Uniqueness of the weights for minimal feedforward nets with a given input-output map

Héctor J. Sussmann

Rutgers University

Abstract — *We show that, for feedforward nets with a single hidden layer, a single output node, and a "transfer function"* Tanh s*, the net is uniquely determined by its input-output map, up to an obvious finite group of symmetries (permutations of the hidden nodes, and changing the sign of all the weights associated to a particular hidden node), provided that the net is irreducible, i.e. that there does not exist an inner node that makes a zero contribution to the output, and there is no pair of hidden nodes that could be collapsed to a single node without altering the input-output map.*

Rutgers Center for Systems and Control, May 1991

Revised October 1991

Research supported in part by the Air Force Office of Scientific Research (AFOSR-91-0343).

The author thanks Eduardo Sontag for suggesting the problem and for his helpful comments and ideas, and an anonymous referee for suggesting how to improve the exposition at several points.

Requests for reprints should be sent to Héctor J. Sussmann, Department of Mathematics, Rutgers University, New Brunswick, NJ 08903.

§1. Introduction .

Feedworward neural nets with a single hidden layer have been shown to exhibit a number of remarkable properties, that makes them good candidates for nonlinear curve-fitting. They satisfy universal approximation properties (Cybenko, 1989, Hornik et al., 1989). More recently, it has been proved that they are actually *better* approximators than many other classes (e.g. polynomials) in that one needs a smaller number of parameters in order to approximate an arbitrary function in a certain class within a given error (Barron, 1991a, 1991b). Questions of interpolation using one-hidden layer nets, and in particular issues dealing with counting how many units are needed in order to achieve a given objective, are discussed by Sontag (1991b). (It is worth remarking that single-hidden layer nets are severely restricted in their approximation properties with respect to discontinuous target functions, as explained in Sontag, 1991a.)

We study the following question: to what extent is a feedforward net uniquely determined by its corresponding input-output map? Hecht-Nielsen (1989) pointed out that there are some obvious sources of non-uniqueness, arising from internal symmetries such as the possibility of relabeling the hidden nodes. He asked whether the small, discrete symmetry group $\mathcal{G}$ that he exhibited was actually the full symmetry group, i.e. whether the input-output map of the net determines the net modulo the symmetries in $\mathcal{G}$.

We give an affirmative answer to this question, by proving a uniqueness theorem: two "irreducible" nets with the same input-output map are related by a transformation in $\mathcal{G}$.

The irreducibility condition is needed because there is another source of nonuniqueness, coming from the fact that a net may contain nodes that make no contribution whatsoever to the output, e.g. nodes whose outgoing connection weights are equal to zero. This means that there are cases when a net can be *reduced* (i.e. some of its nodes can be removed) without changing the input-output map. Clearly, a uniqueness theorem can only hold for *irreducible nets.*

In this note we will consider feedforward nets with a single hidden layer, and a transfer function $\sigma(s) = \operatorname{Tanh} s$. For such nets, we will give a precise definition of reducibility and of the group $\mathcal{G}$ of internal symmetries, and show rigorously that an irreducible net is uniquely determined, up to symmetries in $\mathcal{G}$, by its input-output map.

As a byproduct of our results, it will follow that *an irreducible net is minimal.* Here we call a net "minimal" if its input-output map cannot be obtained from another net with fewer hidden nodes. So minimality is not, in principle, an easily checkable property. On the other hand, "irreduciblity" is defined by listing three very simple situations under which a net could be made smaller, and then calling the net irreducible if none of these situations occurs. So irreducibility is easily checked. The fact that irreducibility is equivalent to minimality means that there is no mechanism, other than the three listed in the definition of reducibility, that can be used to reduce the number of hidden nodes without changing the input-output map.

The paper is organized as follows: in §2 we define the main concepts and state the uniqueness theorem. The proof of the theorem is given in §3.

§2. Statement of the main results .

We will only consider feedforward nets with m input nodes, one layer of hidden neurons, one output node, and "transfer function" $\sigma(s) = \mathrm{Tanh}\, s$, so we will just use the word "net" to refer to this particular kind of neural net. The set of all the nets in this sense that have n hidden nodes will be denoted by $\mathcal{N}_{m,n}$, so that

1. $\mathcal{N}_{m,n}$ is the set of all networks with m input units, one output unit, n hidden units, and activation function Tanh,
2. a *net* is a member of $\mathcal{N}_{m,n}$ for some m, n.

