FEEDFORWARD NETS FOR INTERPOLATION AND CLASSIFICATION[*]

Eduardo D. Sontag
Department of Mathematics, Rutgers University, New Brunswick, NJ 08903
E-mail: sontag@hilbert.rutgers.edu

January 1991, rev. June 1991; to appear in *J.Comp.Syst.Sci*

Abstract

This paper deals with single-hidden-layer feedforward nets, studying various aspects of classification power and interpolation capability. In particular, a worst-case analysis shows that direct input to output connections in threshold nets double the recognition but not the interpolation power, while using sigmoids rather than thresholds allows doubling both. For other measures of classification, including the Vapnik-Chervonenkis dimension, the effect of direct connections or sigmoidal activations is studied in the special case of two-dimensional inputs.

1 Introduction

In this paper we deal with the computational capabilities of certain interconnections of simple processors ("neurons"). These are arranged in a layered network, each processor calculating a scalar function θ (the *activation* or *response* function) of its aggregate input. Such interconnections, often called *feedforward neural nets*, have attracted interest as a potentially useful model of parallel computation. Figure 1 illustrates a typical network of this type.

The input fed to any given processor is an affine combination of the ouputs of all the processors that connect to it, weighted according to real-valued coefficients. The output of the last processor is taken as the output of the net. The weights, together with the interconnection pattern and the choice of θ, determine completely the function computed by the net. Sometimes one may want to allow *direct connections* from inputs to outputs, bypassing the intermediate layers. One such connection is indicated in Figure 1 with a dotted line. Studying the effect of such connections is one of our objectives. Precise definitions are given later, but we first wish to describe in simple terms some of the questions to be asked.

Here we shall be concerned exclusively with three-layer nets. That is, there are input nodes, an output node, and one layer in between. The processors in the intermediate layer are called *hidden units*. For simplicity, and since basic computing abilities are not affected in any significant way, we assume that the output processor computes just a linear combination of its input (no θ is applied, as illustrated in the Figure).

[*]Supported in part by US Air Force Grant AFOSR-880235 and by Siemens Corporate Research.

There are various types of activation functions that may be of interest. In theoretical computer science studies (circuit complexity) one deals most often with threshold gates, corresponding to a discontinuous function $\theta = \mathcal{H}$, the *Heaviside* function that takes the value 1 for positive arguments and is 0 otherwise. On the other hand, the so-called backpropagation technique so popular in practice assumes sigmoidal responses, as a differentiable θ is needed for applying gradient descent techniques. Sigmoidal functions θ are essentially smooth approximations to the Heaviside; an axiomatic definition is given later. Another of our objectives here is to compare the possible power of nets that use Heaviside activation units to those using sigmoids.

The capabilities that we are most interested in have to do with classification and interpolation power of the functions computed by feedforward nets. Nets with one hidden layer are known to be in principle sufficient for arbitrary recognition tasks. This follows from by now well-known approximation theorems (see e.g. [6], [7], but see also [4], [5], [14] for other problems where two layers are required instead). However, what is far less clear is *how many* processors are needed for achieving a given recognition, interpolation, or approximation objective. This is of importance in its practical aspect, since having rough estimates of how many processors will be needed is essential when applying backpropagation. It is also relevant when evaluating generalization properties, as larger nets tend to lead to poorer generalization; see the remarks in that regard in Section 4.1 below. It is well-known and easy to prove (see later) that one can interpolate values at any $n+1$ points using an n-processor net, and in particular that any $n+1$-point set can be partitioned arbitrarily into two classes by such nets. Among other facts, we point out here that allowing direct input to output connections permits doubling the recognition power to $2n$, and the same result is achieved if sigmoids are used but such direct connections are not allowed. Further, we remark that approximate interpolation of $2n-1$ points is also possible, provided that sigmoidal units be employed, but direct connections in threshold nets do not suffice in this case.

There are many alternative possible measures of recognition capabilities for nets. These range from the above-mentioned case of partitioning arbitrary sets to asking what is the cardinality of the largest set that can be arbitrarily partitioned using nets with a fixed architecture, the Vapnik-Chervonenkis or *VC* dimension of this architecture, as well as many other measures in between. We will study a few of these measures. In particular, it is known ([3]) that the VC dimension of threshold nets with a fixed number of hidden units is at least proportional to the number or inputs. If sigmoids or direct connections are allowed, we give lower bounds, for the two input case, at least doubling the VC dimension estimate known for Heaviside nets with no direct input to output connections.

One intuitive explanation for the apparent discrepancy between the fact that sigmoids approximate Heavisides but the former have richer approximation properties is due to the fact that the approximation in question is at what may be called "high gain," that is, for large incoming weights. For small weights, however, using sigmoids one can get also approximations to linear maps (the tangents). This adds a considerable amount of separation power. For interpolation, the intuition is different, and it is based on a continuity assumption on the sigmoid.

This paper is organized as follows. First we present basic definitions of the various classification measures and of nets. We then state the main results on classification, and after that we provide the proofs. These proofs combine simple combinatorial and geometric arguments. A further section shows that some of the bounds obtained are sharp in the case of a piecewise-linear activation function. Then we deal with issues of interpolation. We prove general results as well as study several particular activation functions. The last section summarizes the main conclusions and poses open problems.

To close this introduction, we wish to remark that this is a continuation of previous work dealing with the theme of comparing threshold and sigmoidal feedforward nets. In [15] and [16] we studied nets with *no* hidden layers. In the first reference we proved that the gradient descent procedure may get stuck in spurious local minima, and in the second we compared this numerical procedure to the classical perceptron learning technique used for Heaviside nets, proving a global convergence theorem under hypothesis analogous to those used for the perceptron algorithm.

A preliminary version of this paper, without proofs, appeared in [13].

2 Dichotomies

Quantifying the classification power of a class of functions (such as those computable by nets with a fixed architecture and a fixed number of processors) can be based on the idea of "shattering" of sets, described next. In this approach, a class of functions is considered to be more powerful than another if it can be used to implement arbitrary partitions on sets of larger cardinality. This is made precise as follows.

Fix a positive integer N. A *dichotomy* or *two-coloring* (S_-, S_+) on a set $S \subseteq \mathbb{R}^N$ is a partition $S = S_- \bigcup S_+$ of S into two disjoint subsets. A function $f : \mathbb{R}^N \to \mathbb{R}$ will be said to *implement* this dichotomy if it holds that

$$f(u) > 0 \text{ for } u \in S_+ \;\text{ and }\; f(u) < 0 \text{ for } u \in S_- \;.$$

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^N$ to $\mathbb{R}$, assumed to be nontrivial, in the sense that for each point $u \in \mathbb{R}^N$ there is some $f_1 \in \mathcal{F}$ so that $f_1(u) > 0$ and some $f_2 \in \mathcal{F}$ so that $f_2(u) < 0$. This class *shatters* the set $S \subseteq R^N$ if each dichotomy on S can be implemented by some $f \in \mathcal{F}$.

For any class of functions $\mathcal{F}$ as above, we consider here the following measures of classification power. First we introduce $\overline{\mu}$ and $\underline{\mu}$, dealing with "best" and "worst" cases respectively:

$$\overline{\mu}(\mathcal{F})$$

denotes the largest integer $l \geq 1$ (possibly ∞) so that there is at least *some* set S of cardinality l in $\mathbb{R}^N$ which can be shattered by $\mathcal{F}$, while

$$\underline{\mu}(\mathcal{F})$$

is the largest integer $l \geq 1$ (possibly ∞) so that *every* set of cardinality l can be shattered by $\mathcal{F}$. Note that by definition, $\underline{\mu}(\mathcal{F}) \leq \overline{\mu}(\mathcal{F})$ for every class $\mathcal{F}$.

In particular, the definitions imply that no set of cardinality $\overline{\mu}(\mathcal{F}) + 1$ can be shattered, and that there is at least some set of cardinality $\underline{\mu}(\mathcal{F}) + 1$ which cannot be shattered. The integer $\overline{\mu}$ is usually called the *Vapnik-Chervonenkis (VC) dimension* of the class $\mathcal{F}$ (see for instance [3]), and appears in formalizations of learning in the distribution-free sense.

A set may fail to be shattered by $\mathcal{F}$ because it is very special (see the example below with colinear points). In that sense, a more robust measure is useful:

$$\mu(\mathcal{F})$$

is the largest integer $l \geq 1$ (possibly ∞) for which the class of sets S that can be shattered by $\mathcal{F}$ is dense, in the sense that given every l-element set $S = \{s_1, \ldots, s_l\}$ there are points $\widetilde{s}_i$

arbitrarily close to the respective s_i's such that $\widetilde{S} = \{\widetilde{s}_1, \ldots, \widetilde{s}_l\}$ can be shattered by $\mathcal{F}$. Note that

$$\underline{\mu}(\mathcal{F}) \leq \mu(\mathcal{F}) \leq \overline{\mu}(\mathcal{F}) \tag{1}$$

for all $\mathcal{F}$.

To obtain an upper bound m for $\mu(\mathcal{F})$ one needs to exhibit an open class of sets of cardinality $m+1$ none of which can be shattered.

Take as an example the class $\mathcal{F}$ consisting of all affine functions $f(x) = ax + by + c$ on $\mathbb{R}^2$. Since any three points can be shattered by an affine map provided that they are not colinear (just choose a line $ax + by + c = 0$ that separates any point which is colored different from the rest), it follows that $3 \leq \mu$. On the other hand, no set of four points can ever be dichotomized, which implies that $\overline{\mu} \leq 3$ and therefore the conclusion

$$\mu = \overline{\mu} = 3$$

for this class. (The negative statement can be verified by a case by case analysis: if the four points form the vertices of a 4-gon color them in "XOR" fashion, alternate vertices of the same color; if 3 form a triangle and the remaining one is inside, color the extreme points differently from the remaining one; if all colinear then use an alternating coloring). Finally, since there is some set of 3 points which cannot be dichotomized (any set of three colinear points is like this), but every set of two can,

$$\underline{\mu} = 2\ .$$

We shall say that $\mathcal{F}$ is *robust* if whenever S can be shattered by $\mathcal{F}$ also every small enough perturbation of S can be shattered. (More precisely, one defines a topology on unordered l-sets as follows: if S is a set of l elements, then a basis of open neighborhoods of S is given by the class of all sets of the form $\widetilde{S} = \{\widetilde{s}_1, \ldots, \widetilde{s}_l\}$ so that $|s_i - \widetilde{s}_i| < \varepsilon$ for each i. Then robustness of $\mathcal{F}$ means that for each l, the class of l-element sets that can be shattered is open.) For a robust class and $l = \mu(\mathcal{F})$, every set in an open dense subset in the above topology, i.e. *almost every* set of l elements, can be shattered. All classes considered in this note are robust. If the elements of $\mathcal{F}$ are continuous, then $\mathcal{F}$ is robust.

