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, "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] Funahashi, K., "On the approximate realization of continuous mappings by neural networks," *Proc. Int. Joint Conf. on Neural Networks*, IEEE Publications, 1988, p. I.641-648.

[8] Hornik, K.M., M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks* **2**(1989): 359-366.

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

[10] Sontag, E.D., "On the recognition capabilities of feedforward nets," Report SYCON-90-03, *Rutgers Center for Systems and Control*, April 1990.

[11] Sontag, E.D., "Sigmoids distinguish better than Heavisides," *Neural Computation* **1**(1989): 470-472.

[12] Sontag, E.D., "Feedback Stabilization Using Two-Hidden-Layer Nets," Report SYCON-90-11, *Rutgers Center for Systems and Control*, October 1990.

[13] 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.

[14] Sontag, E.D., and H.J. Sussmann, "Backpropagation separates where perceptrons do," *Neural Networks*, 1991, to appear.

points, one can achieve an open dense set of values; the proof involves looking first at points with rational coordinates, and using that on such points one is dealing basically with rational functions (after a diffeomorphism), plus some theory of semialgebraic sets. We conjecture that one should be able to interpolate at $2k$ 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. For some examples, it is quite easy to get $2k$ points. For instance,

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

for the piecewise linear sigmoid π. The inequality $\underline{\lambda}(k, \pi) \geq 2k$ can be proved by arranging successive points in pairs and interpolating two at a time. The converse inequality follows from Lemma 8.1 and Equation (4). □

10 Further Remarks

The main conclusion from Theorem 1 is that sigmoids at least double recognition power for arbitrary sets. We conjecture that

$$\frac{\overline{\mu}(k, \sigma, N)}{\overline{\mu}(k, \mathcal{H}, N)} \;=\; 2 + O\left(\frac{1}{k}\right)$$

for all N; this is true for $N = 1$ and is strongly suggested by Theorem 3 (the first bound appears to be quite tight). Unfortunately the proof of this theorem is based on the result from [2] regarding arrangements of points in the plane, a fact which does not generalize to dimension three or higher.

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 threshols nets may be infinitely large (see [10]).

In the recent work [12], 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.

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 9.1. We now prove the second equality.

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 is 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.

Conversely, assume that $\{x_1, x_2, \ldots, x_{k+2}\}$ are $k+2$ given points, with $x_1 < \ldots < x_{k+2}$, and $y_1, \ldots, y_{k+2}$ are desired interpolation values. Let $\omega_0 + v_0 u$ be the line that interpolates (x_1, y_1) and (x_2, y_2). Consider the values $\widetilde{y}_3, \ldots, \widetilde{y}_{k+2}$ defined by $\widetilde{y}_i := y_i - (\omega_0 + v_0 x_i)$. There is then a piecewise constant map

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

that equals zero at x_1 and x_2 and equals $\widetilde{y}_i$ at each of the k points x_i, $i = 3, \ldots, k+2$. The sum of this map and the given linear map provide the wanted interpolation (exactly). ∎

Remark 9.3 Thus we can approximately interpolate any $2k-1$ points using k sigmoidal processors. It is not hard to prove as a corollary that, for the standard sigmoid, this approximate interpolation property holds in the following stronger sense: for an open dense set of $2k-1$

Now we prove Proposition 9.1. Assume that we have already proved that for any two *increasing* sequences of real numbers

$$x_0 < x_1 < \ldots < x_{2n} \text{ and } z_0 < z_1 < \ldots < z_{2n} \tag{5}$$

there is some (k, θ)-net so that

$$|f(x_i) - z_i| < \varepsilon/2 \tag{6}$$

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

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

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 (6) holds for each i. By Lemma 5.3, there are some numbers a, b, c so that

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

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

Thus, we must prove the result for the particular case of increasing sequences (5), 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 (5) and any $\varepsilon > 0$, there exists an (n, θ)-net f so that

$$|f(x_i) - z_i| < \varepsilon \text{ for each } i = 0, \ldots, 2n \tag{7}$$

and

$$|f(x) - z_{2n}| < \varepsilon \text{ for all } x \geq x_{2n} . \tag{8}$$

For $n = 1$ this follows from Lemma 9.2, 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

$$|f_1(x_i) - z_i| < \varepsilon/2 \text{ for each } i = 0, \ldots, 2n-2 \tag{9}$$

and

$$|f_1(x) - z_{2n-2}| < \varepsilon/2 \text{ for all } x \geq x_{2n-2} . \tag{10}$$

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 9.2, 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

$$|f_2(x)| < \varepsilon/2 \text{ for all } x < x_{2n-1} - \delta \tag{11}$$

and

$$|f_2(x) - \beta| < \varepsilon/2 \text{ for all } x > x_{2n-1} + \delta . \tag{12}$$

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.

be the largest integer $k \geq 1$, possibly infinite, so that for every set of data as above (3) can be solved. Note that, obviously,

$$\underline{\lambda}(\mathcal{F}) \leq \underline{\mu}(\mathcal{F}) . \tag{4}$$

One could also define analogues of μ and $\overline{\mu}$, but we do not do so, as no results will be provided for them.

