On the equivalence of Two Layered Perceptrons with Binary Neurons

Marcelo Blatt and Eytan Domany

Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot 76100, Israel

Ido Kanter

Department of Physics, Bar Ilan University, 52900 Ramat Gan, Israel

July 9, 1995

Abstract

We consider two-layered perceptrons consisting of N binary input units, K binary hidden units and one binary output unit, in the limit $N \gg K \geq 1$. We prove that the weights of a regular irreducible network are uniquely determined by its input-output map up to some obvious global symmetries. A network is regular if its K weight vectors from the input layer to the K hidden units are linearly independent. A (single layered) perceptron is said to be irreducible if its output depends on everyone of its input units;

and a two-layered perceptron is irreducible if the $K+1$ perceptrons that constitute such network are irreducible. By global symmetries we mean, for instance, permuting the labels of the hidden units. Hence, two irreducible regular two-layered perceptrons that implement the same Boolean function must have the same number of hidden units, and must be composed of equivalent perceptrons.

1 Introduction

In most applications dealing with learning and pattern classification, neural networks are viewed as input-output devices whose parameters (*i.e.* weights) are tuned in order to fit the training data. Fairly efficient learning algorithms exist for feed-forward networks with continuous-valued units (back-propagation [Rumelhart *et al.* 1986]) as well as for the binary case (CHIR [Grossman *et al.* 1989; Nabutovsky *et al.* 1990]). A natural question that arises is: to what extent are the weights of a feed-forward network determined by its input-output map. Hecht-Nielsen (1990) pointed out that there are some sources of non-uniqueness, arising from the internal symmetries such as permutation of the hidden units. He emphasized the importance of understanding the structure of equivalence classes of the weights which give rise to similar mappings, in order to reduce the volume in weight space where search (*i.e.* learning) is performed; instead of searching the entire space of weights, it is sufficient to pick one representative for each class.

Recently, Sussman (1992) showed that a two-layered perceptron, whose activation function is the hyperbolic tangent, is uniquely determined by its input-output mapping up to a finite group of symmetries. Albertini and Sontag (1993) generalized this result for activation functions $s(x)$ satisfying $s(0) = 0$, $s'(0) \neq 0$ and $s''(0) = 0$. This work was extended by Kurkova and Kainen (1994) for the case of asymptotically bounded, non-constant activation functions.

The issue of two-layered perceptrons of binary units was analyzed by Priel *et al.* (1994). They pointed out that any two-layered perceptron (2LP) with a finite number of hidden units is equivalent to the union of a finite number of *restricted* 2LP's. A restricted 2LP has fixed second layer weights. For instance, the committee machine (all the second layer weights are equal to one) is a restricted 2LP. Priel *et al.* found that all 2LP's with three hidden units are equivalent either to the committee machine or to a ruler machine (the output is determined by *a single* hidden unit). This means that for every 2LP with three hidden units there exists either a committee machine of three hidden units or a single layer perceptron, that implements the same Boolean function. For five hidden units the most general 2LP is equivalent to one of four possible restricted two layer perceptrons. They have also shown that two different restricted machines implement two completely distinct sets of Boolean functions in the limit of large number of inputs.

This paper deals with two-layered perceptrons with K binary units and N binary inputs. The problem we address is to understand to what extent the input-output map of the network, which is a Boolean function from $\{-1,1\}^N$ to $\{-1,1\}$, determines its weights and number of hidden units, K.

Various definitions are introduced in Section 2; a net is *minimal* if no 2LP with fewer hidden units implements the same Boolean function; *irreducible* when it has no redundant units; and *regular* if the first-layer weight vectors are linearly independent.

The main result of the paper is to show that for large $N \gg K$, a regular irreducible network is minimal and its weights are uniquely determined by its input-output map, up to some obvious symmetries.

In Section 3 we derive an expression for the *generalization error*, ϵ_g, which is a measure of similarity of the functions implemented by two 2PL's[a]. This expression is used in Section 4 to prove our main result.