Precisely, then, a *net* consists of the specification of (a) m *input nodes*, labeled $1, \ldots, m$, (b) n *hidden nodes*, labeled $1, \ldots, n$, (c) one *output node*, (d) an array $W = \{w_{ij}\}$ of weights for the connections from input nodes to hidden nodes, and (e) an array $C = \{c_j\}$ of weights for the connections from the hidden nodes to the output node. The w_{ij} are indexed by $i \in \{0, \ldots, m\}$, $j \in \{1, \ldots, n\}$, so that w_{ij} is the weight of the connection from the i-th input node to the j-th hidden node, and w_{0j} is the "bias" for the j-th hidden node. Similarly, the c_j are indexed by $j \in \{0, \ldots, n\}$.

Given an input vector $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$, the *net input* $\nu_j(x)$ of the j-th hidden node is the number

$$\nu_j(x) = w_{0j} + \sum_{i=1}^{m} w_{ij} x_i \ .$$

The *output* $y_j(x)$ of the j-th hidden node is

$$y_j(x) = \sigma(\nu_j(x)) = \mathrm{Tanh}(\nu_j(x)) \ .$$

The *net input* of the output node is

$$\mu(x) = c_0 + \sum_{j=1}^{m} c_j y_j(x) \ .$$

The *output* $z(x)$ of the net is the number $z(x) = \sigma(\mu(x))$. The function $x \to z(x)$ is the *input-output map* of the net.

For a fixed m, two nets are *I-O equivalent* if their corresponding input-output maps are the same. Our goal is to show that two I-O equivalent nets must necessarily be the same net, up to some simple internal symmetries.

Fix m. To a net with n hidden nodes we associate the n linear affine functions $\mathbb{R}^m \ni x \to \nu_j(x) \in \mathbb{R}$, $j = 1, \ldots, n$. We call two linear affine functions α, β on $\mathbb{R}^m$ *sign-equivalent* if $|\alpha(x)| = |\beta(x)|$ for all $x \in \mathbb{R}^m$. (This is easily seen to be equivalent to the condition that either $\alpha(x) = \beta(x)$ for all x or $\alpha(x) = -\beta(x)$ for all x.) We call a net *reducible* if one of the following conditions hold:

(I) one of the c_j, for $j = 1, \ldots, n$, vanishes;

(II) there exist two different indices $j_1, j_2 \in \{1, \ldots, n\}$ such that the functionals ν_{j_1}, ν_{j_2} are sign-equivalent.

(III) one of the functionals ν_j is a constant.

If a net N is reducible, then it is I-O equivalent to another net with fewer hidden nodes. Indeed, if N is reducible because (I) holds, and $j \in \{1, \ldots, n\}$ is such that $c_j = 0$, then the j-th node makes no contribution to the net input $\mu(x)$, so we can simply remove it without changing the output. Now suppose that N is reducible because (II) holds. Let j_1, j_2 be such that ν_{j_1} and ν_{j_2} are sign-equivalent. Let $\nu_{j_1}(x) \equiv \rho\nu_{j_2}(x)$, where $\rho = 1$ or $\rho = -1$. Then the combined contribution of the nodes j_1 and j_2 to $\mu(x)$ is $c_{j_1}\sigma(\nu_{j_1}(x)) + c_{j_2}\sigma(\nu_{j_2}(x))$, i.e. $c_{j_1}\sigma(\rho\nu_{j_2}(x)) + c_{j_2}\sigma(\nu_{j_2}(x))$, which is equal to $\rho c_{j_1}\sigma(\nu_{j_2}(x)) + c_{j_2}\sigma(\nu_{j_2}(x))$, i.e. to $(\rho c_{j_1} + c_{j_2})\sigma(\nu_{j_2}(x))$. (Here we used the fact that σ is an odd function, so $\sigma(\rho s) = \rho\sigma(s)$.) So we could obtain the same input-output map by just removing Node j_1 and changing the weight of the connection from Node j_2 to the output from c_{j_2} to $c'_{j_2} = \rho c_{j_1} + c_{j_2}$. Finally, if N is reducible because (III) holds, and ν_j is constant with value κ, then we can remove the j-th hidden node and change c_0 to $c_0 + \sigma(\kappa)$.

