10 References

1. Arai, M., "Mapping abilities of three-layer neural networks," *Proc. IJCNN Int.Joint Conf.on Neural Networks, Washington, 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, "On a geometric problem related to neural net learning," in preparation.

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

4. Bultman, W., "On the design of threshold circuits for some basic functions," Tech. Report from Dept. of Math. and CS, University of Illinois at Chicago, March 1990.

5. Cybenko, G., "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, and Systems* **2**(1989): 303-314.

6. Hanson, S., personal communication.

7. Hornik, K.M., M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, 1989: to appear.

8. O'Neil, P.E., "Hyperplane cuts of an n-cube," *Discrete Math.* **1**(1971): 193-195.

9. Rudin, W., *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

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.

To conclude this note, we wish to remark that there are "universal" functions θ satisfying (S1)-(S2) and as differentiable as wanted, even analytic, such that, for each N and each dichotomy on any finite set $S \subseteq \mathbb{R}^N$, this dichotomy can be implemented by a $(1, \theta)$-net. In other words, $\underline{\mu}(1, \theta, N) = \infty$. Of course, the function θ is so complicated that there is no reasonably "finite" implementation for it. This remark is only of theoretical interest, to indicate that, unless further restrictions are made on (S1)-(S2), much better bounds can be obtained. We now proceed with the construction of θ. By an argument as before, it is enough to consider the case $N = 1$.

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 $(1 - 1/y_j^i)$, multiplied by suitable exponential functions to guarantee convergence, see [9], 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 6.1. 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 $\tau := -\ln(\rho)$.

then, since θ' is bijective on each of $x < 0$ and $x > 0$, it would follow that $x - 1 = -x - 1$ or $x - 1 = x + 1$, a contradiction. Since $\gamma'(1) = \theta'(0) - \theta'(-2) > 0$, γ' is positive for $x > 0$, proving (3). Also, for all $x, y \in \mathbb{R}$,

$$\gamma(x) < \gamma(y) \text{ if and only if } |x| < |y| \,. \tag{4}$$

This is because the result is true for positive x, y, by (3), and in general $\gamma(|x|) = \gamma(x)$ and $\gamma(|y|) = \gamma(y)$ since γ is even by definition.

From (4) we conclude that, for all $x, y \in \mathbb{R}$,

$$\gamma(x - y) - \gamma(x + y) < 0$$

if and only if x and y have equal signs. Given a Boolean vector $(x_1, \ldots, x_n)$, with odd n, the sign of $x_1 + \ldots + x_n - n/2$ is positive if and only if $\mathrm{MAJ}\,(x_1, \ldots, x_n) = 1$. We conclude that

$$\begin{aligned} f_n(x_1, \ldots, x_n, y_1, \ldots, y_n) \;&:=\; \gamma(x_1 + \ldots + x_n - y_1 - \ldots - y_n) \\ &-\; \gamma(x_1 + \ldots + x_n + y_1 + \ldots + y_n - n) \end{aligned}$$

implements ϕ_n. In terms of the original θ, we can write

$$\begin{aligned} f_n(x_1, \ldots, x_n, y_1, \ldots, y_n) \;&=\; \theta(x_1 + \ldots + x_n - y_1 - \ldots - y_n - 1) \\ &+\; \theta(-x_1 \ldots - x_n + y_1 + \ldots + y_n - 1) \\ &-\; \theta(x_1 + \ldots + x_n + y_1 + \ldots + y_n - n - 1) \\ &-\; \theta(-x_1 \ldots - x_n - y_1 - \ldots - y_n + n - 1) \end{aligned}$$

which is the desired $(4, \theta)$-net.

9 Remarks

For arbitrary functions θ satisfying (S1) and (S2), the number $\underline{\mu}$ can be improved considerably from $2k$, (in fact reduced to 1, see below), but this improvement is useless and only of purely academic interest. On the other hand, we conjecture that for the important practical case $\theta(x) = \sigma(x)$ it is close to optimal.

For the piecewise linear function π, at least, one has equality $\underline{\mu}(k, \sigma, N) = 2k$. To prove this last 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.