3 Nets

A "net" is a function of a certain type, corresponding to the idea of feedforward interconnections, via additive links, of processors each of which has a scalar response or *activation function* θ.

Definition 3.1 Let $\theta : \mathbb{R} \to \mathbb{R}$ be any function. A function $f : \mathbb{R}^N \to \mathbb{R}$ is *computable by a single-hidden-layer net with k hidden processors of type θ and N inputs*, or just f *is a (k, θ)-net*, if there are real numbers

$$w_0, w_1, \ldots, w_k, \tau_1, \ldots, \tau_k$$

and vectors

$$v_0, v_1, \ldots, v_k \in \mathbb{R}^N$$

such that, for all $u \in \mathbb{R}^N$,

$$f(u) \ = \ w_0 \ + \ v_0 \cdot u \ + \ \sum_{i=1}^{k} w_i \, \theta(v_i \cdot u - \tau_i) \tag{2}$$

where the dot indicates inner product. *A net with no direct i/o connections* is one for which $v_0 = 0$.

For fixed θ, and under mild assumptions on θ, such nets can be used to approximate uniformly arbitrary continuous functions on compacts. See for instance [6], [7]. In particular, they can be used to implement arbitrary dichotomies. A typical choice is for θ to be the *standard sigmoid*

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

or equivalently, up to translations and change of coordinates, the hyperbolic tangent $\tanh(x)$. Another usual choice is the hardlimiter, threshold, or *Heaviside* function

$$\mathcal{H}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

which can be approximated well by $\sigma(\gamma x)$ when the "gain" γ is large. Yet another possibility is the use of the piecewise linear function

$$\pi(x) = \begin{cases} -1 & \text{if } x \leq -1 \\ 1 & \text{if } x \geq 1 \\ x & \text{otherwise.} \end{cases} \tag{4}$$

Most analysis has been done for $\mathcal{H}$ and no direct connections, but numerical techniques typically use the standard sigmoid (3) (or equivalently tanh). The activation π will be useful as an example for which sharper bounds can be obtained. The examples σ and π, but not $\mathcal{H}$, are particular cases of the following more general type of activation function:

Definition 3.2 A function $\theta : \mathbb{R} \to \mathbb{R}$ will be called a *sigmoid* if these two properties hold:

(S1) $t_+ := \lim_{x\to+\infty} \theta(x)$ and $t_- := \lim_{x\to-\infty} \theta(x)$ exist, and $t_+ \neq t_-$.

(S2) There is some point c such that θ is differentiable at c and $\theta'(c) = \eta \neq 0$. □

Note that we do not require, as it is not needed for the results to be given, that a sigmoid be monotonic.

Remark 3.3 Property (S1) could be replaced by the stronger property that in addition $t_+ = 1$ and $t_- = 0$. This would not change anything, because for any θ satisfying (S1),

$$\widetilde{\theta}(x) := \frac{\theta(x) - t_-}{t_+ - t_-}$$

satisfies the stronger property and is so that (k, θ)-nets are the same as $(k, \widetilde{\theta})$-nets. Similarly, when convenient we may asssume without loss of generality that $t_+ = 1$ and $t_- = -1$. □

Remark 3.4 All the examples above lead to robust classes, in the sense defined earlier. More precisely, assume that θ is continuous except for at most finitely many points x, and it is left continuous at such x, and let $\mathcal{F}$ be the class of (k, θ)-nets, for any fixed k. Then $\mathcal{F}$ is robust, and the same statement holds for nets with no direct connections. This can be proved as follows. Assume that $S = \{s_1, \ldots, s_l\}$ can be shattered by $\mathcal{F}$, and consider any fixed dichotomy and any

$f \in \mathcal{F}$ implementing it. For this f, take all the points $a_{ij} := v_i.s_j - \tau_i$, $i = 1, \ldots, k$, $j = 1, \ldots, l$. By the assumption on θ, there is some small enough $\varepsilon > 0$ so that $a_{ij} - \varepsilon$ is a point of continuity for θ for all i, j, and

$$\widetilde{f}(u) \;=\; w_0 \,+\, v_0 \cdot u \,+\, \sum_{i=1}^{k} w_i\, \theta(v_i \cdot u - \tau_i - \varepsilon)$$

still implements the dichotomy. Since $\widetilde{f}$ is continuous near each point of S, the corresponding dichotomy will be implementable for each set sufficiently close to S. Intersecting over all possible dichotomies, there results a neighborhood of S in which every set can be shattered. If $v_0 = 0$, one has the same result for the case of no direct connections. □

4 Statement of Classification Results

We let

$$\mu(k, \theta, N)$$

denote $\mu(\mathcal{F})$, where $\mathcal{F}$ is the class of (k, θ)-nets in $\mathbb{R}^N$ with *no direct connections*, and similarly for $\underline{\mu}$ and $\overline{\mu}$, and a superscript d is used for the class of arbitrary such nets (with possible direct connections from input to output). The lower measure $\underline{\mu}$ is independent of dimension:

Lemma 4.1 For each k, θ, N, $\underline{\mu}(k, \theta, N) = \underline{\mu}(k, \theta, 1)$ and $\underline{\mu}^d(k, \theta, N) = \underline{\mu}^d(k, \theta, 1)$.

This justifies denoting these quantities just as $\underline{\mu}(k, \theta)$ and $\underline{\mu}^d(k, \theta)$ respectively, as we do from now on, and giving proofs only for $N = 1$. The easy Lemma 4.1 and the next remark are both proved in Section 5 below.

Lemma 4.2 For any sigmoid θ, and for each k, N,

$$\mu(k+1, \theta, N) \geq \mu^d(k, \mathcal{H}, N)$$

and similarly for $\underline{\mu}$ and $\overline{\mu}$.

The main results on classification will be as follows.

Theorem 1 *For any sigmoid θ, and for each k,*

$$\begin{aligned} \underline{\mu}(k, \mathcal{H}) &= k + 1 \\ \underline{\mu}^d(k, \mathcal{H}) &= 2k + 2 \\ \underline{\mu}(k, \theta) &\geq 2k\ . \end{aligned}$$

Theorem 2 *For each k,*

$$\begin{aligned} 4\left\lfloor \frac{k}{2} \right\rfloor \leq \mu(k, \mathcal{H}, 2) &\leq 2k + 1 \\ \mu^d(k, \mathcal{H}, 2) &\leq 4k + 3\ . \end{aligned}$$

Theorem 3 *For any sigmoid θ, and for each k,*

$$\begin{array}{rcl} 2k+1 & \leq & \overline{\mu}(k,\mathcal{H},2) \\ 4k+3 & \leq & \overline{\mu}^d(k,\mathcal{H},2) \\ 4k-1 & \leq & \overline{\mu}(k,\theta,2)\ . \end{array}$$

These results are proved below. The first inequality in Theorem 2 follows from the results of Baum ([3]), who in fact established a lower bound of $2N\lfloor\frac{k}{2}\rfloor$ for $\mu(k,\mathcal{H},N)$ (and hence for $\overline{\mu}$ too), for every N, not just $N=2$ as in the Theorem above.

Because of Lemma 4.2, the last statements in Theorems 1 and 3 are consequences of the previous two.

4.1 A Simple Example on Generalization

The relevance of the above results to questions of learning and generalization can be illustrated through a simple and intuitive example. A careful analysis of the issues involved belongs to the realm of learning theory, but the example should be sufficient to illustrate how the choice of processors influences the generalization capabilities of nets.

Assume given "training data" consisting of a set of points in the real axis labeled "O" or "X" as in Figure 2 (ignore for now the question mark). Suppose that a learning algorithm has succeded in loading this data into a two-processor Heaviside-net with no direct input to output connections, that is, in finding some $(2,\mathcal{H})$-net f with no connections that implements the indicated dichotomy (say, positive at the X's and negative at the O's). Observe that no possible $(1,\mathcal{H})$-net can load the data, but there are infinitely many possible such $(2,\mathcal{H})$-nets. The relevant fact for our example is that *all* of these $(2,\mathcal{H})$-nets will give the *same* generalization at a new point placed in the position indicated by the question mark in the Figure. All of them will classify this point as "X" as follows from the general arguments in the paper.

On the other hand, if as in standard numerical approaches one uses sigmoidal processors and one finds a $(2,\theta)$-net that loads the same training data, the generalization is not unique. There are some such nets (assuming that θ satisfies the properties (S1) and (S2)) for which $f>0$ where the question mark is, but there are also infinitely many such nets for which $f<0$ there. The actual generalization will depend on the initial conditions used in the gradient descent ("backpropagation") algorithm, and not on intrinsic properties of the data.

5 Some Basic Lemmas

We now prove Lemmas 4.1 and 4.2. The former is an immediate consequence of this fact, applied with $\mathcal{F}_N = (k,\theta)$-nets on N inputs:

Lemma 5.1 Let $\{\mathcal{F}_N, N\geq 1\}$ be a class of functions so that $g(u)=f(u,0,\ldots,0)\in\mathcal{F}_1$ whenever $f\in\mathcal{F}_N$ and so that $g(u)=f(v\cdot u)$ is in $\mathcal{F}_N$ whenever $f\in\mathcal{F}_1$ and v is any fixed vector in $\mathbb{R}^N$. Then, $\underline{\mu}(\mathcal{F}_N)=\underline{\mu}(\mathcal{F}_1)$ for all N.

Proof. Pick any N and any function θ. Take any subset $S\subseteq\mathbb{R}$ of cardinality at most $\underline{\mu}(\mathcal{F}_N)$, and any dichotomy (S_-,S_+) on this set. Consider now the set $\widetilde{S}_- := \{(u,0,\ldots,0)\in\mathbb{R}^N \mid u\in S_-\}$, and similarly $\widetilde{S}_+$ and $S = S_- \bigcup S_+$. Let $f \in \mathcal{F}_N$ implement the dichotomy $(\widetilde{S}_-, \widetilde{S}_+)$. Then $f(\cdot, 0, \ldots, 0) : \mathbb{R} \to \mathbb{R}$ implements the original dichotomy. Thus $\underline{\mu}(\mathcal{F}_1) \geq \underline{\mu}(\mathcal{F}_N)$.