An net which is not reducible will be called *irreducible.* A net with n hidden nodes will be called *minimal* if it is not I-O equivalent to a net with fewer hidden nodes. We have shown above that a minimal net is necessarily irreducible. It will follow from our main theorem that, conversely, an irreducible net is minimal.

The are some obvious transformations that can be applied to a net N without changing its input-output map. For instance, suppose we pick a hidden node j and change the sign of all the weights w_{ij} for $i = 0, 1, \ldots, m$, and also change the sign of c_j. Since σ is odd, this will not alter the contribution of this node to the total net input $\mu(x)$. Let us use $\theta_j(N)$ to denote the net resulting from this transformation.

Another possibility is to interchange two hidden nodes, that is, to take two hidden nodes j_1 and j_2 and relabel j_1 as j_2 and j_2 as j_1, taking care to also relabel the corresponding weights. (Precisely: consider a new net with weights w_{ij}^{new}, c_j^{new}, such that $w_{ij_1}^{new} = w_{ij_2}$, $w_{ij_2}^{new} = w_{ij_1}$, $c_{j_1}^{new} = c_{j_2}$, $c_{j_2}^{new} = c_{j_1}$.) Call the resulting net $\tau_{j_1 j_2}(N)$. The maps $\tau_{j_1 j_2}$, θ_j generate a finite group $\mathcal{G}_{m,n}$ of transformations of the set $\mathcal{N}_{m,n}$. (The precise number of elements of this group is $2^n n!$.) Call two nets in $\mathcal{N}_{m,n}$ *equivalent* if they are related by a transformation in the group $\mathcal{G}_{m,n}$. It is then clear that two equivalent nets are I-O equivalent. Our main result says that the converse is true for irreducible nets:

Theorem 2.1. *Let N_1, N_2 be irreducible I-O equivalent nets in $\mathcal{N}_{m,n_1}$, $\mathcal{N}_{m,n_2}$. Then (i) $n_1 = n_2$ and (ii) N_1 and N_2 are equivalent.*

We will prove this theorem in the next section. First, we remark that the theorem implies:

Corollary 2.1. *An irreducible net is minimal.*

PROOF. Let N be irreducible. If N was not minimal, there would be another net N' with fewer hidden nodes that computes the same input-output function. If N' is not irreducible, then we could reduce the number of nodes, as explained above, without changing the input-output map. If the resulting net is not irreducible, we can reduce again. Continue this process until no further reduction is possible. We then get an irreducible net $\hat{N}$ which is I-O equivalent to N but has fewer nodes than N. By the theorem, $\hat{N}$ and N have the same number of hidden nodes. This contradiction completes the proof. ∎

§3. Proof of Theorem 2.1 .

Assume N_1 and N_2 are nets with n_1, n_2 hidden nodes. Use w_{ij}^1, c_j^1, w_{ij}^2, c_j^2, to denote the weights of N_1 and N_2. Similarly, use $\nu_j^1(x)$, $y_j^1(x)$, $\mu^1(x)$, $z^1(x)$, $\nu_j^2(x)$, $y_j^2(x)$, $\mu^2(x)$, $z^2(x)$, with an obvious meaning. Our hypothesis is that $z^1(x) = z^2(x)$ for all x. Since σ is one-to-one, this implies that $\mu^1(x) = \mu^2(x)$ for all x. So the following identity holds:

$$c_0^1 + \sum_{j=1}^{n_1} c_j^1 \sigma(\nu_j^1(x)) = c_0^2 + \sum_{j=1}^{n_2} c_j^2 \sigma(\nu_j^2(x)) \ . \tag{3.1}$$

Let $J = \{1, 2, \ldots, n_1 + n_2\}$, and define $\varphi_j = \nu_j^1$ for $1 \leq j \leq n_1$, $\varphi_j = \nu_{j-n_1}^2$ for $n_1 + 1 \leq j \leq n_1 + n_2$, $a_0 = c_0^1 - c_0^2$, $a_j = c_j^1$ for $1 \leq j \leq n_1$, $a_j = -c_{j-n_1}^2$ for $n_1 + 1 \leq j \leq n_1 + n_2$. Then (3.1) becomes:

$$a_0 + \sum_{j \in J} a_j \sigma(\varphi_j(x)) = 0 \ . \tag{3.2}$$

Lemma 3.1. *Let J be a finite set and let $\{\varphi_j\}_{j \in J}$ be a family of nonconstant linear affine functions on $\mathbb{R}^m$, no two of which are sign-equivalent. Then the functions $\sigma \circ \varphi_j$, $j \in J$ and the constant function 1 are linearly independent.*

PROOF. Assume that $\breve{a} + \sum_{j \in J} a_j \sigma \circ \varphi_j \equiv 0$. We have to prove that $\breve{a}$ and all the a_j vanish.