2 The Model and Definitions

We consider feed forward networks with N input units, one layer of K hidden units, a single output unit, and $s(x) = \text{sign}(x)$ as the *transfer function*. A network of this kind, the two-layered perceptron, is completely defined by the

[a] $\epsilon_g = 0$ means that the input-output map implemented by two networks is exactly the same

specification of: *i)* K vectors $\vec{W}_l \in \Re^N$, with W_{li} ($1 \leq i \leq N$ and $1 \leq l \leq K$) denoting the weight of the connection of the i^{th} input unit to the l^{th} hidden unit, and *ii)* the second layer weight vector $\vec{w} \in \Re^K$, where w_l ($1 \leq l \leq K$) stands for the connection of the l^{th} hidden unit to the output. Given an input vector $\vec{x} = (x_1, \ldots, x_{\text{N}}) \in \{-1, 1\}^N$, the *local field*, h_l, given by

$$h_l(\vec{x}) = \sum_{i=1}^{N} W_{li}\ x_i = \vec{W}_l \cdot \vec{x}$$

is the *input* on the l^{th} hidden unit. The *output* of each unit is given by the transfer function; *i.e.*

$$s_l(\vec{x}) = \text{sign}(h_l) \qquad \qquad \text{for } 1 \leq l \leq K\ .$$

Each hidden unit can be regarded as the output of a single layer perceptron. The state taken by the hidden layer in response to an input is called the *Internal Representation* (IR) corresponding to this input. The output unit can also be considered as a single layer perceptron, of weights $\vec{w}$, whose output is determined by the IR, so the net output is:

$$y\,(\vec{x}) = \text{sign}(\vec{w}\cdot\vec{s}(\vec{x}))$$

In this scheme, a 2LP of N inputs and K hidden units is composed by K threshold-less perceptrons of N inputs units whose outputs are just the input units of the remaining perceptron.

The Boolean function $y : \{-1,1\}^N \to \{-1,1\}$ is the *input-output map* of the 2LP.

For a fixed N, two 2LP's are *equivalent* if their corresponding input-output maps are the same. This definition is also used for the equivalence of two perceptrons.

The symmetry operations mentioned in the abstract and in the last part of Section 1 are some obvious transformations that can be applied to a 2LP without affecting the input-output map. These symmetries are: *i)* Changing the sign of all the weights entering and leaving a hidden unit. Such an operation is a "gauge" symmetry. *ii)* Permuting the labels of the hidden units. *iii)* Replacing any one of the constituent perceptrons by another one that is equivalent to it [Dertouzos 1964; Priel *et al.* 1994]. *iv)* The norm of the weight vectors can be changed.

For reasons that will become clear later we normalize (without loss of generality) so that $\|\vec{W}_l\| = \sqrt{N}$ for $l = 1, \ldots, K$.

An input-unit l of a perceptron is called *redundant* if its output does not depend on it; *e.g.*, if for all possible inputs $\vec{\sigma} \in \{-1,1\}^A$ we have $y(\sigma_1, \ldots, \sigma_l, \ldots, \sigma_{\mathrm{A}}) = y(\sigma_1, \ldots, -\sigma_l, \ldots, \sigma_{\mathrm{A}})$ ($A = N$ for the perceptrons composing the first layer and $A = K$ for the perceptron of the second). For instance, a input unit is redundant if the weight associated to it is null. A perceptron that does not have redundant hidden units is called *irreducible*. Extending this definition for a 2LP, we say that a 2LP is *irreducible* if its $K+1$ perceptrons are irreducible.

We call a 2LP *regular* if the first layer weight vectors $\vec{W}_l$ $(1 \leq l \leq K)$ are linearly independent. In particular, if $K \ll N$ this is the generic case, since a finite number (K) of vectors with $N \gg K$ components will be linearly dependent only in a subset of measure zero of cases.

