the number of weights and assume that $k \leq 2^{(n+2)/2}$. Then the VC dimension of $\mathcal{N}$ is at least*

$$\frac{W}{12} \log\left(\frac{k}{8}\right).$$

*For fixed input dimension $n \geq 2$ the VC dimension of $\mathcal{N}$ satisfies the bounds $\Omega(k)$ and $\Omega(W)$.*

*Proof.* Clearly, a DOG neuron can be simulated by two Gaussian RBF Neurons. Thus, by virtue of Theorem 11 there is a network $\mathcal{N}$ with $m + q$ input nodes and one hidden layer of $2(h + 2^q)$ Gaussian RBF neurons that shatters some set of cardinality $hq(m + 1)$. Choosing $h = \lfloor k/4 \rfloor, q = \lfloor \log(k/4) \rfloor$, and $m = n - \lfloor \log(k/4) \rfloor$ we obtain similarly to Corollary 7 the claimed lower bound in terms of $n$ and $k$.

Furthermore, the stated bound in terms of $W$ and $k$ follows by analogy to the reasoning in Corollary 8. Finally, the bound for fixed input dimension is obvious, as in the proof of Corollary 5. $\square$

Some radial basis function networks studied theoretically or used in practice have no adjustable width parameters (for instance Broomhead and Lowe, 1988; Powell, 1992). Therefore, a natural question is whether the previous result also holds for networks with fixed width parameters. The values of the width parameters for Theorem 13 arise from the widths of DOG neurons specified in Theorem 11. The two width parameters of each DOG neuron have the form $\sigma$ and $\sigma/\sqrt{2}$ where $\sigma$ is common to all DOG neurons and is only required to be sufficiently small. Hence, we can choose a single $\sigma$ that is sufficiently small for all dichotomies to be induced. Thus, for the RBF network we not only have that the width parameters can be fixed, but even that there need to be only two different width values—solely depending on the architecture and not on the particular dichotomy.

**Corollary 14.** *Let $\mathcal{N}$ be a Gaussian RBF network with $n$ input nodes and $k$ hidden nodes satisfying the conditions of Theorem 13. Then there exists a real number $\sigma_{k,n} > 0$ such that the VC dimension bounds stated in Theorem 13 hold for $\mathcal{N}$ with each RBF neuron having fixed width $\sigma_{k,n}$ or $\sigma_{k,n}/\sqrt{2}$.*

With regard to Theorem 13 we further remark that $k$ has been previously established as lower bound for RBF networks by Anthony and Holden (1994). Further, also Theorem 19 of Lee et al. (1995) in connection with the result of Erlich et al. (1997) implies the lower bound $\Omega(nk)$, and hence $\Omega(k)$ for fixed input dimension. By means of Theorem 13 we are now able to present a lower bound that is even superlinear in $k$.

**Corollary 15.** *Let $n \geq 2$ and $\mathcal{N}$ be the network with $k = 2^{n+1}$ hidden Gaussian RBF neurons. Then $\mathcal{N}$ has VC dimension at least*

$$\frac{k}{3} \log\left(\frac{k}{8}\right).$$

*Proof.* Since $k = 2^{n+1}$, we may substitute $n = \log k - 1$ in the first bound of Theorem 13. Hence, the VC dimension of $\mathcal{N}$ is at least

$$\left\lfloor \frac{k}{4} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor \cdot \left(\log k - \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor\right) \;\geq\; 2\left\lfloor \frac{k}{4} \right\rfloor \cdot \left\lfloor \log\left(\frac{k}{4}\right) \right\rfloor.$$

As in the proof of Corollary 8 we use that $\lfloor k/4 \rfloor \geq k/6$ and $\lfloor \log(k/4) \rfloor \geq \log(k/8)$. This yields the claimed bound. $\square$

# 5 Bounds for Single Neurons

In this section we consider the three discrete variants of a local receptive field neuron and the Gaussian RBF neuron. We show that their VC dimension is at most linear. Furthermore, this bound is asymptotically tight.