Write $\varphi_j(x) = \lambda_j(x) + \lambda_j^0$, where λ_j is a nonzero linear functional and λ_j^0 is a constant. (The fact that $\lambda_j \neq 0$ follows because φ_j is not a constant.) Our hypothesis guarantees that, if $j_1 \neq j_2$, then either (i) the functionals λ_{j_1} , λ_{j_2} are not sign-equivalent, or (ii) if $\lambda_{j_1} \equiv \pm\lambda_{j_2}$, then the corresponding constant terms $\lambda_{j_k}^0$ do not satisfy $\lambda_{j_1}^0 = \pm\lambda_{j_2}^0$ with the same choice of sign.

Define an equivalence relation on J by calling two elements j_1, j_2 of J equivalent if the corresponding linear functions λ_{j_1}, λ_{j_2}, are sign-equivalent. Let $\mathcal{E}$ be the set of equivalence classes. Pick a j_E for each $E \in \mathcal{E}$. As E varies over the classes in $\mathcal{E}$, no two of the functionals λ_{j_E} are sign-equivalent. So, for each pair E_1, E_2 of distinct members of $\mathcal{E}$, the set of points $x \in \mathbb{R}^m$ where $|\lambda_{j_{E_1}}(x)| \neq |\lambda_{j_{E_2}}(x)|$ is open and dense in $\mathbb{R}^m$. Also, for each E, the set of $x \in \mathbb{R}^m$ such that $\lambda_{j_E}(x) \neq 0$ is open and dense in $\mathbb{R}^m$, because λ_{j_E} is not $\equiv 0$. So we may pick an x such that $\lambda_{j_E}(x) \neq 0$ for all $E \in \mathcal{E}$, and $|\lambda_{j_{E_1}}(x)| \neq |\lambda_{j_{E_2}}(x)|$

for all pairs E_1, E_2 of distinct members of $\mathcal{E}$. Let $\tilde{\lambda}_j = \rho_j\lambda_j$, where $\rho_j = \pm 1$, the sign being chosen so that $\tilde{\lambda}_j(x) > 0$. Let $\tilde{\lambda}_j^0 = \rho_j\lambda_j^0$, $\tilde{\varphi}_j = \rho_j\varphi_j$, $\tilde{a}_j = \rho_j a_j$. Then

$$\breve{a} + \sum_{j\in J} \tilde{a}_j \sigma \circ \tilde{\varphi}_j \equiv 0 \tag{3.3}$$

as well, and our proof will be complete if we show that $\breve{a}$ and all the $\tilde{a}_j$ vanish. Moreover, if the $\tilde{a}_j$ vanish, then (3.3) shows that $\breve{a} = 0$ as well. So it will be enough to show that the $\tilde{a}_j$ vanish.

If $E \in \mathcal{E}$, then the functionals $\tilde{\lambda}_j$ for $j \in E$ are all sign-equivalent, and satisfy $\tilde{\lambda}_j(x) > 0$. So in fact all the $\tilde{\lambda}_j$ are equal to one and the same functional, that we will call $\tilde{\lambda}_E$. If $j, j' \in E$ and $j \neq j'$, then we know that $\lambda_j = \alpha\lambda_{j'}$, where $|\alpha| = 1$ and $\lambda_j^0 \neq \alpha\lambda_{j'}^0$. On the other hand, $\tilde{\lambda}_j(x) = \tilde{\lambda}_{j'}(x)$ and $\tilde{\lambda}_j(x) = \rho_j\lambda_j(x)$, $\tilde{\lambda}_{j'}(x) = \rho_{j'}\lambda_{j'}(x)$, so $\rho_{j'} = \rho_j\alpha$. Since $\lambda_j^0 \neq \alpha\lambda_{j'}^0$, we conclude that $\tilde{\lambda}_j^0 \neq \tilde{\lambda}_{j'}^0$. So we have shown that *the numbers $\tilde{\lambda}_j^0$, as j varies over an $E \in \mathcal{E}$, are all different.*