Conversely, take any dichotomy (S_-, S_+) on $S \subseteq \mathbb{R}^N$, with S of cardinality $l \leq \underline{\mu}(\mathcal{F}_1)$. If S consists of the distinct vectors $u_1, \ldots, u_l$, then there exists some vector v so that all the numbers $v \cdot u_i$ are distinct (in fact, a random vector will have this property with probability one). Indeed, the set of all such separating v's is the intersection of the finitely many open dense sets

$$\{v \,|\, v \cdot (u_i - u_j) \neq 0\}$$

(one for each $i \neq j$) and is hence nonempty. Pick any v like this, and let $y_i := v \cdot u_i$ for each i. The dichotomy (S_-, S_+) induces a dichotomy of the set $\widetilde{S} = \{y_i, i = 1, \ldots, l\}$ corresponding to values $v \cdot u_i, u_i \in S_+$ and $v \cdot u_i, u_i \in S_-$. Let f be a (k, θ)-net that implements the dichotomy on $\widetilde{S}$. Then $g(u) := f(v \cdot u)$ is a (k, θ)-net which implements the original dichotomy. Thus $\underline{\mu}(\mathcal{F}_N) \geq (\mathcal{F}_1)$. ∎

In order to prove Lemma 4.2, we need a few simple facts. The first two are basically just restatements of (S1) and (S2) respectively:

Lemma 5.2 Let θ be a sigmoid, with $t_+ = 1$ and $t_- = 0$, and pick any compact subset $K \subseteq \mathbb{R}$ not containing zero. Then,

$$\theta(\lambda x) - \mathcal{H}(x) \to 0 \ \text{ as } \ \lambda \to +\infty$$

uniformly on $x \in K$.

Proof. There is some $\alpha > 0$ such that $\alpha \leq |x|$ for all $x \in K$. By property (S1), there is for each $\varepsilon > 0$ some $\eta > 0$ such that $|\theta(y) - \mathcal{H}(y)| < \varepsilon$ if $|y| > \eta$. Thus if $\lambda > \eta/\alpha$ it follows that also $|\theta(\lambda x) - \mathcal{H}(\lambda x)| < \varepsilon$ for all $x \in K$; since $\mathcal{H}(\lambda x) = \mathcal{H}(x)$, the desired conclusion follows. ∎

Lemma 5.3 Let c, η be constants as in (S2), and assume that $K \subseteq \mathbb{R}$ is compact. Then

$$\frac{\lambda}{\eta}\left[\theta\left(\frac{1}{\lambda}x - \frac{c - c\lambda}{\lambda}\right) - \theta(c) + \frac{\eta}{\lambda}c\right] - x \quad \to \quad 0 \quad \text{as} \ \ \lambda \to +\infty$$

uniformly on $x \in K$.

Proof. Multiplying by η, we need to show that

$$\frac{1}{1/\lambda}\left[\theta\left(c + \frac{x-c}{\lambda}\right) - \theta(c) - \eta\Big(\frac{x-c}{\lambda}\Big)\right] \quad \to \quad 0$$

uniformly on $x \in K$. The numbers $\varepsilon = \frac{x-c}{\lambda}$ are small as $\lambda \to +\infty$, uniformly on K; thus what is desired is that

$$\theta(c + \varepsilon) - \theta(c) - \eta\varepsilon = o(\varepsilon)$$

which is precisely property (S2). ∎

Now Lemma 4.2 follows from the following one:

Lemma 5.4 If f is a $(k, \mathcal{H})$-net, S is a finite subset of $\mathbb{R}^N$, θ is a sigmoid, and $\varepsilon > 0$ is given, then there is a $(k+1, \theta)$-net g with no direct input to output connections so that $|f(u) - g(u)| < \varepsilon$ for each $u \in S$.

Proof. Without loss of generality (cf. Remark 3.3), we assume $t_+ = 1$ and $t_- = 0$. Arguing as in Remark 3.4, we can assume that the expressions $v_i \cdot s_j - \tau_i$ are all nonzero. Let K be the set consisting of all these expressions. For each term of the form

$$w_i \,\mathcal{H}(v_i \cdot u - \tau_i)$$

we can use an approximation

$$w_i \,\theta(\lambda v_i \cdot u - \lambda \tau_i)$$

for large enough λ, by Lemma 5.2. For the linear term $v_0 \cdot u$, on the other hand,

$$v_0 \cdot u \;\approx\; -\frac{\lambda}{\eta}\theta(c) + c - \frac{\lambda}{\eta}\theta\Big(\frac{1}{\lambda}v_0 \cdot u - \frac{c - c\lambda}{\lambda}\Big)$$

for large λ, by Lemma 5.3. ∎

6 Upper Bounds

We first establish the upper bounds $\underline{\mu}(k, \mathcal{H}) \leq k + 1$ and $\underline{\mu}^d(k, \mathcal{H}) \leq 2k + 2$. Let f be as in (2) with $\theta = \mathcal{H}$. The points $u \in \mathbb{R}$ where

$$v_i u = \tau_i, v_i \neq 0, i = 1, \ldots, k,$$

determine a partition of $\mathbb{R}$ into at most $k + 1$ disjoint intervals where f must be linear, or constant in the case of no direct connections. Thus in the latter case there can be at most k sign alternations, and in the first (direct connections allowed) at most $2k+1$ (since in each of the $k + 1$ intervals, f can alternate sign at most once). Consider the set of points $S = \{0, 1, \ldots, l\}$, and color these in an alternating fashion:

$$S_+ := \{0, 2, \ldots, l\}\ , \quad S_- := \{1, 3, \ldots, l - 1\}$$

(if l even; similarly if l is odd). There are l sign alternations. If f implements this dichotomy on S and is of the above form, it must follow from the above discussion that $l \leq 2k + 1$ (or $l \leq k$ if there are no direct connections). Using $l = 2k + 2$ (or $k + 1$ respectively), we conclude that there is a set of cardinality $2k + 3$ (or $k + 2$ respectively) which cannot be shattered, and the bounds are proved.

Consider now a set S of l points arranged as the vertices of a regular l-gon in the plane, and assume that l is even. Dichotomize S by using an alternating coloring. If f is a $(k, \mathcal{H})$-net, let $L_1, \ldots, L_k$ be the lines $v_i \cdot u = \tau_i$, $i = 1, \ldots, k$. (Assume all v_i are nonzero; otherwise this just contributes to the constant term, with a smaller k.) Each line L_i crosses the l-gon in at most two edges; perturbing weights if needed, we may assume that no L_i passes through a vertex.

If f has no direct connections ($v_0 = 0$), it must have the same value on any two adjacent vertices that are not separated by some L_i, contradicting the fact that f implements the dichotomy unless every edge is so separated. Thus if $l = 2k + 2$ it is impossible to dichotomize with no direct connections. The same argument works with all small enough perturbations of the set S, so

$$\mu(k, \mathcal{H}, 2) \leq 2k + 1$$

as desired for Theorem 2.

If instead direct connections are allowed, we argue as follows, with the same set S and the same dichotomy. Consider the connected components of

$$\mathbb{R}^2 \setminus \bigcup_{i=1}^{k} L_i \ .$$

Since each line L_i crosses the l-gon in at most two edges, there are at most $2k$ "segments" of successive vertices in each component. Assume that some three successive vertices u_1, u_2, u_3 are in the same component. The restriction of f to this component is a linear map, so the only way for f to implement the above dichotomy on S is if the zero locus of f crosses both edges (u_1, u_2) and (u_2, u_3). (In addition, there cannot be any set of four such successive vertices, as these cannot be separated linearly.) Moreover, this zero locus is of the form

$$w_0 \, + \, v_0 \cdot u \, + \, \gamma \ = \ 0$$

where γ is a constant that depends on the particular component. Since all these lines are parallel, there can be at most *two* of them. Thus the vertices are arranged into at most two segments of three successive vertices plus at most $2k-2$ segments of at most 2 vertices each. It follows that

$$l \leq 2(3) + 2(2k-2) = 4k+2 \ .$$

Consider the sets of cardinality $l = 4k+4$ obtained as small perturbations of the above S. By an analogous argument, none of these can be shattered by $(k, \mathcal{H})$-nets. So there is some open class of sets of that cardinality none of which can be shattered, which implies that

$$\mu^d(k, \mathcal{H}, 2) \leq 4k+3$$

as desired for Theorem 2.

7 Lower Bounds

We use the notation "$I < J$" for intervals to mean that $x < y$ whenever $x \in I$ and $y \in J$. A trivial but useful technical result is as follows.

Proposition 7.1 Let $I_1 < J_1 < I_2 < J_2 < \ldots < I_k < J_k$ be closed finite subintervals of $\mathbb{R}$, and denote $I := \bigcup I_i$, $J := \bigcup J_i$. Then there exists a $(k-1, \mathcal{H})$-net f so that $f(x) > 0$ for $x \in I$ and $f(x) < 0$ for $x \in J$.

Proof. Let $c_1, \ldots, c_k$ and $\tau_1, \ldots, \tau_{k-1}$ be any numbers separating the intervals:

$$I_1 < c_1 < J_1 < \tau_1 < I_2 < c_2 < J_2 < \tau_2 < I_3 < \ldots < \tau_{k-1} < I_k < c_k < J_k$$

(with the obvious notation). Now pick $w_0 := c_1$ and for $i = 1, \ldots, k-1$, $w_i := c_{i+1} - c_i$. Finally, let $v_0 := -1$ and $v_i := 1$ for $i = 1, \ldots, k-1$. This gives rise to a $(k-1, \mathcal{H})$-net f. Since for any $x \in I_i \bigcup J_i$ it holds that $f(x) = c_i - x$, the desired property holds. ∎

We now prove the first two conclusions in Theorem 1. The upper bounds were already established, so it is necessary to show that any $2k+2$ or $k+1$-element set can be dichotomized, depending on whether direct connections are allowed or not.

Let $S \subseteq \mathbb{R}$ have cardinality $l = 2k$, $S = \{y_1, \ldots, y_l\}$, with $y_1 < y_2 < \ldots < y_l$. Now take any dichotomy (S_-, S_+) of S. We shall assume that $y_1 \in S_+$; otherwise the argument is the same (multiply the obtained f by -1). Thus there are disjoint closed finite intervals as in the statement of Proposition 7.1 such that each $y_i \in S_+$ is in some interval I_j and each $y_i \in S_-$ is in some interval J_j. Any f as in the conclusion of the Lemma dichotomizes. This completes the proof that $\underline{\mu}^d(k, \mathcal{H}, N) = 2k + 2$.