## 5.1 Discrete Neurons

We assume in the following that the output of the ternary CSRF neuron is thresholded at $1/2$ or any other fixed value from the interval $(0, 1]$, to obtain output values in $\{0, 1\}$. Thus, we can treat the binary and ternary CSRF neuron similarly. (If the threshold is chosen from the interval $[-1, 0]$ this corresponds to a negated binary RBF neuron and, hence, has the VC dimension of the latter.)

**Theorem 16.** *The VC dimension of a binary RBF neuron with $n$ inputs is equal to $n+1$. The VC dimension of a (binary and ternary) center-surround neuron with $n$ inputs is at least $n+1$ and at most $4n+5$.*

*Proof.* The class of functions computed by a binary RBF neuron with $n$ inputs can be identified with the class of balls in $\mathbb{R}^n$. Dudley (1979) shows that the VC dimension of this class is equal to $n+1$ (see also Wenocur and Dudley, 1981; Assouad, 1983). This gives the result for the RBF neuron.

Clearly, a binary and ternary center-surround neuron can simulate the RBF neuron by adjusting the center radius to 0. This implies the lower bound $n+1$. For the upper bound consider a center-surround neuron with $n$ inputs and assume, without loss of generality, that its output is binary. Let $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ be the center and $a, b \in \mathbb{R}$ the radii. Then, if $f : \mathbb{R}^n \to \{0, 1\}$ is the function computed by the neuron, on some input vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ it satisfies

$$\begin{aligned}
f(x) = 1 \quad &\Longleftrightarrow \quad \|x - c\| \geq a \quad \text{and} \quad \|x - c\| \leq b \\
&\Longleftrightarrow \quad (x_1 - c_1)^2 + \cdots + (x_n - c_n)^2 \geq a^2 \quad \text{and} \\
&\qquad\quad (x_1 - c_1)^2 + \cdots + (x_n - c_n)^2 \leq b^2 \\
&\Longleftrightarrow \quad \|x\|^2 - 2c_1x_1 - \cdots - 2c_nx_n \geq a^2 - c_1^2 - \cdots - c_n^2 \quad \text{and} \\
&\qquad\quad -\|x\|^2 + 2c_1x_1 + \cdots + 2c_nx_n \geq -b^2 + c_1^2 + \cdots + c_n^2.
\end{aligned}$$

Each of the last two inequalities defines a halfspace in $\mathbb{R}^{n+1}$, both with weights $1, -2c_1, \ldots, -2c_n$ or the negative thereof, and with thresholds $a^2 - c_1^2 - \cdots - c_n^2$ or $-b^2 + c_1^2 + \cdots + c_n^2$, respectively. Thus, we have that the number of dichotomies induced by a binary center-surround neuron on some finite subset of $\mathbb{R}^n$ is not larger than the number of dichotomies induced by intersections of parallel halfspaces on some subset of $\mathbb{R}^{n+1}$ with the same cardinality, where the additional input component is obtained as $\|x\|^2$ for every input vector. Hence, a set shattered by a center-surround neuron in $\mathbb{R}^n$ gives rise to a set of the same cardinality shattered by intersections of parallel halfspaces in $\mathbb{R}^{n+1}$. According to Theorem 19, which is given in the appendix, the VC dimension of the class of intersections of parallel halfspaces in $\mathbb{R}^{n+1}$ is at most $4n+5$. This entails the bound for the center-surround neuron. $\square$

We remark that, in contrast to the RBF neuron, the exact values for the VC dimension of center-surround neurons are not known yet.

## 5.2 Gaussian RBF Neurons

It is easy to see that a thresholded Gaussian RBF neuron, that is, one with a fixed output threshold, is equivalent to a binary RBF neuron. Hence by Theorem 16, its VC dimension is equal to the number of its parameters. The pseudo dimension generalizes the VC dimension to real-valued function classes and is defined as follows.