As E varies over $\mathcal{E}$, the numbers $\tilde{\lambda}_E(x)$ are positive and different. If $j \in E$, we have

$$\begin{aligned}\sigma(\tilde{\varphi}_j(tx)) &= \frac{e^{2\tilde{\varphi}_j(tx)} - 1}{e^{2\tilde{\varphi}_j(tx)} + 1} \\ &= \frac{\xi_j e^{2t\tilde{\lambda}_E(x)} - 1}{\xi_j e^{2t\tilde{\lambda}_E(x)} + 1} \ ,\end{aligned} \tag{3.4}$$

where $\xi_j = e^{2\tilde{\lambda}_j^0}$. (In particular, the ξ_j for j in a fixed $E \in \mathcal{E}$, are all different.) Since $\tilde{\lambda}_E(x) > 0$, and $\xi_j > 0$, we can conclude that

$$\lim_{t\to+\infty} \sigma(\tilde{\varphi}_j(tx)) = 1 \ . \tag{3.5}$$

It then follows from (3.3) and (3.5) that

$$\breve{a} + \sum_{j\in J} \tilde{a}_j = 0 \quad . \tag{3.6}$$

If we subtract (3.6) from (3.3) we get

$$\sum_{j\in J} \tilde{a}_j \zeta_j \equiv 0 \ , \tag{3.7}$$

where $\zeta_j = \sigma \circ \tilde{\varphi}_j - 1$, so that $\zeta_j(q) = -2\psi_j(q)$, and

$$\psi_j(q) = \frac{1}{1+\xi_j e^{2\tilde{\lambda}_j(q)}} \ . \tag{3.8}$$

So we have $\sum_{j\in J} \tilde{a}_j \psi_j \equiv 0$.

Now order the classes $E \in \mathcal{E}$ in a finite sequence $(E_1, E_2, \ldots, E_r)$, chosen so that $\tilde{\lambda}_{E_1}(x) < \tilde{\lambda}_{E_2}(x) < \cdots < \tilde{\lambda}_{E_r}(x)$. Let $v_k = \tilde{\lambda}_{E_k}(x)$. We then have

$$\psi_j(-tx) = (1 + \xi_j e^{-2tv_k})^{-1} \ ,$$

if $j \in E_k$. For each j, let $k(j)$ be the k such that $j \in E_k$. For $t > 0$ and sufficiently large, we have $0 < \xi_j e^{-2tv_{k(j)}} < 1$, and so we can expand $\psi_j(-tx)$ in a convergent power series:

$$\psi_j(-tx) = \sum_{s=0}^{\infty} (-1)^s \xi_j^s e^{-2tsv_{k(j)}} \ . \tag{3.9}$$

If we multiply (3.9) by $\tilde{a}_j$ and sum over j, we get

$$0 = \sum_{j\in J} \sum_{s=0}^{\infty} (-1)^s \tilde{a}_j \xi_j^s e^{-2tsv_{k(j)}} \ ,$$

i.e.

$$0 = \sum_{s=0}^{\infty} \sum_{k=1}^{r} e^{-2tsv_k} \sum_{j\in E_k} (-1)^s \tilde{a}_j \xi_j^s \ . \tag{3.10}$$

We rewrite (3.10) as

$$0 = \sum_{v\in\mathbf{R}, v\geq 0} e^{-2tv} \sum_{s\in\{0,1,\ldots\}, k\in\{1,\ldots,r\}, sv_k = v} \ \sum_{j\in E_k} (-1)^s \tilde{a}_j \xi_j^s \ . \tag{3.11}$$

The index of summation v is a nonnegative real number, but in fact the only v's occurring in the summation are those that can be expressed as integral multiples of some v_k. So these v's form a discrete subset Δ of the half-line $[0, \infty[$. If we order the elements of Δ as a sequence $v^1, v^2, \ldots$, such that $0 < v^1 < v^2 < \cdots$, then it follows easily from (3.11) that all the coefficients

$$\kappa^\ell \stackrel{\text{def}}{=} \sum_{s\in\{0,1,\ldots\}, k\in\{1,\ldots,r\}, sv_k = v^\ell} \ \sum_{j\in E_k} (-1)^s \tilde{a}_j \xi_j^s \tag{3.12}$$

vanish. (This is proved by induction on ℓ: if all the κ^ℓ for $\ell < \bar{\ell}$ are equal to zero, then (3.11) implies that $0 = \kappa^{\bar{\ell}} \mathrm{e}^{-2tv^{\bar{\ell}}} + o(\mathrm{e}^{-2tv^{\bar{\ell}}})$ as $t \to +\infty$, so $\kappa^{\bar{\ell}} = 0$.)