In the case of no direct connections, we use a set of cardinality $l = k + 1$. The same construction reduces the problem to the separation of intervals

$$I_1 < J_2 < I_3 < J_4 < \ldots < I_k < J_{k+1}$$

and this can be easily achieved with a combination of k Heavisides, proving $\underline{\mu}(k, \mathcal{H}, N) = k + 1$.

We now indicate how to prove the first two statements in Theorem 3. These are consequences of a result that appeared in [2], which we cite next:

Result. Pick any integer $n \geq 1$. Let S be the set consisting of the vertices of the convex regular n-gon in the plane. Assume that a dichotomy of S is given. Then, there exists some vector v such that the dot products $v.u$, $u \in S$, fall into at most

$$\left\lfloor \frac{n}{2} \right\rfloor + 1$$

intervals such that each interval contains only elements of the type $v \cdot u$, $u \in S_-$ or only elements of the type $v \cdot u$, $u \in S_+$. ∎

Take any k, and apply this result with $n := 4k + 3$. There result $2k + 2$ intervals. By Lemma 7.1, the intervals can be separated by some $(k, \mathcal{H})$-net (with direct connections), and again $g(u) := f(v \cdot u)$ can be used. For nets with no direct connections, the same argument can be applied using $n := 2k + 1$ points; the resulting $k + 1$ intervals can be separated using k processors and no connections.

8 Some Particular Activation Functions

Consider the last inequality in Theorem 1. For arbitrary sigmoids, this is far too conservative, as the number $\underline{\mu}$ can be improved considerably from $2k$, even made infinite (see below). We conjecture that for the important practical case $\theta(x) = \sigma(x)$ it is close to optimal, but the only upper bounds that we have are still too high. For the piecewise linear function π, at least, one has equality:

Lemma 8.1 $\underline{\mu}(k, \pi) = 2k$.

Proof. To prove this fact it is enough to show that

$$f(x) \;=\; w_0 \;+\; \sum_{i=1}^{k} w_i \, \pi(\gamma_i x - \tau_i)$$

can not implement the dichotomy of $\{1, 2, \ldots, l\}$ into odds and evens unless $2k \geq l$. Since f is continuous, this will in turn follow from the fact that f cannot be zero on more than $2k - 1$ disjoint closed intervals.

Indeed, assume without loss of generality that all $\gamma_i > 0$ (terms with $\gamma_i = 0$ can be absorbed into w_0), and let T be the set of all numbers of the form $\frac{1}{\gamma_i}(\tau_i - 1)$ and $\frac{1}{\gamma_i}(\tau_i + 1)$,

$$T := \{t_1, \ldots, t_{2k}\}$$

with $t_1 \leq t_2 \leq \ldots$. On each interval $[t_i, t_{i+1}]$, as well as on $(-\infty, t_1]$ and $[t_{2k}, +\infty)$, the function f is linear.

Assume that f vanishes precisely on the disjoint closed intervals $I_j, j = 1, \ldots, m$. If for some j, j' it were the case that both I_j and $I_{j'}$ intersect one of the above intervals of linearity, then f would have to be identically zero on that interval and therefore $j = j'$. Furthermore, f is constant on $(-\infty, t_1]$ and $[t_{2k}, +\infty)$, so if any I_j intersects one of these it follows that no other $I_{j'}$ can intersect $[t_1, t_2]$ or $[t_{2k-1}, t_{2k}]$ respectively. In conclusion, $m \leq 2k - 1$, as wanted. ∎

We show next that there exist sigmoids θ, as differentiable as wanted, even real-analytic, where all classification measures are infinite. Of course, such a function θ must necessarily be so complicated that there is no reasonably "finite" implementation for it. This remark is mainly of theoretical interest, to indicate that, unless further restrictions are made on (S1)-(S2), far better bounds can be obtained.

Lemma 8.2 There is some sigmoid θ, which can be taken to be an analytic function, so that $\underline{\mu}(1, \theta) = \infty$.

Proof. First consider all possible ordered sequences of rational numbers

$$\sigma_i = (q_1^i, \ldots, q_{n_i}^i), \quad n_i \geq 1, \quad q_1^i < \ldots < q_{n_i}^i$$

enumerated in any fixed way. Next define $\rho_1 := 1$ and pick a sequence $\{\rho_r\}$ so that, for every $l = 1, \ldots, n_r$,

$$y_l^r := \rho_r e^{q_l^r} > \rho_i e^{q_j^i} + 1$$

for all $i = 1, \ldots, r - 1$ and $j = 1, \ldots, r_i$. By construction.

$$y_1^1 < \ldots < y_{n_1}^1 < y_1^2 < \ldots < y_{n_2}^2 < y_1^3 < \ldots$$

and the set of all y_j^i's is a discrete subset of $\mathbb{R}$. One can then construct basically the infinite product of the monomials $(1 - x/y_j^i)$, multiplied by suitable exponential functions to guarantee convergence, see [12], Theorem 15.9; this results in a real-analytic function α which has simple zeroes at the y_j^i's and no other zeroes. It follows that α alternates sign on the intervals between consecutive y_j^i's, since otherwise a local minimum would result at some y_j^i and hence a zero of multiplicity at least two. All $y_j^i \neq 0$, so $\alpha(0) \neq 0$. Composing if necessary with tanh, we may and will assume that α is bounded. Define

$$\theta(x) := \frac{1}{1 + e^x} \alpha(e^x) .$$

Since α is bounded,

$$\lim_{x \to +\infty} \theta(x) = 0 \neq \alpha(0) = \lim_{x \to -\infty} \theta(x)$$

and therefore θ satisfies (S1)-(S2).

Now let $I_1 < J_1 < I_2 < J_2 < \ldots < I_k < J_k$ be as in Proposition 7.1. We wish to show that there is some number τ so that $f(x) = \theta(x-\tau)$ is positive on the I_i's and negative on the J_i's —or viceversa, in which case $-\theta(x-\tau)$ will be the desired net instead. Pick any $2k+1$ rational numbers separating the intervals,

$$q_1 < I_1 < q_2 < J_1 < \ldots < J_k < q_{2k+1}$$

and let σ_i be the sequence $(q_1, \ldots, q_{2k+1})$. Denote $\rho := \rho_i$. Since α has constant and alternating signs on the intervals $(\rho e^{q_i}, \rho e^{q_{i+1}})$, and since ρe^x is in such an interval whenever $x \in (q_i, q_{i+1})$, the desired conclusion follows taking $\tau := -\ln(\rho)$. ∎

The above construction was somewhat complicated because we wanted $\underline{\mu}(1,\theta) = \infty$. If only μ and $\overline{\mu}$ are desired to be infinite, one may also take far simpler examples, such as $\cos x$ —modified slightly in order to obtain property (S1). More interesting perhaps is the fact that one may find such examples, with infinite VC dimension, even if the extra requirement that θ be *strictly increasing* is also imposed. For instance, consider

$$\theta(x) \;:=\; \frac{1}{\pi}\arctan x \;+\; \frac{\cos x}{\alpha(1+x^2)} \;+\; \frac{1}{2} \tag{5}$$

where α is any fixed number larger than 2π. This function has limits $1, 0$ at $\pm\infty$, and is (real-)analytic. Moreover, its derivative is everywhere positive, since it can be written as

$$\frac{1}{\alpha\pi(1+x^2)^2}\left[\frac{\alpha}{2}s(x) \;+\; (x^2+1)\left(\frac{\alpha}{2} - \pi\sin x\right)\right]$$

where

$$s(x) \;=\; \left(x^2 - \frac{4\pi x\cos x}{\alpha} + 1\right) \;,$$

and this last expression is itself always positive because as a quadratic form in x its discriminant satisfies

$$\Delta/4 \;=\; \left(\frac{2\pi\cos x}{\alpha}\right)^2 \;-\; 1 \;< 0 \;.$$

A plot of this function θ (with $\alpha = 100$) is given in Figure 3.

We now prove that

$$\mu(2,\theta,1) = \infty$$

(so also the VC dimension $\overline{\mu}(2,\theta,1) = \infty$) for the function in Equation (5). Consider the auxiliary function

$$\rho(x) \;:=\; \theta(x) + \theta(-x) - 1 \;=\; \frac{2\cos x}{\alpha(1+x^2)} \tag{6}$$

and fix an arbitrary positive integer l. We need to obtain a dense class of sets S, each having cardinality l, such that each S can be shattered by $(1,\rho)$-nets (and hence also by $(2,\theta)$-nets, which is the desired conclusion).

Indeed, take any set S consisting of rationally independent points $x_1, \ldots, x_l$. From, e.g., Theorem 3.2 and Lemma 2.7 in [10], we may conclude that the values of $(wx_1, \ldots., wx_l)$ modulo 2π are dense in $[0, 2\pi]^l$, as w ranges over $\mathbb{R}$ (in fact, even restricting to positive integer multiples w one would still obtain density). It follows that the vectors of the form

$$(\cos(wx_1), \ldots, \cos(wx_l))$$

form a dense subset of $[-1,1]^l$. Thus the vector

$$(\rho(wx_1),\ldots,\rho(wx_l))$$

can achieve any desired sequence of signs, by picking appropriate weights w. This gives the shattering result.

9 Interpolation

In this Section we deal with the following approximate interpolation problem. Given a sequence of k (distinct) points $u_1,\ldots,u_k$ in R^N, any $\varepsilon>0$, and any sequence of real numbers $y_1,\ldots,y_k$, as well as some class $\mathcal{F}$ of functions from $\mathbb{R}^N$ to $\mathbb{R}$, we ask if there exists some

$$f\in\mathcal{F} \text{ so that } |f(u_i)-y_i|<\varepsilon \text{ for each } i\,. \tag{7}$$

Let

$$\underline{\lambda}(\mathcal{F})$$

be the largest integer $k\geq 1$, possibly infinite, so that for every set of data as above (7) can be solved. Note that, obviously,

$$\underline{\lambda}(\mathcal{F}) \;\leq\; \underline{\mu}(\mathcal{F})\,. \tag{8}$$

We may also introduce $\lambda(\mathcal{F})$ and $\overline{\lambda}(\mathcal{F})$ in a manner analoguous to that of μ and $\overline{\mu}$. However, no nontrivial results will be provided for them except for some relatively minor remarks.

By exactly the same argument as in proving Lemma 5.1, $\underline{\lambda}$ is independent of the dimension N when applied to nets. Thus we let $\underline{\lambda}^d(k,\theta)$ and $\underline{\lambda}(k,\theta)$ be respectively the values of $\underline{\lambda}(\mathcal{F})$ when applied to (k,θ)-nets with or without direct connections, for any input dimension, and we always assume in proofs that $N=1$.