**Definition 3.** *Let $\mathcal{F}$ be a class of functions mapping $\mathbb{R}^n$ to $\mathbb{R}$. The* pseudo dimension *of $\mathcal{F}$ is the cardinality of the largest set $S \subseteq \mathbb{R}^{n+1}$ shattered by the class* $\{(x, y) \mapsto \operatorname{sgn}(f(x) - y) : f \in \mathcal{F}\}$.

The pseudo dimension is a stronger notion than the VC dimension in that an upper bound on the pseudo dimension of some function class also yields the same bound on the VC dimension of the thresholded class, whereas the converse need not necessarily be true. Thus, there is no general way of inferring the pseudo dimension of a Gaussian RBF neuron from the VC dimension of a binary RBF neuron. Nevertheless, the pseudo dimension of a single Gaussian RBF neuron is linear as we show now.

**Theorem 17.** *The pseudo dimension of a Gaussian RBF neuron with $n$ inputs is at least $n + 1$ and at most $n + 2$.*

*Proof.* The lower bound easily follows from the facts that a thresholded Gaussian RBF neuron can simulate any binary RBF neuron, that a binary RBF neuron has VC dimension $n+1$ (see Theorem 16), and that the VC dimension is a lower bound for the pseudo dimension.

We obtain the upper bound as follows: According to a well known result (see, e.g., Haussler, 1992, Theorem 5), since the function $z \mapsto \exp(-z)$ is continuous and strictly decreasing, the pseudo dimension of the class

$$\left\{ x \mapsto \exp\left(-\frac{\|x - c\|^2}{\sigma^2}\right) \ : \ c \in \mathbb{R}^n, \sigma \in \mathbb{R} \setminus \{0\} \right\}$$

is equal to the pseudo dimension of the class

$$\left\{ x \mapsto \frac{\|x - c\|^2}{\sigma^2} \ : \ c \in \mathbb{R}^n, \sigma \in \mathbb{R} \setminus \{0\} \right\},$$

which is, by Definitions 1 and 3, equal to the VC dimension of the class

$$\left\{ (x, y) \mapsto \operatorname{sgn}\left(\frac{\|x - c\|^2}{\sigma^2} - y\right) \ : \ c \in \mathbb{R}^n, \sigma \in \mathbb{R} \setminus \{0\} \right\}.$$

This class can also be written as

$$\left\{ (x, y) \mapsto \operatorname{sgn}(\|x\|^2 - 2c \cdot x + \|c\|^2 - \sigma^2 y) \ : \ c \in \mathbb{R}^n, \sigma \in \mathbb{R} \setminus \{0\} \right\}.$$

Each function in this class has the form $(x, y) \mapsto \operatorname{sgn}(f(x) + g(x, y))$ with $f(x) = \|x\|^2$ and $g$ being an affine function in $n + 1$ variables. Hence, the VC dimension of this class cannot be larger than the VC dimension of the class

$$\{\operatorname{sgn}(f + g) \ : \ g \text{ affine}, g : \mathbb{R}^{n+1} \to \mathbb{R}\}.$$

Wenocur and Dudley (1981) show that if $G$ is a $d$-dimensional vector space of real-valued functions then $\{\operatorname{sgn}(f + g) \ : \ g \in G\}$ has VC dimension $d$ (see also Anthony and Bartlett, 1999, Theorem 3.5). Thus the upper bound follows since the class of affine functions in $n + 1$ variables is a vector space of dimension $n + 2$. □

# 6 Upper Bounds for Networks of Discrete Neurons

The following result shows that one-hidden-layer networks of discrete local receptive field neurons have a VC dimension bounded by $O(W \log k)$. This implies that the lower bounds established in Section 3 are asymptotically tight. For the proof we employ a method from a similar result for threshold networks.