We now fix a $k \in \{1, \ldots, r\}$, and assume that we have already proved that the $\tilde{a}_j$ vanish for $j \in E_{k'}$, $k' < k$. Let α be the number of indices j belonging to E_k. Choose integers $s > 0$, $h > 0$, such that the α numbers sv_k, $(s+h)v_k$, $(s+2h)v_k$, …, $(s+(\alpha-1)h)v_k$ are not integral multiples of $v_{k'}$ for any $k' \in \{k+1, \ldots, r\}$. (To see that such a choice is possible, let B be the subset of $\{k+1, \ldots, r\}$ consisting of those k' such that the quotient $v_{k'}/v_k$ is rational. Write this quotient as $p_{k'}/q_{k'}$, where $p_{k'}$, $q_{k'}$ are relatively prime positive integers, and $p_{k'} > q_{k'} \geq 1$, because $v_k < v_{k'}$ if $k' > k$. Then sv_k cannot be an integer multiple of $v_{k'}$ unless s is divisible by $p_{k'}$. So, if we pick $h = \prod_{k' \in B} p_{k'}$ and $s = 1 + h$, we see that, if δ is an integer, then $s + \delta h$ can never be divisible by $p_{k'}$, because $p_{k'} > 1$. So this s and h have the desired property for all the indices $k' \in B$. Moreover, it is clear that $(s + \delta h)v_k$ cannot be an integer multiple of $v_{k'}$ if $v_{k'}/v_k$ is irational. So in fact s and h work for all the indices $k' \in \{k+1, \ldots, r\}$.

Now let δ be an integer such that $0 \leq \delta < \alpha$. Then $(s + \delta h)v_k$ is equal to v^ℓ for some ℓ. The corresponding coefficient κ^ℓ is equal to the sum

$$\sum_{s' \in \{0,1,\ldots\}, k' \in \{1,\ldots,r\}, s'v_{k'} = v^\ell} \; \sum_{j \in E_{k'}} (-1)^{s'} \tilde{a}_j \xi_j^{s'} \ .$$

This sum contains no contribution coming from values of k' such that $k' < k$, because we are assuming that all the corresponding $\tilde{a}_j$ vanish. A k' such that $k' > k$ can only contribute to the sum if $s'v_{k'} = v^\ell$ for some integer s', and this can only happen if $(s + \delta h)v_k$ is an integral multiple of $v_{k'}$. Since this is impossible by our choice of s and h, we conclude that the sum only contains contributions coming from $k' = k$. In this case, the only possible value of s' is $s + \delta h$. So

$$0 = \kappa^\ell = \sum_{j \in E_k} (-1)^{s+\delta h} \tilde{a}_j \xi_j^{s+\delta h} \ .$$

Therefore

$$\sum_{j \in E_k} \tilde{a}_j \xi_j^{s+\delta h} = 0$$

for $\delta = 0, 1, \ldots, \alpha - 1$. Write $b_j = \tilde{a}_j \xi_j^s$, $g_j = \xi_j^h$. Then

$$\sum_{j \in E_k} b_j g_j^\delta = 0 \;\text{ for }\; \delta = 1, \ldots, \alpha - 1 \ . \tag{3.13}$$

Now, we know that the numbers ξ_j, for $j \in E_k$, are all different. So the g_j are all different. So (3.13) is a system of α linear homogeneous equations for the b_j, whose coefficients form a Vandermonde matrix. Since the determinant of this matrix is $\neq 0$, because the g_j are all different, we conclude that all the b_j vanish. But this implies that the $\tilde{a}_j$ vanish as well.