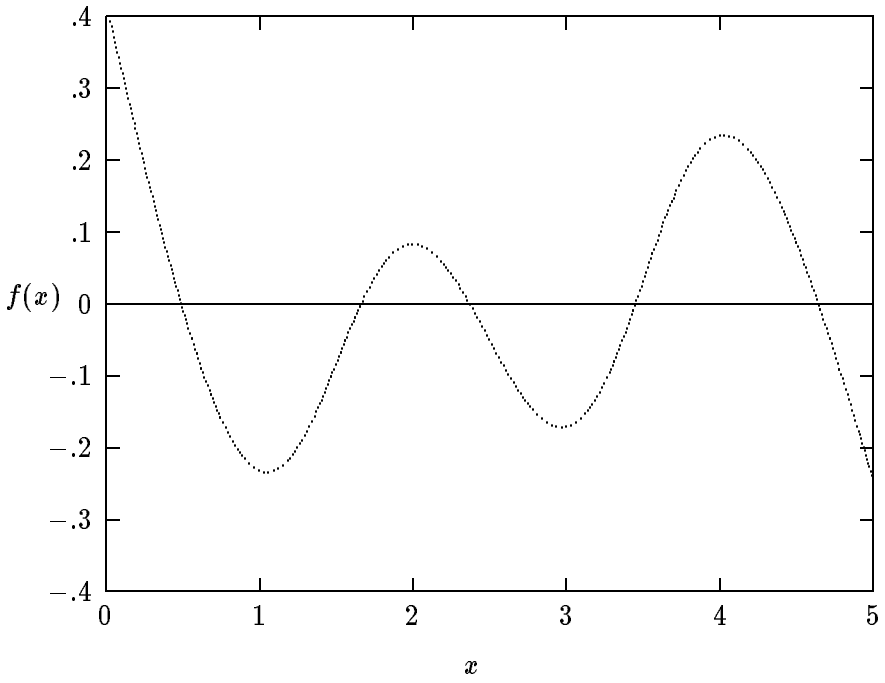

Figure 1: 5-bit parity with 3 sigmoids

be the "majority" function that equals "1" if the majority of the x_i's are one, and zero otherwise, and let

$$\mathrm{XOR}\,(x, y)$$

be the sum mod 2 of x and y. Then we let

$$\phi_n(x_1, \ldots, x_n, y_1, \ldots, y_n) := \mathrm{XOR}\,(\mathrm{MAJ}\,(x_1, \ldots, x_n), \mathrm{MAJ}\,(y_1, \ldots, y_n))$$

(a Boolean function of $2n$ variables).

These functions were studied by [4], who showed that ϕ_n is implementable using $(k, \mathcal{H})$-nets with k growing at most polynomially on n. It seems highly unlikely that a *fixed* k will work when using Heavisides, however. (If this conjecture is proved, one would have an example of a class of Boolean functions for which between the gap between sigmoids and Heavisides is infinite.)

The construction will work for any function θ that is differentiable and satisfies the following property (as the sigmoid does):

$$\theta' \text{ is even and it strictly decreases for } x > 0.$$

Let θ be like this, and introduce $\gamma(x) := \theta(x-1) + \theta(-x-1)$. We claim that

$$\gamma \text{ is strictly increasing for } x > 0\ . \tag{3}$$

Indeed, note first that γ' only vanishes at $x = 0$, because if for some nonzero x it holds that

$$\theta'(x-1) = \theta'(-x-1)\ \Big(= \theta'(x+1)\Big)$$

Lemma 7.3 If f is a $(k, \mathcal{H})$-net, S is a finite subset of $\mathbb{R}^N$, 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. Arguing as in Remark 3.1, we can assume that the expressions $v_i.s_j - \tau_i$ are all nonzero. Let K be the set consisting of all these expressions. For each term of the form

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

we can use an approximation

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

for large enough λ, by Lemma 7.1. For the linear term $v_0.u$, on the other hand,

$$v_0.u \;\approx\; -\frac{\lambda}{\mu}\theta(c) + c - \frac{\lambda}{\mu}\theta\left(\frac{1}{\lambda}v_0.u - \frac{c - c\lambda}{\lambda}\right)$$

for large λ, by Lemma 7.2). ∎

8 Some Boolean Problems

The results given in this paper are mostly about dichotomies of arbitrary sets in the plane, but contact with an active area of theoretical computer science, that of circuit complexity, results when specializing to functions defined on Boolean hypercubes $\{0,1\}^n$. Given a function $\phi : \{0,1\}^n \to \{0,1\}$, we say that a net f implements this function if f implements the dichotomy induced by ϕ, that is, $f(u) > 0$ (respectively, < 0,) if and only if $\phi(u) = 1$ (respectively, $= 0$,) for all Boolean vectors u.

The mapping

$$(x_1, \ldots, x_n) \mapsto x_1 + \ldots + x_n$$

from the Boolean n-cube into the interval $\{0, \ldots, n\}$ transforms the parity problem into the problem of dichotomizing the latter set into odds from evens, which can be done with an $(n, \mathcal{H})$-net with no connections. A smaller n may be sufficient for the Boolean parity problem, but it is unknown if this is the case (see for instance [8] for references to this problem, which is classical in combinatorics; in any case, it is easy to prove directly that n is best possible for $n = 2, 3$).