**Theorem 18.** *Suppose $\mathcal{N}$ is a network with one hidden layer of binary RBF neurons and a linear output node. Let $k$ denote the number of hidden nodes and $W = nk + 2k + 1$ the number of weights. Then the VC dimension of $\mathcal{N}$ is at most $2W \log((2k+2)/\ln 2)$. If the hidden nodes are binary or ternary CSRF neurons, the VC dimension is at most $2W \log((4k+2)/\ln 2)$.*

*Proof.* A binary RBF neuron with $n$ inputs has VC dimension $n+1$ (see Theorem 16). The output node of $\mathcal{N}$ is a linear neuron with $k$ inputs, thus it has VC dimension $k+1$. Since each node of $\mathcal{N}$ has a VC dimension equal to the number of its parameters, it follows by reasoning similarly as in Theorem 6.1 of Anthony and Bartlett (1999) that the number of dichotomies induced by $\mathcal{N}$ on a set of cardinality $m$, where $m \geq W$, is at most $(em(k+1)/W)^W$. (Note that $\mathcal{N}$ has $k+1$ computation nodes.) This implies that the VC dimension is at most $2W \log((2k+2)/\ln 2)$.

Consider now the case that the hidden nodes are CSRF neurons. Clearly, a weighted binary or ternary CSRF neuron can be simulated by a weighted combination of two binary RBF neurons. Thus, $\mathcal{N}$ can be simulated by a network $\mathcal{N}'$ with $2k$ binary RBF neurons as hidden nodes. Now observe that each CSRF neuron gives rise to two RBF neurons with the same center. Thus, although the number of nodes and connections in $\mathcal{N}'$ has increased, the number of parameters is the same as in $\mathcal{N}$. In other words, $\mathcal{N}'$ is a network with equivalences among its weights. Combining a method due to Shawe-Taylor (1995) for networks with equivalences with the above-mentioned derivation by Anthony and Bartlett (1999), we obtain that $\mathcal{N}'$ induces at most $(em(2k+1)/W)^W$ dichotomies on a set of cardinality $m$. This results in a VC dimension not larger than $2W \log((4k+2)/\ln 2)$. $\square$

# 7 Conclusions

Local receptive fields occur in many kinds of biological and artificial neural networks. We have studied here several models of local receptive field neurons and have established superlinear VC dimension lower bounds for networks with one hidden layer. Although, compared with the previously known linear bounds, at first sight the gain by a logarithmic factor seems exiguous, there are at least two arguments showing that it constitutes a significant improvement. First, the VC dimension is a rather coarse measure. Increasing it by one amounts to doubling the number of functions computed by the network. Second, in a network with the VC dimension linearly bounded from above by the number of weights, each weight can be considered responsible for a particular input vector. Superlinearity implies that each weight manages to get hold of a number of input vectors that increases with the network size. Thus, in networks with superlinear VC dimension the neurons have found a very effective way to cooperate and coordinate their computations.

The VC dimension yields bounds on the complexity of learning for several models of learnability. For instance, bounds on the computation time or the number of examples required for learning can often be expressed in terms of the VC dimension. If the VC dimension provides a lower bound in a model of learning then the superlinear lower bounds given here yield new lower bounds on the complexity of learning using local receptive field neural networks. Of course, if the VC dimension serves as upper bound in a model, there is no immediate consequence. But then one may be encouraged to find other measures that more tightly quantify the complexity of learning in these models.

For the discrete versions of local receptive field neurons we have shown that the superlinear lower bounds for networks are asymptotically tight. The currently available methods for RBF and DOG networks give only rise to the upper bound $O(W^2k^2)$ for these networks. This bound, however, is also valid for networks of unrestricted depth and for networks of sigmoidal neurons. The problems of narrowing the gaps between upper and lower bounds for RBF and sigmoidal networks with one hidden layer seem therefore to be closely related. We have also established tight linear bounds for the VC dimension of single discrete neurons and for the pseudo dimension of the Gaussian RBF neuron. The VC and pseudo dimension of the DOG neuron can be shown to be at most quadratic. We conjecture that also the DOG neuron has linear VC and pseudo dimension, but the methods currently available do not seem to permit an answer.