It follows by induction from the above that all the $\tilde{a}_j$, for all $j \in E_k$, for all k, are equal to zero. As explained above, this proves our lemma. ∎

We now return to the proof of the theorem. Assume that not all the a_j are zero. Then the lemma and (3.2) imply that either (i) one of the φ_j is a constant, or (ii) two of the φ_j are sign-equivalent. The first possibility is excluded because both nets under consideration are irreducible. (Cf. Condition (III) of the definition of reducibility.) So (ii) must hold. But, since no two φ_j coming from the same net can be sign-equivalent (by Condition (II) of the definition of reducibility), there must exist j_1, j_2 such that $1 \leq j_1 \leq n_1$, $n_1 + 1 \leq j_1 \leq n_1 + n_2$, such that φ_{j_1} and φ_{j_2} are sign-equivalent. Moreover, for no other index $j \in J$ can φ_j be sign-equivalent to φ_{j_1}. So we can split the left-hand side of (3.2) into two parts, namely, (a) the sum $a_{j_1}\sigma(\varphi_{j_1}(x)) + a_{j_2}\sigma(\varphi_{j_2}(x))$ and (b) the other terms. The other terms involve the function 1 and functions φ_j that are not constant and not sign-equivalent to φ_{j_1}. So, by the lemma, the sum $a_{j_1}\sigma(\varphi_{j_1}(x)) + a_{j_2}\sigma(\varphi_{j_2}(x))$ must vanish. This means that either

$$\nu^1_{j^1} \equiv \nu^2_{\hat{j}^1} \text{ and } c^1_{j^1} = c^2_{\hat{j}^1}$$

or

$$\nu^1_{j^1} \equiv -\nu^2_{\hat{j}^1} \text{ and } c^1_{j^1} = -c^2_{\hat{j}^1} \ ,$$

where we have written $j^1 = j_1$ and $\hat{j}^1 = j_2 - n_1$.

Using this, we can rewrite (3.2) removing the contribution from j_1 and j_2, and apply the lemma again to the resulting identity. This will give rise to a second pair of hidden nodes j^2, $\hat{j}^2$ such that either

$$\nu^1_{j^2} \equiv \nu^2_{\hat{j}^2} \text{ and } c^1_{j^2} = c^2_{\hat{j}^2}$$

or

$$\nu^1_{j^2} \equiv -\nu^2_{\hat{j}^2} \text{ and } c^1_{j^2} = -c^2_{\hat{j}^2} \ .$$

This procedure can be continued until we end up with an identity where all the remaining coefficients a_j vanish. By the irreducibility of the nets, at this point there cannot be any terms left other than a_0, because all the c_j for $j \neq 0$ are supposed to be nonzero (cf. Condition (I)). So $a_0 = 0$, i.e. $c^1_0 = c^2_0$. Moreover, we have constructed sequences $j^1, j^2, \ldots, j^k$, $\hat{j}^1, \hat{j}^2, \ldots, \hat{j}^k$, of nodes of N_1, N_2, such that each ν_{j^ℓ} is sign-equivalent to $\nu_{\hat{j}^\ell}$ and to no other ν_j coming from N_2, and that the j^ℓ, $\hat{j}^\ell$ exhaust all the nodes of N_1, N_2. This means that $n_1 = n_2 = k$. If we make a permutation of the nodes of N_2 (which transforms N_2 into an equivalent net) we may assume that $\hat{j}^\ell = j^\ell$ for each ℓ, so that in fact $\nu^1_j \equiv \pm\nu^2_j$ for each j, and $c^1_j = \pm c^2_j$, with the same choice of sign in both. So N_1 and N_2 are equivalent, as stated. ∎

REFERENCES

Barron, A. R. (1991a). Approximation bounds for superpositions of a sigmoidal function. *Proceedings of the IEEE International Symposium on Information Theory.* IEEE Press.

Barron, A. R. (1991b). Universal approximation bounds for superpositions of a sigmoidal function. Technical Report #58, Statistics Department, University of Illinois at Urbana-Champaign.

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

Hecht-Nielsen, R. (1989) Theory of the Backpropagation Neural Network. *Proceedings of the International Joint Conference on Neural Networks, Washington, 1989.* IEEE Publications, NY, 593–605.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks* **2**, 359-366.

Sontag, E.D. (1991a). Capabilities of four- vs three-layer nets, and control applications. *Proceedings of the Conference on Information Sciences and Systems.* John Hopkins University, 558-563.

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