When using sigmoids, however, parity can be implemented using $\lceil \frac{n+1}{2} \rceil$ neurons, as follows by the same reduction and Theorem 1. (Recently, Steve Hanson reported to us [6] that such smaller nets had been observed in actual backprop experiments.) Because the result is somewhat unexpected, it is worth calculating explicitely one example. Applying the constructions in this note, the 5-bit parity problem can be solved by the following $(3, \mathcal{H})$ net with no connections. Let

$$f(x) := 10.4 + 2\sigma(3x - 4.5) + 2\sigma(3x - 10.5) - 20\sigma(x/5)\,.$$

Then, $f(x_1 + \ldots + x_5)$ computes the desired parity. (See Figure 1.)

We wish to construct now an example of a sequence ϕ_n of Boolean functions on n variables which can be implemented by nets with a fixed number (four) of sigmoidal neurons, that is, a $(4, \sigma)$-net. Let n be odd and let

$$\mathrm{MAJ}\,(x_1, \ldots, x_n)$$

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 to appear in [2], which we state next.

Result. Pick any integer $1 \leq k$. 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 $v.u$, $u \in S_-$ or only elements of the type $v.u$, $u \in S_+$. ∎

Take any k, and apply this result with $n := 4k+3$. There result $2k+2$ intervals. By Lemma 6.1, the intervals can be separated by some $(k, \mathcal{H})$-net (with direct connections), and again $g(u) := f(v.u)$ 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 neurons and no connections.

7 Approximating by Sigmoids

In order to prove Lemma 4.1 (for arbitrary "sigmoids" θ that satisfy the conditions given earlier), we need a few simple lemmas. The first two are basically just restatements of (S1') and (S2) respectively:

Lemma 7.1 For each compact subset $K \subseteq \mathbb{R}$ not containing zero,

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

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

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

uniformly on $x \in K$.

Proof. Multiplying by μ, we need to show that

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

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

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

which is precisely property (S2). ∎

Now lemma 4.1 follows from the following one:

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.

6 Lower Bounds

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

Proposition 6.1 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 have cardinality $l = 2k$. First the problem is reduced to the case $N = 1$, by finding a vector v whose inner products with all the elements of S are distinct, and replacing S by the set of these inner products (this is closely related to [1], which looked at somewhat similar questions). Precisely, if S consists of the distinct vectors $u_1, \ldots, u_l$, then there exists some vector v so that all the numbers $v.u_i$ are distinct. Indeed, the set of all such separating v's is the intersection of the finitely many open dense sets

$$\{v \mid v.(u_i - u_j) \neq 0\}$$

(one for each $i \neq j$) and is hence nonempty.

Pick one such v, and let $y_i := v.u_i$ for each i; without loss of generality assume that $y_1 < y_2 < \ldots < y_l$. Now take any dichotomy (S_-, S_+) of S. This induces a partition of the set of y_i's into two subsets, corresponding to values $v.u_i, u_i \in S_+$ and $v.u_i, u_i \in S_-$. We shall assume that y_1 is of the first of these types; otherwise the argument is the same (multiply the obtained f by -1). Thus there are disjoint closed finite intervals as in the statement of Lemma 6.1 such that each y_i of the first type is in some interval I_j and each y_i of the second type is in some interval J_j. Let f be as in the conclusion of the Lemma. Then $g(u) := f(v.u)$ is a $(k-1, \mathcal{H})$-net which implements the dichotomy. 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

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

5 Upper Bounds

Pick first $N = 1$, and let f be as in (2) with $\theta = \mathcal{H}$. The points u 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). So there is a set of cardinality $2k+2$ (or $k+1$ respectively) which cannot be shattered. A similar example holds in each space $\mathbb{R}^N$, simply by considering points of the form $(u, 0, \ldots, 0)$ and the above dichotomy on the u's. This shows the upper bounds $\underline{\mu}(k, \mathcal{H}, N) \leq k+1$ and $\underline{\mu}^d(k, \mathcal{H}, N) \leq 2k+2$.

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 $v_i.u = \tau_i$, $i = 1, \ldots, k$. (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. 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 dichotomiy 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

$$w_0 + v_0.u + \gamma = 0$$

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 $l \leq 3.2 + 2.(2k-2) = 4k+2$. 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

Lemma 4.1 For each k, N,

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

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

Theorem 1 *For each k, N,*

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

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 each k,*

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

These results are proved below. Further results, dealing with the particular case of Boolean functions, are also discussed. 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. We conjecture, but have as yet been unable to prove, that direct connections or sigmoids should also improve these bounds by at least a factor of 2, just as in the two-dimensional case and in the worst-case analysis.

Because of Lemma 4.1, the last statements in Theorems 1 and 3 are consequences of the previous two. In fact, the Lema and therefore these results are valid for any function θ that satisfies the following two conditions, not just for the standard sigmoid discussed in the introduction:

(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 $\theta'(c) = \mu \neq 0$.