In the constructions of the sets being shattered we have permitted arbitrary real vectors. It is not hard to see that rational numbers suffice. It would be interesting to know what happens for even more restrictive inputs such as, for instance, Boolean vectors. We have also allowed that the centers of the local receptive field neurons can be placed anywhere in the input domain. We do not know if the results hold when the centers may not freely float around.

The superlinear bounds involve constant factors that are the largest known for any standard neural network with one hidden layer. This fact could be interpreted as evidence that the cooperative computational capabilities of local receptive field neurons are even higher than those of other neuron types. This statement, however, must be taken with a pinch of salt since the constants in these bounds are not yet known to be tight.

Gaussian units are just one type of radial basis function neuron. The method we have developed for obtaining superlinear lower bounds is of quite general nature. We expect it therefore to be applicable for other RBF networks as well. The main clue in the result for RBF networks was first to consider CSRF and DOG networks. With this idea we have established a new kind of link between neurophysiological models and artificial neural networks. This link extends the paradigm of neural computation by demonstrating that models originating from neuroscience do not only lead to powerful computing mechanisms but can also be essential in theory, that is, in *proofs* concerning the computational power of those mechanisms.


## Acknowledgment

This work has been supported in part by the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2, No. 27150.


# Appendix: A VC Dimension Upper Bound for Intersections and Unions of Parallel Halfspaces

We consider the function classes defined by intersections and unions of parallel halfspaces and derive an upper bound on the VC dimension of these classes. This bound was used in Theorem 16. A general way of bounding the VC dimension of classes that are constructed from finite intersections and unions has been established by Blumer et al. (1989). In particular, they show that if we form a new class having as members intersections of $s$ functions from a class with VC dimension $d$, then the VC dimension of the new class is less than $2ds \log(3s)$ (Blumer et al., 1989, Lemma 3.2.3). The following, new calculation for parallel halfspaces results in a bound with improved constants.

**Theorem 19.** *The function class consisting of intersections of parallel halfspaces in $\mathbb{R}^n$ has VC dimension at most $4n+1$. The same holds for the class of unions of parallel halfspaces.*

*Proof.* The proof is given for intersections; the result then follows for unions by duality. Clearly, it is sufficient to consider intersections of only two parallel halfspaces. We use $\mathbb{R}^{n+2}$ to represent the joint parameter domain for the halfspaces. The first $n$ components defining the weights are shared by both halfspaces. The components $n+1$ and $n+2$ correspond to their separate thresholds. The main step is to derive an upper bound on the number of dichotomies induced on any set $S \subseteq \mathbb{R}^n$ of cardinality $m$. We assume without loss of generality that $S$ is in general position. (If not then the elements can be perturbed to obtain a set in general position with a number of dichotomies no less than for the original set. See, e.g., Anthony and Bartlett, 1999, p. 34.)

First, we give an upper bound on the number of dichotomies induced by pairs of parallel halfspaces where each halfspace is non-trivial. Here, we say that a halfspace is trivial if it induces one of the dichotomies $(\emptyset, S)$ or $(S, \emptyset)$. Such a bound is obtained in terms of the number of connected components into which the parameter domain is partitioned by certain hyperplanes arising from the elements of $S$. Every input vector $(s_1, \ldots, s_n) \in S$ gives rise to the two hyperplanes

$$\{x \in \mathbb{R}^{n+2} : s_1 x_1 + \cdots + s_n x_n - x_{n+1} = 0\},$$
$$\{x \in \mathbb{R}^{n+2} : s_1 x_1 + \cdots + s_n x_n - x_{n+2} = 0\},$$