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 $N = 1$.

The main technical fact is as follows:

Proposition 9.1 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 9.2 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 showed analogously. ∎

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 $(1 - 1/y_j^i)$, 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 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 $\tau := -\ln(\rho)$. ∎

9 Interpolation

In this Section we deal with the following approximate interpolation problem. Given a sequence of k (distinct) points $x_1, \ldots, x_k$ 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

$$f \in \mathcal{F} \text{ so that } |f(x_i) - y_i| < \varepsilon \text{ for each } i\,. \tag{3}$$

Let

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

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 7.1, 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 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. ∎

It is worth remarking that there are sigmoids θ, as differentiable as wanted, even real-analytic, where all classification measures are infinite. 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. (If only μ and $\overline{\mu}$ are desired to be infinite, one may also take the simpler example $\theta(x) = \sin(x)$. Note that for any l rationally independent real numbers x_i, the vectors of the form $(\sin(\gamma_1 x_1), \ldots, \sin(\gamma_l x_l))$, with the γ_i's real, form a dense subset of $[-1, 1]^l$, so all dichotomies on $\{x_1, \ldots, x_l\}$ can be implemented with $(1, \sin)$-nets.)

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 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 Lemma 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 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

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

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

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

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

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 5.3 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 4.2 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 $v_i.s_j - \tau_i$ 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 5.2. 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 5.3. ∎

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,*

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

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

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 $g(u) = f(v.u)$ 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 $\underline{\mu}(\mathcal{F}_1) > \underline{\mu}(\mathcal{F}_N)$.

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 $v.u_i$ 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

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

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

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

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

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 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.

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 a *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 $u \in \mathbb{R}^N$,

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

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

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 [7], [6], [8]. In particular, they can be used to implement arbitrary dichotomies. A typical choice is for θ to be the *standard sigmoid*

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

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

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

Most analysis has been done for $\mathcal{H}$ and no direct connections, but numerical techniques typically use the standard sigmoid (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 $t_+ \neq t_-$.

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, 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.

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 interpolation, and the paper concludes with various remarks. Many open problems are mentioned at various points.

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 [13] and [14] 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, appears in the Proceedings of the 1990 Conference on Neural Information Processing Systems.

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

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

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 $f \in \mathcal{F}$.

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.

connections is one of our objectives. Precise definitions are given later, but we now 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).

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 negative 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 the approximation theorems proved by [7], [6], and [8] (see also [4], [5], [12] for approximation problems where two layers are required). 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 both in its practical aspects (having rough estimates of how many processors will be needed is essential when applying backpropagation) and in evaluating generalization properties (larger nets tend to lead to poorer generalization). It is known and easy to prove (see for instance [1], [5]) that one can basically 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 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.

FEEDFORWARD NETS FOR INTERPOLATION AND CLASSIFICATION

Eduardo D. Sontag
SYCON - Center for Systems and Control
Department of Mathematics
Rutgers University, New Brunswick, NJ 08903
(908)932-3072; E-mail: sontag@hilbert.rutgers.edu

ABSTRACT

This paper deals with single-hidden-layer feedforward nets, studying various measures of classification power and interpolation capability. Results are given showing that direct input to output connections in threshold nets double the recognition but not the interpolation power, while using sigmoids rather than thresholds allows (at least) doubling both.

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.

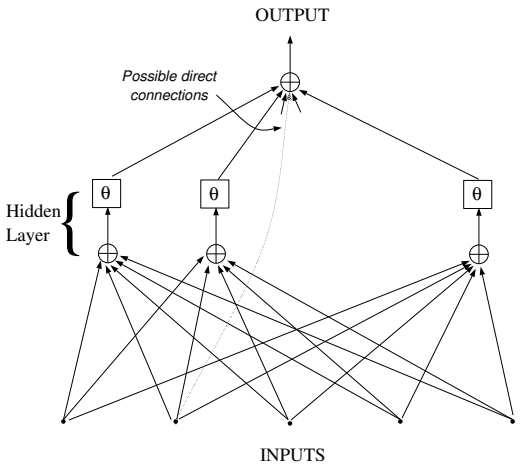

Figure 1: *Feedforward Net.*

The input fed to any given processor is a combination of the ouputs of all the processors that connect to it, weighted according to real-valued coefficients. (There are also special weights or "biases" associated to each processor; we mention these later). 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