A net with K hidden units is called *minimal* if it is not equivalent to a net with fewer hidden units. By definition, a minimal network is necessarily irreducible. We will show that a regular irreducible network is minimal. Note that regularity and irreducibility are conditions imposed on first and second layer weights separately, while to be minimal is a requirement for the whole 2LP.

3 The Generalization Error

The generalization error, ϵ_g, is an index of similarity of the input-output map implemented by two networks. ϵ_g is the fraction of the input space for which two networks give different outputs. Clearly two nets are equivalent if and only if $\epsilon_g = 0$, *e.g.* when their generalization error vanishes. In this work we use this definition in order to prove whether or not two networks are equivalent.

First we analyze the case of two single layer perceptrons and then turn to the case of 2LP's, to establish the conditions such that ϵ_g vanishes. We obtain an expression for ϵ_g which is used in Section 4 to prove our uniqueness theorem.

The equivalence between two (single layer) perceptrons was studied in the framework of *threshold logic* [Dertouzos, 1964; Lewis II and C.L. Coates, 1967]. In the case of a finite number of inputs N, there is no simple expression for ϵ_g. Explicit classes of equivalence up to $N = 6$ are presented by Dertouzos (1964). In the large N limit, *i.e.* $N \gg 1$, the well known result (Györgyi and Tishby 1990, Opper *et al.* 1990, Seung *et al.* 1992) for the generalization error between two irreducible perceptrons whose weights are $\vec{W}_1$ and $\vec{W}_2$ (with norm $\sqrt{N}$) is:

$$\epsilon_g\Big(\vec{W}_1; \vec{W}_2\Big) = \frac{1}{\pi} \arccos\left(\frac{\vec{W}_1 \cdot \vec{W}_2}{N}\right) + \mathcal{O}\Big(\frac{1}{N}\Big)$$

In this limit the generalization error between two perceptrons vanishes as the angle between them goes to zero. Since we assumed that the norm is fixed, we conclude that if two perceptrons implement the same input-output map, then $\vec{W}_1 = \vec{W}_2$.

Let us now consider two 2LP, $\mathcal{N}^{(1)}$ and $\mathcal{N}^{(2)}$, both with N inputs and, respectively, $K^{(1)}$ and $K^{(2)}$ hidden units. We assume that $\mathcal{N}^{(1)}$ is regular. Use $W_{li}^{(1)}$, $w_l^{(1)}$ to denote the weights of $\mathcal{N}^{(1)}$ and $W_{li}^{(2)}$, $w_l^{(2)}$ those corresponding to $\mathcal{N}^{(2)}$. Similarly, use $h_l^{(1)}$, $h_l^{(2)}$, $s_l^{(1)}$, $s_l^{(2)}$, $y^{(1)}(\vec{x})$ and $y^{(2)}(\vec{x})$, with the same convention.

The generalization error between them is given by:

$$\epsilon_g\Big(\mathcal{N}^{(1)}; \mathcal{N}^{(2)}\Big) \;=\; \Big\langle\!\Big\langle \Theta\Big(-y^{(1)}\, y^{(2)}\Big) \Big\rangle\!\Big\rangle \tag{1}$$

where $\Theta(\cdot)$ is the Heaviside step function and $\langle\!\langle\cdots\rangle\!\rangle \;=\; \frac{1}{2^N}\sum_{x_1=-1,1}\cdots\cdots\sum_{x_{\rm N}=-1,1}\cdots$ indicates the average over input space.