that is, their representations in $\mathbb{R}^{n+2}$ are the vectors

$$(s_1, \ldots, s_n, -1, 0) \quad \text{and} \quad (s_1, \ldots, s_n, 0, -1),$$

respectively. All hyperplanes are homogeneous, that is, they pass through the origin. It is clear that for every connected component of $\mathbb{R}^{n+2}$ arising from this partition, the two functions induced on $S$ by the pair of halfspaces represented in this way are the same for all vectors belonging to this component. Thus, the number of connected components provides an upper bound on the number of induced dichotomies.

A well-known result attributed to Schläfli (1901) states that $m$ homogeneous hyperplanes in general position partition $\mathbb{R}^n$ into exactly

$$2 \sum_{i=0}^{n} \binom{m-1}{i} \tag{12}$$

connected components (see also Anthony and Bartlett, 1999, Lemma 3.3). Hence, the set $S$, giving rise to $2m$ hyperplanes, partitions $\mathbb{R}^{n+2}$ into at most

$$2 \sum_{i=0}^{n+1} \binom{2m-1}{i} \tag{13}$$

connected components. Not all of them, however, represent pairs of non-trivial halfspaces. For every trivial dichotomy of the first halfspace we can have as many dichotomies induced by the second halfspace as there are dichotomies possible by a single halfspace in $\mathbb{R}^{n+1}$ on a set of cardinality $m$; and likewise for every trivial dichotomy of the second halfspace. Hence, using Schläfli's count (12) we may subtract from (13) the amount of

$$\left(4 \sum_{i=0}^{n} \binom{m-1}{i}\right) + \left(4 \sum_{i=0}^{n} \binom{m-1}{i}\right) - 4. \tag{14}$$

The term $-4$ at the end results from the fact that the four combinations of trivial halfspaces are counted by both sums. Note also that the number given in (14) is not a bound, but is precise since $S$ is in general position. Up to this point, the pairs of halfspaces also include redundant combinations where the intersection is empty or one halfspace is a subset of the other. Clearly, for every non-redundant pair there are three redundant ones. Therefore, we can exclude the latter dividing the number by 4. Thus, an upper bound for the number of pairs of non-trivial halfspaces with non-empty intersections is obtained by subtracting one fourth of (14) from (13) giving

$$\left(\frac{1}{2}\sum_{i=0}^{n+1}\binom{2m-1}{i}\right)-\left(2\sum_{i=0}^{n}\binom{m-1}{i}\right)+1. \tag{15}$$

That the intersection of two halfspaces is non-empty does not imply that the dichotomy induced on $S$ is non-empty. Therefore, we are allowed to exclude these cases. Each pair of non-trivial halfspaces with empty intersection on $S$ gives rise to two non-trivial dichotomies that can be induced by a single halfspace. Thus, we may subtract half the number of non-trivial dichotomies induced by a single halfspace, which is

$$\left(\sum_{i=0}^{n}\binom{m-1}{i}\right)-1. \tag{16}$$

Finally, we take those pairs into account where at least one halfspace induces a trivial dichotomy. In this case the dichotomy can be induced by a single halfspace, that is, we may add the amount given by (12). All in all, an upper bound is provided by (15) minus (16) plus (12), yielding

$$\left(\frac{1}{2}\sum_{i=0}^{n+1}\binom{2m-1}{i}\right)-\left(\sum_{i=0}^{n}\binom{m-1}{i}\right)+2. \tag{17}$$

Assuming $1 \leq n+1 \leq 2m-1$ without loss of generality, we use the estimates

$$\frac{1}{2}\sum_{i=0}^{n+1}\binom{2m-1}{i} \quad < \quad \frac{1}{2}\left(\frac{e(2m-1)}{n+1}\right)^{n+1}$$