9.1 Interpolating With k Processors

We first remark that the inequality

$$\underline{\lambda}(k,\theta) \;\geq\; k+1 \tag{9}$$

holds under minimal assumptions on θ. Moreover, *exact* interpolation at $k+1$ points with a (k,θ)-net is in general possible. (See for instance [1], [11] for previous proofs of this result, under somewhat more restrictive assumptions.) The following very easy technical fact is all that is required; it says that for any linear class of functions which solves the approximate interpolation problem on a set S, it is enough to use k generators of this class in order to interpolate at k points.

Lemma 9.1 Let $\{f_\delta:\mathbb{R}^N\to\mathbb{R},\ \delta\in\Delta\}$ be a set of functions, and let $\mathcal{F}$ be the linear space of functions $\mathbb{R}^N\to\mathbb{R}$ spanned by this set. Let

$$S=\{u_1,\ldots,u_k\}$$

be a finite set of points in $\mathbb{R}^N$ so that the following property holds: for each $\varepsilon>0$ and each sequence of real numbers $y_1,\ldots,y_k$, there is some $f\in\mathcal{F}$ so that (7) holds. Then the following stronger property also holds: For every sequence of real numbers $y_1,\ldots,y_k$, there are

$$\delta_1,\ldots,\delta_k\in\Delta \quad\text{and}\quad w_1,\ldots,w_k\in\mathbb{R}$$

such that, writing

$$f \ := \ \sum_{i=1}^{k} w_i f_{\delta_i} \ ,$$

it holds that $f(u_i) = y_i$ for each $i = 1, \ldots, k$.

Proof. Let S be as given, and denote, for each sequence $\delta_1, \ldots, \delta_l$ of elements of Δ:

$$\Phi(\delta_1, \ldots, \delta_l) \ := \ \begin{pmatrix} f_{\delta_1}(u_1) & \cdots & f_{\delta_l}(u_1) \\ \vdots & \ddots & \vdots \\ f_{\delta_1}(u_k) & \cdots & f_{\delta_l}(u_k) \end{pmatrix} \ \in \ \mathbb{R}^{k \times l} \ .$$

Pick an $\varepsilon > 0$ such that

$$C \in \mathbb{R}^{k \times k} \, , \|C - I\| < \varepsilon \ \ \Rightarrow \ \ C \text{ nonsingular } , \tag{10}$$

where we denote $\|A\| := \sum |a_{ij}|$ for any matrix A. Now consider the approximate interpolation problem

$$f(u_1) = 1 \, , \ f(u_j) = 0 \, , \ j \neq 1 \, .$$

Let

$$f \ = \ \sum_{i=1}^{l} w_i f_{\delta_i}$$

solve this problem with tolerance ε. Thus

$$\Phi(\delta_1, \ldots, \delta_l) \, W_1$$

is at distance less than ε from the first column $\text{col}\,(1, 0, \ldots, 0)$ of an identity matrix, where $W_1 = \text{col}\,(w_1, \ldots, w_l)$. Repeating with each interpolation problem

$$f(u_i) = 1 \, , \ f(u_j) = 0 \, , \ j \neq i$$

and padding with zeroes as necessary the corresponding vectors $W_1, W_2, \ldots$, there results the existence of a set of indices $\{\delta_1, \ldots, \delta_q\}$ and a matrix $W \in \mathbb{R}^{q \times k}$, for some integer q, so that $C = \Phi(\delta_1, \ldots, \delta_q) \, W$ satisfies (10). We conclude that the matrix $\Phi(\delta_1, \ldots, \delta_q)$ has rank k. Picking a subset of k linearly independent columns, after reordering we may assume that

$$V \ := \ \Phi(\delta_1, \ldots, \delta_k)$$

is nonsingular. Now the equalities $f(u_i) = y_i$ can be achieved by simply solving for w the linear equation $Vw = \text{col}\,(y_1, \ldots y_k)$. ∎

Remark 9.2 In the above proof, note that if $\delta_1^0, \ldots, \delta_r^0$ are such that the corresponding matrix $\Phi(\delta_1^0, \ldots, \delta_r^0)$ is already known to have rank r then one may always take the first r columns of V to correspond to these indexes $\delta_i^0, i = 1, \ldots, r$ (just add these columns to $\Phi(\delta_1), \ldots, \Phi(\delta_q)$, and then pick a basis that includes them). In particular, we may apply the Lemma to the case of $(k-1, \theta)$-nets (use the constant function $f \equiv 1$ as the first element), and this will give equation (9) assuming that approximate interpolation employing *any* number of processors is possible.□

Remark 9.3 For nets with direct connections allowed, and if S contains at least $N+1$ affinely-independent elements, the projections $f(u) = u_i$ can be used as initial basis elements in addition to the constant $f \equiv 1$. So only $k - N - 1$ processors are needed in that case. This shows that for sets in general position, $k + N + 1$ points can be exactly interpolated using (k, θ)-nets with direct connections. □

The only property needed for the above two remarks to apply is that of approximate interpolation at any set S, with no prior constraint on how many processors are used –see for instance [5] for approximate interpolation. This property holds in particular if θ is so that nets with processors of type θ are dense in the set of continuous functions on a real interval (with the uniform convergence topology). Examples of such θ's are $\mathcal{H}$ or ([7]) any continuous bounded nonconstant function.

9.2 Interpolating with $k/2$ processors

The main technical fact is as follows:

Proposition 9.4 Assume that θ is a continuous sigmoid. Given any $2n + 1$ (distinct) points $x_0, \ldots, x_{2n}$ in R, any $\varepsilon > 0$, and any sequence of real numbers $y_0, \ldots, y_{2n}$, there exists some $(n + 1, \theta)$-net f such that $|f(x_i) - y_i| < \varepsilon$ for each i.

Before proving this Proposition, we establish an easy technical result:

Lemma 9.5 Let θ be a continuous sigmoid. Assume given real numbers $p, q, \alpha, \beta, \varepsilon, \delta$ so that $\varepsilon > 0$, $\delta > 0$, and $\alpha < q < \beta$. Then, there exists some real numbers a, b, c, d so that, if $f(x) := d + a\theta(bx + c)$, then the following properties hold:

1. $f(p) = q$.
2. $|f(x) - \alpha| < \varepsilon$ for all $x \leq p - \delta$.
3. $|f(x) - \beta| < \varepsilon$ for all $x \geq p + \delta$.

Proof. We assume that $t_+ = 1$ and $t_- = -1$ in the definition of sigmoid (cf. Remark 3.3). Let $\rho > 0$ be smaller than $\beta - q$, $q - \alpha$, and ε. Consider the function

$$g(\xi) := \frac{\beta - \alpha}{2}\theta(\xi) + \frac{\beta + \alpha}{2} .$$

Note that $g(\xi)$ approaches α, β at $-\infty, +\infty$, so there is some $K > 0$ so that $|g(\xi) - \alpha| < \rho$ if $\xi \leq -K$ and $|g(\xi) - \beta| < \rho$ if $\xi \geq K$. Pick any $\gamma > 2K/\delta$ and define for this γ, $f_0(x) := g(\gamma x)$. Then,

$$|f_0(x) - \alpha| < \rho \text{ if } x \leq -\delta/2$$

and

$$|f_0(x) - \beta| < \rho \text{ if } x \geq \delta/2 .$$

As $f_0(\delta/2) > \beta - \rho > q$ and $f_0(-\delta/2) < \alpha + \rho < q$, by continuity of f_0 (here we use that θ is continuous) there must be some $u \in (-\delta/2, \delta/2)$ so that $f_0(u) = q$. Finally, we let

$$f(x) := f_0(x + u - p) .$$

Clearly this satisfies $f(p) = q$. For any $x \leq p - \delta$ it holds that $z := x + u - p \leq -\delta/2$, so $|f(x) - \alpha| = |f_0(z) - \alpha| < \rho < \varepsilon$, as desired. The property for $x \geq \delta/2$ is shown analogously. ∎

Now we prove Proposition 9.4. Assume that we have already proved that for any two *increasing* sequences of real numbers

$$x_0 < x_1 < \ldots < x_{2n} \;\text{ and }\; z_0 < z_1 < \ldots < z_{2n} \tag{11}$$

there is some (n, θ)-net so that

$$|f(x_i) - z_i| < \varepsilon/2 \tag{12}$$

for each i. The result then follows from here. Indeed, given the original data, we may assume that the x_i are already in increasing order (reorder them, if necessary). Now pick any real d so that

$$d \;>\; \frac{y_i - y_{i+1}}{x_{i+1} - x_i} \tag{13}$$

for all $i = 0, \ldots, 2n-1$. Letting $z_i := x_i d + y_i$, these are now in increasing order. Let f be so that equation (12) holds for each i. By Lemma 5.3, there are some numbers a, b, c, e so that

$$|a + e\theta(bx_i + c) + dx_i| \;<\; \varepsilon/2$$

for each i. Then $|f(x_i) + a + e\theta(bx_i + c) - y_i| < \varepsilon$ is a $(n+1, \theta)$-net as wanted.

Thus, we must prove the result for the particular case of increasing sequences (11), which we do via an argument somewhat analogous to that used in [5] for showing the (weaker) fact that one can approximately interpolate n points using $n-1$ processors. We show inductively:

Given data (11) and any $\varepsilon > 0$, there exists an (n, θ)-net f so that

$$|f(x_i) - z_i| < \varepsilon \;\text{ for each }\; i = 0, \ldots, 2n \tag{14}$$

and

$$|f(x) - z_{2n}| < \varepsilon \;\text{ for all }\; x \geq x_{2n}\,. \tag{15}$$

For $n = 1$ this follows from Lemma 9.5, by choosing $p = x_1$, $q = z_1$, $\alpha = z_0$, $\beta = z_2$, and δ less than $x_1 - x_0$ and $x_2 - x_1$. Assume now that an $(n-1, \theta)$-net f_1 has been obtained for $x_0, \ldots, x_{2n-2}$ and $z_0, \ldots, z_{2n-2}$, and so that

$$|f_1(x_i) - z_i| < \varepsilon/2 \;\text{ for each }\; i = 0, \ldots, 2n-2 \tag{16}$$

and

$$|f_1(x) - z_{2n-2}| < \varepsilon/2 \;\text{ for all }\; x \geq x_{2n-2}\,. \tag{17}$$

Note that this last inequality holds in particular for x_{2n-1} as well as for all $x \geq x_{2n}$. Now let f_2 be as in Lemma 9.5, with δ less than $x_{2n-1} - x_{2n-2}$ and $x_{2n} - x_{2n-1}$, $\alpha = 0$, $\beta = z_{2n} - z_{2n-2}$, $q = z_{2n-1} - z_{2n-2}$, and $p = x_{2n-1}$, and so that

$$|f_2(x)| < \varepsilon/2 \;\text{ for all } x < x_{2n-1} - \delta \tag{18}$$

and

$$|f_2(x) - \beta| < \varepsilon/2 \;\text{ for all } x > x_{2n-1} + \delta\,. \tag{19}$$

It follows that $f := f_1 + f_2$ is as desired for the inductive step. This completes the proof of the Proposition. ∎

We now summarize properties of $\underline{\lambda}$. The next result should be compared with Theorem 1. The main difference is in the second equality. Note that one can prove $\underline{\lambda}(k,\theta) \geq \underline{\lambda}^d(k-1,\mathcal{H})$, in complete analogy with the case of $\underline{\mu}$, also as a consequence of Lemma 5.4, but this is not sufficient anymore to be able to derive the last inequality in the Theorem from the second equality.