Note that the function $\mathcal{H}$ does not satisfy (S2), but the standard sigmoid and $\pi(x)$ of course do. Property (S1) can be replaced by:

(S1') $\lim_{x \to +\infty} \theta(x) = 1$ and $\lim_{x \to -\infty} \theta(x) = 0$

in the sense that, for any θ satisfying (S1),

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

satisfies (S1') and is so that (k, θ)-nets are the same as $(k, \widetilde{\theta})$-nets. So whenever we say that θ satisfies (S1) we may assume without loss of generality that it satisfies (S1') as well.

For fixed θ, and under mild assumptions on θ, such neural nets can be used to approximate uniformly arbitrary continuous functions on compacts. See for instance [5], [7]. In particular, they can be used to implement arbitrary dichotomies.

In neural net practice, one often takes θ to be the *standard sigmoid*

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

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

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

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

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

Most analysis has been done for $\mathcal{H}$ and no direct connections, but backpropagation techniques typically use the sigmoid (or equivalently tanh) and sometimes may also incorporate direct connections.

Remark 3.1 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

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

still implements the dichotomy. Since 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 Results

The main results that we prove are as follows, where

$$\underline{\mu}(k, \theta, N)$$

denotes $\underline{\mu}(\mathcal{F})$ for the class of (k,θ)-nets in $\mathbb{R}^N$ with no direct connections, and similarly for μ and $\overline{\mu}$, and a superscript d is used for the class of arbitrary such nets (with possible direct conections from input to output).

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 make $ax+by+c=0$ separate any point which has color 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 square, color them in "XOR" fashion, opposite vertices of 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, then

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

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, for each $\varepsilon > 0$, $\mathcal{S}$:= 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

We define a "neural net" as a function of a certain type, corresponding to the idea of feedforward interconnections, via additive links, of neurons each of which has a scalar response θ.

Let $\theta : \mathbb{R} \to \mathbb{R}$ be any function. We say that f is a *single hidden layer neural net with k hidden neurons of type θ* (or just that 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 $u \in \mathbb{R}^N$,

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

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

2 Dichotomies

Quantifying the classification power of a class of functions (such as those computable by neural nets with a fixed architecture and a fixed number of neurons) 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.

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

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

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 $f \in \mathcal{F}$.

Here we consider, for any class of functions $\mathcal{F}$ as above, 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 $1 \leq l$ (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 $1 \leq l$ (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 $1 \leq l$ (possibly ∞) for which the set 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

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

ON THE RECOGNITION CAPABILITIES OF FEEDFORWARD NETS

Eduardo D. Sontag*
SYCON - Rutgers Center for Systems and Control
Department of Mathematics
Rutgers University, New Brunswick, NJ 08903
Phone: (201)932-3072 *E-mail*: sontag@hilbert.rutgers.edu

ABSTRACT

In this note we deal with recognition capabilities of various feedforward neural net architectures, analyzing the effect of direct input to output connections and comparing Heaviside (threshold) with sigmoidal response units. The results state, roughly, that allowing direct connections or allowing sigmoidal responses doubles the recognition power of the standard architecture (no connections, Heaviside responses) which is often assumed in theoretical studies.

1 Introduction

In theoretical studies of feedforward neural networks, one often assumes that there are no direct connections from input to output neurons, and that each neuron computes a threshold function of its aggregate input (see for instance [3]). On the other hand, backpropagation uses sigmoidal responses, as a smooth unit is essential in applying gradient descent techniques. In this note, we show that using sigmoids rather than Heavisides results in at least twice the recognition power. This means on the one hand that less neurons are needed than may have been expected, explaining the somewhat surprising behavior found in some numerical experiments ([6]), and on the other hand that in order to assure generalization from examples, more training instances may be needed than would be the case with Heaviside units. If instead of sigmoids one uses again Heavisides but now allowing also direct i/o connections, the situation is analogous to the case with sigmoids.

The widespread intuition that sigmoids approximate Heaviside functions, and hence that their separation properties are somewhat similar, is based on a high-gain approximation of Heavisides. For small gains, however, using sigmoids one can get also approximations to linear maps (the tangents). This adds a considerable amount of separation power.

There are many possible measures of recognition capabilities for nets. These range from what we might call a "best-case" analysis (what is the cardinality of the largest set that can be arbitrarily partitioned using nets with a fixed architecture, the "Vapnik-Chervonenkis (VC) dimension" of this architecture) to a "worst-case" analysis (what is the largest cardinality so that every set of that cardinality can be so partitioned), and many others in between. This note will study a few of these measures. The results are roughly that direct connections or sigmoidal responses have at least twice the recognition power of the standard architecture (no connections, Heaviside responses). Except for worst-case analysis, the results are given for subsets of the plane; the general cases (in arbitrary dimensions) are still open.

*This research was supported in part by Siemens Corporate Research
Report SYCON 90-03, April 1990.