(see, e.g., Anthony and Bartlett, 1999, Theorem 3.7) and

$$\sum_{i=0}^{n}\binom{m-1}{i} \quad \geq \quad 2,$$

whence we obtain that the number of dichotomies is less than

$$\frac{1}{2}\left(\frac{e(2m-1)}{n+1}\right)^{n+1}.$$

Now suppose that $S$ is shattered. Then all $2^m$ dichotomies must be induced, which implies that

$$2^m \quad < \quad 2^n \left( \frac{e(2m-1)}{2(n+1)} \right)^{n+1} .$$

Taking logarithms, this is equivalent to

$$m \quad < \quad n + (n+1) \log \left( \frac{e(2m-1)}{2(n+1)} \right) . \tag{18}$$

It is well known that all real numbers $\alpha, \beta > 0$ satisfy the inequality

$$\ln \alpha \quad \leq \quad \alpha\beta + \ln(1/\beta) - 1$$

(see, e.g., Anthony and Bartlett, 1999, Appendix A.1.1). Substituting $\alpha = (2m-1)/2$ and $\beta = (\ln 2)/(2(n+1))$ yields

$$\ln \left( \frac{2m-1}{2} \right) \quad \leq \quad \frac{(2m-1)\ln 2}{4(n+1)} + \ln \left( \frac{2(n+1)}{e \ln 2} \right) ,$$

implying

$$(n+1) \log \left( \frac{2m-1}{2} \right) \quad \leq \quad \frac{m}{2} - \frac{1}{4} + (n+1) \log \left( \frac{2(n+1)}{e \ln 2} \right) .$$

Using this in inequality (18), it follows that

$$m \quad < \quad \frac{m}{2} + n + (n+1) \log \left( \frac{2}{\ln 2} \right) - \frac{1}{4},$$

which is equivalent to

$$m \quad < \quad n \left( 2 + \log \left( \frac{2}{\ln 2} \right) \right) + \log \left( \frac{2}{\ln 2} \right) - \frac{1}{2}.$$

This implies that $m \leq 4n + 1$. Hence, the cardinality of any set shattered by intersections of parallel halfspaces, and thus the VC dimension of this class, is not larger than this number. $\square$

# References


Anthony, M. (1995). Classification by polynomial surfaces. *Discrete Applied Mathematics*, 61:91–103.

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

Anthony, M. and Holden, S. B. (1994). Quantifying generalization in linearly weighted neural networks. *Complex Systems*, 8:91–114.

Assouad, P. (1983). Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282.

Atick, J. J. and Redlich, A. N. (1993). Convergent algorithm for sensory receptive field development. *Neural Computation*, 5:45–60.

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

Bartlett, P. L. and Williamson, R. C. (1996). The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. *Neural Computation*, 8:625–628.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36:929–965.

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

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334.

Dudley, R. M. (1979). Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31:306–308.

Enroth-Cugell, C. and Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology*, 187:517–552.

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

Glezer, V. D. (1995). *Vision and Mind: Modeling Mental Functions.* Lawrence Erlbaum, Mahwah, New Jersey.

Hartman, E. J., Keeler, J. D., and Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*, 2:210–215.

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

Hawken, M. J. and Parker, A. J. (1987). Spatial properties of neurons in the monkey striate cortex. *Proceedings of the Royal Society of London. Series B*, 231:251–288.

Holden, S. B. and Rayner, P. J. W. (1995). Generalization and PAC learning: Some new results for the class of generalized single-layer networks. *IEEE Transactions on Neural Networks*, 6:368–380.

Howlett, R. J. and Jain, L. C., editors (2001a). *Radial Basis Function Networks 1: Recent Developments in Theory and Applications*. Studies in Fuzziness and Soft Computing. Springer-Verlag, Berlin.

Howlett, R. J. and Jain, L. C., editors (2001b). *Radial Basis Function Networks 2: New Advances in Design*. Studies in Fuzziness and Soft Computing. Springer-Verlag, Berlin.