Theorem 4 *For any continuous sigmoid θ, and for each k,*

$$\begin{aligned} \underline{\lambda}(k,\mathcal{H}) &= k+1 \\ \underline{\lambda}^d(k,\mathcal{H}) &= k+2 \\ \underline{\lambda}(k,\theta) &\geq 2k-1 \;. \end{aligned}$$

Proof. The first equality is easy, and the last one follows from Proposition 9.4. The inequality $\underline{\lambda}^d(k,\mathcal{H}) \geq k+2$ follows from Remark 9.3, as two distinct points are always affinely independent. We now prove the remaining inequality $\underline{\lambda}^d(k,\mathcal{H}) \leq k+2$.

Consider the problem of interpolating at the points

$$\{1,2,\ldots,k+3\}$$

and the respective desired values

$$\{0,1,0,2,0,3,\ldots\}$$

(that is, odds should be mapped to zero, and even numbers of the form $2l$ into l). Assume that

$$f(u) \;=\; w_0 \;+\; v_0 u \;+\; \sum_{i=1}^{k} w_i\,\mathcal{H}(v_i u - \tau_i)$$

would solve the approximate interpolation problem for this data, with, say, $\varepsilon = 0.2$. Without loss of generality, we may take all $v_i = 1$.

Consider the possible points of discontinuity $x = \tau_i$. As $\mathcal{H}$, and therefore also f, is continuous from the left, we may shift f into a map of the form $f(u-\delta)$, $\delta > 0$, with δ small enough so that the new f still interpolates to within ε accuracy, and $\tau_i + \delta$ is not an integer for any i. So we will assume from now on that τ_i is not an integer.

Since there are only k points of the form τ_i, there must be two integer intervals, say $[l,l+1]$ and $[m,m+1]$, with l and m in the range $\{1,2,\ldots,k+2\}$, that contain no such point. In each, f is an affine map

$$f(u) \;=\; v_0 u + \alpha_i$$

(where α_i depends on the interval). The slope v_0 of f on each interval is the same, but as calculated at the endpoints the slopes must be different (since they must be at distance less than one from distinct integers, by the choice of interpolation data). Thus no such net can exist. ∎

Remark 9.6 One may expect that, under weak extra hypotheses, it should be possible to interpolate at $2k$, rather than $2k-1$, points. Note that for $k=2$ this is easy to achieve: just choose the slope d so that some $z_i - z_{i+1}$ becomes zero and the z_i are allowed to be nonincreasing or nondecreasing. The same proof, changing the signs if necessary, gives the wanted net. Below we prove that when $\theta = \pi$, the piecewise linear sigmoid, this bound is always achieved. □

9.3 Interpolation for Some Particular Activations

The previous results show that one can approximately interpolate at any $2k-1$ points using only k sigmoidal processors. We now prove that, for the *standard sigmoid*, this approximate interpolation property holds in the following stronger sense: for an open dense set of $2k-1$ points, one can achieve an open dense set of values. (In this sense, for "generic" data at $2k-1$ points the interpolation problem can be solved.)

Remark 9.7 Note that in general approximate interpolation fails to imply the stronger property. To illustrate the difference, start with any smooth map $\rho : \mathbb{R} \to \mathbb{R}^2$ which has a dense image D. Next let

$$f(x,\delta) := \rho_1(\delta) + x\rho_2(\delta) ,$$

seen as a function of x parameterized by δ. Given any two distinct points $x_1, x_2 \in \mathbb{R}$, the possible pairs $(f(x_1,\delta), f(x_2,\delta))$, as δ varies, describe a dense subset of $\mathbb{R}^2$. This is because

$$\begin{pmatrix} f(x_1,\delta) \\ f(x_2,\delta) \end{pmatrix} = T_{x_1,x_2} \begin{pmatrix} \rho_1(\delta) \\ \rho_2(\delta) \end{pmatrix} ,$$

where

$$T = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}$$

is nonsingular. By Sard's Theorem, however, this set of pairs $(f(x_1,\delta), f(x_2,\delta))$ has measure zero, as the differential of the map $\delta \mapsto f(x_1,\delta), f(x_2,\delta)$ has everywhere rank ≤ 1. Thus, the family of functions $\mathcal{F} = \{f(\cdot,\delta), \delta \in \mathbb{R}\}$ is so that for *every* pair $x_1 \neq x_2$ one can solve the approximate interpolation problems $f(x_1,\delta) = y_1$, $f(x_2,\delta) = y_2$ to any desired accuracy, but on the other hand for *no* possible pair $\{x_1, x_2\}$ does the set of achievable pairs (y_1, y_2) have nonzero measure.

Another example of the same phenomenon, this one closer to nets, uses the sigmoid θ introduced in Equation (5). We claim that for every integer k, any k rationally independent numbers $\{x_1, \ldots, x_k\}$, any k real numbers $\{y_1, \ldots, y_k\}$, and any $\varepsilon > 0$, there is some $(2,\theta)$-net f which satisfies (7). Indeed, consider the function ρ introduced in (6). Take a sequence of real numbers $\omega_n \to \infty$ so that, for each $j = 1, \ldots, k$,

$$\cos \omega_n x_j \;\to\; q\left(\frac{x_j}{x_1}\right)^2 y_j \quad \text{as} \quad n \to \infty ,$$

where q is any number so that $q(x_j/x_1)^2 y_j \in [-1,1]$ for all j. Now let

$$f_n(x) := \frac{\alpha(1 + w_n^2 x_1^2)}{2q}\, \rho(w_n x) ,$$

which is computable by a $(2,\theta)$-net for each fixed n. It follows that $f_n(x_j) \to y_j$ as $n \to \infty$, as desired. Again we conclude that approximate interpolation is possible (except, in this case, for those $\{x_1, \ldots, x_k\}$ that are rationally dependent), but for each given set $\{x_1, \ldots, x_k\}$ the set of possible values $(f(x_1), \ldots, f(x_k))$, as f ranges over all possible $(2,\theta)$-nets, has measure zero if $k > 7$ (the number of parameters).

The content of the next result is that using *special properties* of the standard sigmoid one can obtain the stronger generic interpolation result. □

The result for the standard sigmoid will be a consequence of the following more general technical fact. We write $\operatorname{im} T$ for the image of a mapping T and $\operatorname{int} S$ for the interior of a set S.

Lemma 9.8 Let $\Phi : \mathbb{R}^r \times \mathbb{R}^s \to \mathbb{R}^t$ be a real-analytic mapping. Assume that there is some open dense subset $X_0 \subseteq \mathbb{R}^r$ so that

$$\text{for each fixed } x \in X_0\,,\ \ \operatorname{im} \Phi(x,\cdot) \text{ is dense in } \mathbb{R}^t \tag{20}$$

and that there exists some $x_0 \in X_0$ such that one can factor $\Phi(x_0,\cdot)$ in the following manner:

$$\Phi(x_0,\alpha) \ = \ R(\phi(\alpha)) \quad \text{for all } \alpha \in \mathbb{R}^s \tag{21}$$

where $\phi : \mathbb{R}^s \to \mathbb{R}^q$ is a mapping so that $\operatorname{int} \operatorname{im} \phi \neq \emptyset$, for some positive integer q, and $R : \mathbb{R}^q \to \mathbb{R}^t$ is a rational function having no poles on the image of ϕ.

Then, there exists an open dense subset $X_1 \subseteq \mathbb{R}^r$ so that

$$\text{for each fixed } x \in X_1\,,\ \ \operatorname{im} \Phi(x,\cdot) \text{ contains an open dense subset of } \mathbb{R}^t\,. \tag{22}$$

In particular, this implies that $s \geq t$.

Proof. Consider the map R. As it is rational, its image is a semialgebraic set, that is, a finite union of sets of the form $A_i \bigcap \Phi_i$, where each A_i has the form $\{x \,|\, f(x) > 0\}$ and each Φ_i has the form $\{x \,|\, f(x) = 0\}$, for suitable polynomials f. (This is the Tarski-Seidenberg, or "generalized Sturm's" Theorem; see for instance [8], VI.10.) Since $\Phi(x_0,\cdot) = R \circ \phi$, the image of R must also be dense, so no proper sets of type Φ_i may appear. Thus $\operatorname{im} R$ contains an open set, from which it follows from Sard's Theorem that R must have a nonsingular Jacobian at some point, and hence at *almost all* points in its domain (by analyticity). In particular, R must have a nonsingular Jacobian at some point of $\operatorname{int} \operatorname{im} \phi$, so by the Implicit Mapping Theorem it follows that $\operatorname{int} \operatorname{im} R \circ \phi$ is nonempty too. Again from Sard's Theorem, this time applied to $\Phi(x_0,\cdot)$, this means that $\frac{\partial \Phi}{\partial \alpha}(x_0,\alpha_0)$ has rank t for some $\alpha_0 \in \mathbb{R}^s$ (and in particular $s \geq t$).