$\epsilon_g\Big(\mathcal{N}^{(1)};\mathcal{N}^{(2)}\Big)$ can be expressed in terms of the IR as follows:

$$\epsilon_g\Big(\mathcal{N}^{(1)};\mathcal{N}^{(2)}\Big) \;=\; \sum_{\nu=1}^{2^{K^{(1)}}}\sum_{\eta=1}^{2^{K^{(2)}}} \Theta\Big(-(\vec{w}^{(1)}\cdot\vec{\sigma}^{\nu})\;(\vec{w}^{(2)}\cdot\vec{\sigma}^{\eta})\Big)\;\mathrm{P}(\vec{\sigma}^{\nu};\;\vec{\sigma}^{\eta}) \tag{2}$$

where $\{\vec{\sigma}^{\nu}\}_{1\leq\nu\leq 2^{K^{(1)}}}$ and $\{\vec{\sigma}^{\eta}\}_{1\leq\eta\leq 2^{K^{(2)}}}$ are the sets of IR for $K^{(1)}$ and $K^{(2)}$ hidden units respectively; $\mathrm{P}(\vec{\sigma}^{\nu};\;\vec{\sigma}^{\eta})$ is the fraction of input space for which the two 2LP's get the IR's $\vec{\sigma}^{\nu}$ and $\vec{\sigma}^{\eta}$ simultaneously (*i.e.* in response to the same input):

$$\mathrm{P}\Big(\vec{s}^{(1)};\vec{s}^{(2)}\Big) \;=\; \left\langle\!\!\left\langle \prod_{l=1}^{K^{(1)}}\Theta\left(\frac{h_l^{(1)}s_l^{(1)}}{\sqrt{N}}\right)\;\prod_{l=1}^{K^{(2)}}\Theta\left(\frac{h_l^{(2)}s_l^{(2)}}{\sqrt{N}}\right)\right\rangle\!\!\right\rangle \tag{3}$$

The argument of the Θ-functions were normalized by $\frac{1}{\sqrt{N}}$, for convenience (see below). We now define the (symmetric) correlation matrix:

$$A \;=\; \begin{pmatrix} R^{11} & R^{12} \\ \left(R^{12}\right)^T & R^{22} \end{pmatrix}$$

where the elements of R^{ab} $(a,b\,=\,1,2)$ are the correlations between the weight vectors of $\mathcal{N}^{(a)}$ and $\mathcal{N}^{(b)}$;

$$R^{ab}_{lm} \;=\; \frac{\vec{W}_l^{(a)}\cdot\vec{W}_m^{(b)}}{N} \qquad\quad \text{for}\quad a,b=1,2;\quad 1\leq l\leq K^{(a)};\quad 1\leq m\leq K^{(b)}$$

Clearly, A is a $\Big(K^{(1)}{+}K^{(2)}\Big)\times\Big(K^{(1)}{+}K^{(2)}\Big)$ matrix of rank M, with $K^{(1)}\leq M\leq K^{(1)}+K^{(2)}$; M is the number of independent vectors in the set $\Big\{\vec{W}_l^{(1)},\vec{W}_m^{(2)}\Big\}$ $(1\leq l\leq K^{(1)};\;1\leq m\leq K^{(2)})$.

Introducing the integral expression of the Θ-function

$$\Theta(z) = \int_0^{\infty}\mathrm{d}h\int_{-\infty}^{\infty}\frac{\mathrm{d}\hat{h}}{2\pi}\exp\Big\{i\hat{h}(h-z)\Big\}$$

and the Fourier expansion of the Dirac-delta function

$$\delta(z) = \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{2\pi} \exp(itz)$$

in eq. (3), we obtain [Priel *et al.* 1994]:

$$\mathrm{P}(\tilde{s}) = \left[\int_0^{\infty} \prod_{l=1}^{K^{(1)}+K^{(2)}} \mathrm{d}\tilde{h}_l \; \rho\Big(\tilde{h}\Big) \right] + \mathcal{O}\Big(\frac{1}{N}\Big) \tag{4}$$