Joshi, A. and Lee, C.-H. (1993). Backpropagation learns Marr's operator. *Biological Cybernetics*, 70:65–73.

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

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

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16:37–68.

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

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1997). Correction to "Lower bounds on VC-dimension of smoothly parameterized function classes". *Neural Computation*, 9:765–769.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences USA*, pages 7508–7512.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Liu, S.-C. and Boahen, K. (1996). Adaptive retina with center-surround receptive field. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 678–684, MIT Press, Cambridge, Mass.

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

Maass, W. (1994). Neural nets with super-linear VC-dimension. *Neural Computation*, 6:877–884.

Maass, W. and Schmitt, M. (1999). On the complexity of learning for spiking neurons with temporal coding. *Information and Computation*, 153:26–46.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Freeman, New York.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B*, 207:187–217.

Mead, C. (1989). *Analog VLSI and Neural Systems.* Addison-Wesley, Reading, Mass.

Mead, C. A. and Mahowald, M. A. (1988). A silicon model of early visual processing. *Neural Networks*, 1:91–97.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8:164–177.

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

Natschläger, T. and Schmitt, M. (1996). Exact VC-dimension of Boolean monomials. *Information Processing Letters*, 59:19–20. Erratum *ibid.* 60:107, 1996.

Nicholls, J. G., Martin, A. R., and Wallace, B. G. (1992). Retina and lateral geniculate nucleus. In *From Neuron to Brain: A Cellular and Molecular Approach to the Function of the Nervous System*, chapter 16, pages 559–600. Sinauer Associates, Sunderland, Mass., third edition.

Nilsson, N. J. (1990). *The Mathematical Foundations of Learning Machines.* Morgan Kaufmann, San Mateo, CA.

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

Park, J. and Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, 5:305–316.

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

Powell, M. J. D. (1992). The theory of radial basis function approximation in 1990. In Light, W., editor, *Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms, and Radial Basis Functions*, chapter 3, pages 105–210. Clarendon Press, Oxford.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Rodieck, R. W. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, 5:583–601.

Sakurai, A. (1993). Tighter bounds of the VC-dimension of three layer networks. In *Proceedings of the World Congress on Neural Networks*, volume 3, pages 540–543. Erlbaum, Hillsdale, New Jersey.

Schläfli, L. (1901). *Theorie der vielfachen Kontinuität.* Zürcher & Furrer, Zürich. Reprinted in: Schläfli, L. (1950). *Gesammelte Mathematische Abhandlungen, Band I.* Birkhäuser, Basel.

Schmidhuber, J., Eldracher, M., and Foltin, B. (1996). Semilinear predictability minimization produces well-known feature detectors. *Neural Computation*, 8:773–786.

Schmitt, M. (2000). On the complexity of computing and learning with multiplicative neural networks. Technical Report 274, Fakultät für Mathematik der Ruhr-Universität Bochum.

Shawe-Taylor, J. (1995). Sample sizes for threshold networks with equivalences. *Information and Computation*, 118:65–72.

Tessier-Lavigne, M. (1991). Phototransduction and information processing in the retina. In Kandel, E. R., Schwartz, J. H., and Jessell, T. M., editors, *Principles of Neural Science*, chapter 28, pages 400–418. Prentice Hall, Englewood Cliffs, New Jersey, third edition.

Ward, V. and Syrzycki, M. (1995). VLSI implementation of receptive fields with current-mode signal processing for smart vision sensors. *Analog Integrated Circuits and Signal Processing*, 7:167–179.

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

Yasui, S., Furukawa, T., Yamada, M., and Saito, T. (1996). Plasticity of center-surround opponent receptive fields in real and artificial neural systems of vision. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 159–165, MIT Press, Cambridge, Mass.

Yee, P. and Haykin, S. (2001). *Regularized Radial-Basis Function Networks: Theory and Applications.* Wiley, New York.