Now let

$$X_1 := \left\{ x \in X_0 \,\middle|\, \operatorname{rank} \frac{\partial \Phi}{\partial \alpha}(x,\alpha_0) = t \right\}$$

which is again open dense, because of analyticity of $\frac{\partial \Phi}{\partial \alpha}(\cdot,\alpha_0)$. For each fixed $x \in X_1$ the set

$$\mathcal{A}_x \ := \ \left\{ \alpha \,\middle|\, \operatorname{rank} \frac{\partial \Phi}{\partial \alpha}(x,\alpha) = t \right\}$$

is open dense, by analyticity. Let Φ_x be the restriction of $\Phi(x,\cdot)$ to $\mathcal{A}_x$. The image of $\Phi(x,\cdot)$ is dense (because $x \in X_1 \subseteq X_0$), so density of $\mathcal{A}_x$ in $\mathbb{R}^s$ implies that $\Phi_x(\mathcal{A}_x)$ is also dense. Moreover, Φ_x is an open mapping, since it has nonsingular differential at every point, so the image of Φ_x is an open dense subset of $\mathbb{R}^t$. ∎

Remark 9.9 The above Lemma can be applied to the interpolation problem with the standard sigmoid (3). Just take

$$\Phi((u_1,\ldots,u_{2k-1}),\alpha) \ := \ \begin{pmatrix} f(u_1,\alpha) \\ \vdots \\ f(u_{2k-1},\alpha) \end{pmatrix} \tag{23}$$

where $f(u, \alpha)$ is the map (2) (with no connections, i.e. $v_0 = 0$,) and α is the vector consisting of all the weights w_0 as well as w_i, v_i, τ_i, $i = 1, \ldots, k$ which appear in (2). (Thus $r = (2k-1)N$, $s = 1 + (2+N)k$, and $t = 2k - 1$ in the Lemma.) Property (20) holds by Proposition 9.4, with $X_0 = \mathbb{R}^r$. Property (21) holds if we take as x_0 any vector $(u_1^0, \ldots, u_{2k-1}^0)$ where the u_i^0's are $2k-1$ distinct vectors with integer coordinates. When x_0 is like this, $\Phi(x_0, \cdot)$ can be expressed as a rational function of the w_i's and of the exponentials of both the scalars τ_i's and the coordinates of the vectors v_i's. Thus one can take $q = s$, and the map ϕ in the Lemma is obtained by taking either an identity or an exponential in each coordinate (and is hence a diffeomorphism with its image, which implies that the image has nonempty interior, as needed for the Lemma). □

Another interesting example is that of the piecewise linear sigmoid π introduced in Equation (4). We next show that

$$\underline{\lambda}(k, \pi) \ = \ 2k \tag{24}$$

(not just $2k-1$). The upper bound is a consequence of Lemma 8.1 and Equation (8). To prove that $2k$ points can be interpolated, in fact *exactly*, not just approximately, we modify the proof of Proposition 9.4 as follows.

Assume given an increasing sequence of $2n+2$ points $x_{-1} < x_0 < x_1 < \ldots < x_{2n}$ in $\mathbb{R}$ as well as $2n+2$ desired interpolation values $y_{-1}, y_0, \ldots, y_{2n}$. We will show the existence of an $(n+1, \pi)$-net so that $f(x_i) = y_i$ for $i = -1, \ldots, 2n$.

Without loss of generality, we may assume that $y_{-1} \geq y_0$ (multiply everything by -1 otherwise) and that $x_0 = 0$ (translate if necessary). As before, pick a d so that (13) holds but now ask in addition that $d > \frac{y_0 - y_{-1}}{x_{-1}} \geq 0$, so that not only are the $z_i = x_i d + y_i$ increasing for $i = 0, \ldots, 2n$, but also

$$x_{-1} \ < \ \frac{y_0 - y_{-1}}{d} \ .$$

Now find an (n, π)-net which interpolates $f(x_i) = z_i$ for $i \geq 0$ and also satisfies $f(x) = z_0 = y_0$ for all $x \leq x_0 = 0$ (same proof as before works, except that now one can easily show that the interpolation is exact).

The final interpolation function will have the form $f(x) + q(x)$, where $q(x)$ is a $(1, \pi)$-net. Thus we need

$$q(x_i) \ = \ -x_i d, \ i = 0, 1, \ldots, 2n$$

and $q(x_{-1}) = y_{-1} - y_0$ (this last equality because $f(x_{-1}) = y_0$). This can be achieved by any $(1, \pi)$-net which has the constant value $y_{-1} - y_0$ for all $x \leq \frac{y_0 - y_{-1}}{d}$ and which coincides with the linear map $q(x) = -xd$ for $x \in [\frac{y_0 - y_{-1}}{d}, x_{2n}]$. This completes the proof of (24).

9.4 Particular Sets of Points

In the case of classification, whereas not *every* set of cardinality $k+2$ can be shattered by $(k, \mathcal{H})$-nets, (or $2k+3$ if allowing direct connections,) it is true that *some* sets of cardinality $2k+1$ (or $4k+3$ with direct connections) can be shattered in $\mathbb{R}^2$. It is then natural to ask if a similar situation occurs for interpolation, that is, if by choosing appropriate points in $\mathbb{R}^N$, $N > 1$, one may be able to achieve interpolation at more than $k+1$ points (or $k+2$, if direct connections are allowed). We next show that, at least for the case of Heaviside nonlinearities, such an improvement is essentially impossible.

Fix any set of p points $u_1, \ldots, u_p$ in $\mathrm{I\!R}^N$, and consider the mapping

$$\Phi(\alpha) \; = \; \begin{pmatrix} f(u_1, \alpha) \\ \vdots \\ f(u_p, \alpha) \end{pmatrix}$$

as earlier, where

$$\alpha \; = \; (\vec{w}, \vec{v}, \vec{\tau}) \; \in \mathrm{I\!R}^{(k+1)+Nk+k}$$

is the set of all weights appearing in a $(k, \mathcal{H})$-net with no direct connections. (In the case where direct connections are allowed, the notation will be the same, but now α will be a vector of size $N(k+1)+k$.) For each binary matrix $E = (e_{ij}) \in \{0,1\}^{p \times k}$ let

$$X_E \; := \; \{\alpha \,|\, \mathcal{H}(v_j \cdot u_i - \tau_j) = e_{ij} \text{ for each } i = 1, \ldots, p \text{ and } j = 1, \ldots, k\} \; .$$

Then, the image of Φ is the (*finite*) union of the images of the restrictions to each of the (possibly empty) sets X_E. On the other hand, each such restriction is the map

$$(\vec{w}, \vec{v}, \vec{\tau}) \; \mapsto \; \widehat{E}\vec{w}$$

where

$$\widehat{E} \; = \; \begin{pmatrix} 1 & & \\ \vdots & E & \\ 1 & & \end{pmatrix} \quad \in \{0,1\}^{p \times (k+1)}$$

and hence is a subspace of dimension at most $k+1$. We conclude that the image of Φ is a finite union of subspaces of $\mathrm{I\!R}^p$ of dimension no greater than $k+1$. Thus a dense set of values can only be obtained if $p \; \leq \; k+1$, which is the result that we had for arbitrary points. Thus, with the obvious definition,

$$\overline{\lambda}(k, \mathcal{H}, N) = \underline{\lambda}(k, \mathcal{H}, N) = k+1 \; .$$

In the case of direct connections, the restrictions to each set X_E are of the form ($\widehat{E}\vec{w}+$ linear map on $\mathrm{I\!R}^N$) and thus the image is a subspace of dimension $\leq k+1+N$. Therefore, no matter what the set of points is, one cannot obtain a dense set of values unless $p \leq k+1+N$. Together with Remark 9.3, one concludes that (again with the obvious notation)

$$\lambda^d(k, \mathcal{H}, N) \; = \; \overline{\lambda}^d(k, \mathcal{H}, N) \; = \; k+1+N \; .$$

Though for $N > 1$ this is larger than $\underline{\lambda}^d(k, \mathcal{H}, N) = k+2$, we only gained a constant (independent of the number of processors k) improvement.

10 Conclusions and Remarks

Our main conclusions can be summarized as follows. For any θ, let

$$\underline{\mu}(\theta) \; := \; \liminf_{k \to \infty} \frac{\underline{\mu}(k, \theta)}{k} \; .$$

Thus, roughly speaking, we are guaranteed that $k\underline{\mu}(\theta)$ points can always be shattered when using (k, θ)-nets. Similarly we may define $\underline{\mu}^d$ for the case of direct connections, and we may analogously define $\underline{\lambda}$ for interpolation problems.

With these notations, we proved that

$$\underline{\mu}(\mathcal{H}) = 1\,, \quad \underline{\mu}^d(\mathcal{H}) = 2\,, \quad \underline{\mu}(\theta) \geq 2\,.$$

where the last inequality holds for any sigmoid θ that satisfies (S1)-(S2). (In particular, $\underline{\mu}(\pi) = 2$, but it can even happen that $\underline{\mu}(\theta) = \infty$ for suitable θ's.) In contrast, for interpolation we had:

$$\underline{\lambda}(\mathcal{H}) = 1\,, \quad \underline{\lambda}^d(\mathcal{H}) = 1\,, \quad \underline{\lambda}(\theta) \geq 2\,,$$

where the last inequality holds for any *continuous* sigmoid θ that satisfies (S1)-(S2). (In particular, $\underline{\lambda}(\pi) = 2$; and also $\underline{\lambda}(\theta) \geq 1$ even if continuity or (S1)-(S2) do not hold, but under very weak nonlinearity assumptions.) These results hold independently of the input dimension N. Note that a parameter count would suggest $\underline{\lambda} \leq 3$, and indeed such a bound holds in the case of the standard sigmoid, as results from the facts on "generic" interpolation discussed in Remark 9.9.

Various upper bounds were given for μ. For the VC dimension $\overline{\mu}$, the results given were for the case $N = 2$: Letting

$$\overline{\mu}(\theta) \;:=\; \liminf_{k\to\infty} \frac{\overline{\mu}(k,\theta,2)}{k}$$

and similarly for $\overline{\mu}^d$, we have:

$$\overline{\mu}(\mathcal{H}) \geq 2\,, \quad \overline{\mu}^d(\mathcal{H}) \geq 4\,, \quad \overline{\mu}(\theta) \geq 4$$

(the latter if (S1)-(S2) hold).

It is known from [3] that $\overline{\mu}(\mathcal{H}) \geq N$ for any input dimension N. We conjecture that if (S1)-(S2) hold then

$$\overline{\mu}(\theta) \geq 2N$$

for all N, not just $N = 2$, and similarly that $\overline{\mu}^d(\mathcal{H}) \geq 2N$ in general. Unfortunately, our proofs of these facts for $N = 2$ are based on the result from [2] regarding arrangements of points in the plane, a fact which does not generalize to dimension three or higher (S. Suri has shown —personal communication— that for any k-element set in $\mathbb{R}^3$ there are dichotomies for which no family of less than $\frac{k}{2+\varepsilon}$ parallel hyperplanes can partition the space into single-class regions, where ε is some number smaller than 0.2).

For a measure of interpolation similar to $\overline{\mu}$, we showed that still $\overline{\lambda}(\mathcal{H}) = 1$ and $\overline{\lambda}^d(\mathcal{H}) = 1$ hold, independently of the dimension N. For sigmoids, however, it may happen that $\overline{\lambda}(\theta) = \infty$ (c.f. Remark 9.7).

One may also compare the power of nets with and without connections, or threshold vs sigmoidal processors, on Boolean problems. For instance, it is a trivial consequence from the given results that parity on n bits can be computed with $\lceil \frac{n+1}{2} \rceil$ hidden sigmoidal units and no direct connections, though requiring (apparently, though this is an open problem) n thresholds. In addition, for some families of Boolean functions, the gap between sigmoidal nets and threshold nets may be infinitely large: in [9] the authors prove in particular that the class of functions of $2n$ Boolean variables

$$F_n(x_1,\ldots,x_n,y_1,\ldots,y_n) \;:=\; \mathrm{MAJ}\,(x_1,\ldots,x_n) \oplus \mathrm{MAJ}\,(y_1,\ldots,y_n)$$

(where MAJ indicates majority function) can be computed with $(5,\theta)$-nets, independently of n, as long as θ be "nonlinear" enough (for instance, if θ is twice differentiable and satisfies (S1),) but there is no possible fixed integer l so that every F_n can be computed by some $(l,\mathcal{H})$-net.

In the recent work [14], the author has studied certain representation properties of *two*-hidden-layer nets, in comparison to the single-layer nets studied here (see also [4]). The results in that reference do not deal with numbers of units needed, however.

References

[1] Arai, M., “Mapping abilities of three-layer neural networks,” *Proc. Int. Joint Conf. on Neural Networks*, Washington, DC, June 18-22, 1989, IEEE Publications, 1989, pp. I-419/424.

[2] Asano,T., J. Hershberger, J. Pach, E.D. Sontag, D. Souivaine, and S. Suri, “Separating Bi-Chromatic Points by Parallel Lines,” *Proceedings of the Second Canadian Conference on Computational Geometry*, Ottawa, Canada, 1990, p. 46-49.

[3] Baum, E.B., “On the capabilities of multilayer perceptrons,” *J.Complexity* **4**(1988): 193-215.

[4] Blum, E.K., and L. Kwan Li, “Approximation theory and feed-forward networks,” to appear.

[5] Chester, D., “Why two hidden layers and better than one,” *Proc. Int. Joint Conf. on Neural Networks*, Washington, DC, Jan. 1990, IEEE Publications, 1990, p. I.265-268.

[6] Cybenko, G., “Approximation by superpositions of a sigmoidal function,” *Math. Control, Signals, and Systems* **2**(1989): 303-314.

[7] Hornik, K., “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, **4**(1991): 251-257.

[8] Jacobson, N., *Lectures in Abstract Algebra*, Van Nostrand, Princeton, 1964.

[9] Maass, M., G. Schnitger, and E.D. Sontag, “On the computational power of sigmoid versus Boolean threshold circuits,” *Proc. of the 32nd Annual IEEE Conference on Foundations of Computer Science*, San Juan, Puerto Rico, Oct. 1991.

[10] Mañé, R., *Ergodic Theory and Differentiable Dynamics,* Springer-Verlag, NY, 1987.

[11] Shrivastava, Y., and S. Dasgupta, “Neural networks for exact matching of funtions on a discrete domain,” *Proc. IEEE Conf. Dec. and Control*, Honolulu, Dec. 1990, pp. 1719-1724.

[12] Rudin, W., *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

[13] Sontag, E.D., “Remarks on interpolation and recognition using neural nets,” in *Advances in Neural Information Processing Systems 3* (R.P. Lippmann, J. Moody, and D.S. Touretzky, eds), Morgan Kaufmann, San Mateo, CA, 1991, pp. 939-945.

[14] Sontag, E.D., “Feedback Stabilization Using Two-Hidden-Layer Nets,” in *Proc. Amer. Automatic Control Conference*, Boston, June 1991, pp. 815-820.

[15] Sontag, E.D., and H.J. Sussmann, “Backpropagation can give rise to spurious local minima even for networks without hidden layers,” *Complex Systems* **3**(1989): 91-106.

[16] Sontag, E.D., and H.J. Sussmann, “Backpropagation separates where perceptrons do,” *Neural Networks*, **4**(1991): 243-249.

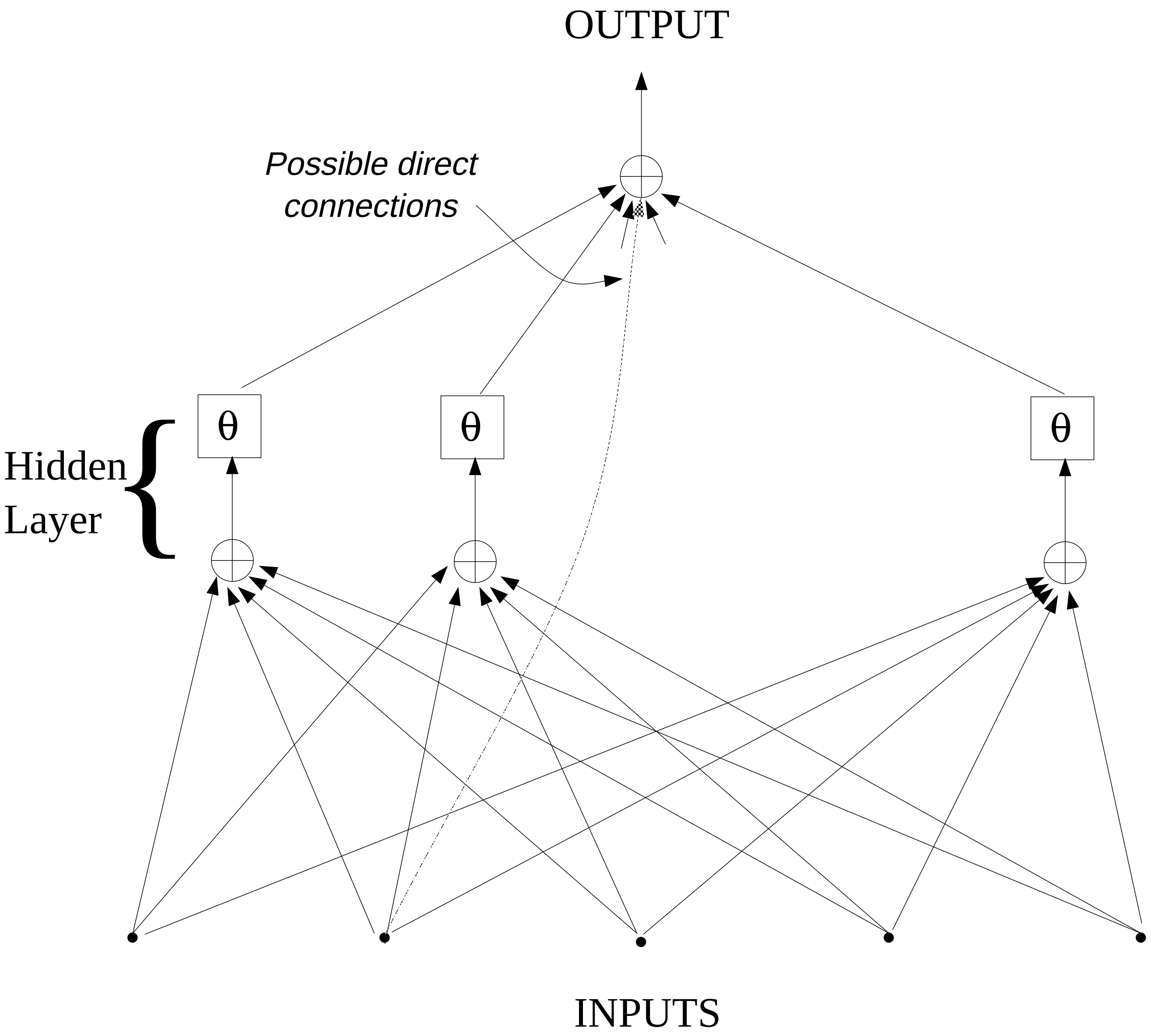

Figure 1: *Feedforward Net.*

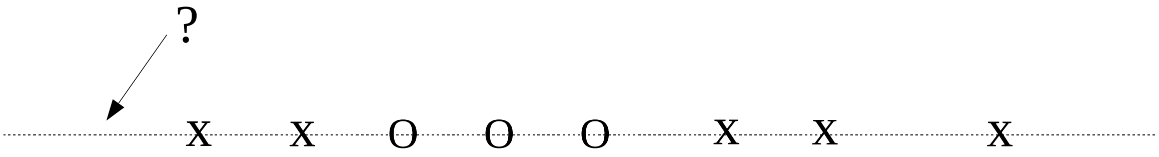

Figure 2: *Labeled Samples.*

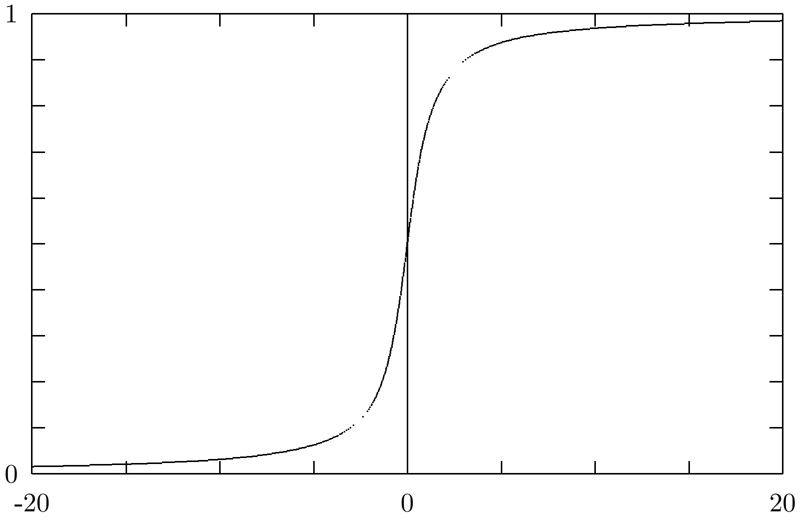

Figure 3: *Sigmoid leading to infinite VC dimension.*