with

$$\rho\Big(\tilde{h}\Big) = \prod_{m=1}^{M} \left[\frac{1}{\sqrt{2\pi \; \lambda_m}} \exp\left\{ - \frac{\Big(\vec{\psi}_m \cdot \tilde{h}\Big)^2}{2\lambda_m} \right\} \right] \prod_{M+1}^{K^{(1)}+K^{(2)}} \delta\Big(\vec{\psi}_m \cdot \tilde{h}\Big) \tag{5}$$

we denoted by $\tilde{h}$ the vector whose $K^{(1)}+K^{(2)}$ components are given by $\tilde{h}_l = h_l^{(1)}$ for $1 \leq l \leq K^{(1)}$ and $\tilde{h}_{K^{(1)}+l} = h_l^{(2)}$ for $1 \leq l \leq K^{(2)}$; the same convention is used to define $\tilde{s}$. $\vec{\psi}_m$ are the eigenvectors of the matrix $B_{lm} = \tilde{s}_l A_{lm} \tilde{s}_m$, and λ_m are the corresponding eigenvalues $(1 \leq m \leq K^{(1)}+K^{(2)})$. Clearly, A and B have the same eigenvalues, which can be labeled so that $\lambda_m \neq 0$ if $1 \leq m \leq M$ and $\lambda_m = 0$ for $M < m \leq K^{(1)}+K^{(2)}$.

Note that $\rho\Big(\tilde{h}\Big)$ is the *joint distribution* of the $K^{(1)}+K^{(2)}$ local fields given an internal representation for each of the two 2PL's. The eigenvectors $\vec{\psi}_m$ with eigenvalue $\lambda_m = 0$ span a subspace Ω. If the vector $\tilde{h}$ has a non-vanishing component in Ω, it has zero probability density, $\rho\Big(\tilde{h}\Big) = 0$. On the other hand all $\tilde{h}$ that are perpendicular to Ω appear with non-vanishing probability.

In the simplest case, the set $\left\{\vec{W}_l^{(1)}, \vec{W}_m^{(2)}\right\}$ $(1 \leq l \leq K^{(1)};\ 1 \leq m \leq K^{(2)})$ is linearly independent; $M = K^{(1)}+K^{(2)}$ and hence $\det(B) \neq 0$. Thus the integrand (5) can be written as:

$$\rho\left(\tilde{h}\right) = \frac{1}{\sqrt{(2\pi)^{K^{(1)}+K^{(2)}} \det(B)}} \ \exp\left[-\frac{1}{2} \sum_{m,n=1}^{K^{(1)}+K^{(2)}} \tilde{h}_m \left(B^{-1}\right)_{mn} \tilde{h}_n\right] . \qquad (6)$$

In general, not all vectors in the set $\left\{\vec{W}_l^{(1)}, \vec{W}_m^{(2)}\right\}$ are linearly independent. It is simple to see that the rank of A equals the number independent $\vec{W}$ vectors. Since $\mathcal{N}^{(1)}$ is regular and permuting the labels of the hidden units of $\mathcal{N}^{(2)}$ does not alter its input-output mapping, we can assume without loss of generality that $\left\{\vec{W}_m^{(1)}, \vec{W}_l^{(2)}\right\}$ $(1 \leq m \leq K^{(1)};\ 1 \leq l \leq M - K^{(1)})$ are linearly independent, and that the remaining weight vectors of $\mathcal{N}^{(2)}$ are a linear combination of these M;

$$\vec{W}_l^{(2)} = \sum_{m=1}^{K^{(1)}} a_{lm}\, \vec{W}_m^{(1)} + \sum_{m=1}^{M-K^{(1)}} b_{lm}\, \vec{W}_m^{(2)} \qquad \text{for} \quad M - K^{(1)} < l \leq K^{(2)} \quad (7)$$

Integrating (4) over $h_l^{(2)}$, $M - K^{(1)} < l \leq K^{(2)}$ we get:

$$\mathrm{P}(\tilde{s}) = \left[\int_0^\infty \prod_{l=1}^{M} \mathrm{d}\tilde{h}_l\, \rho'\left(\tilde{h}\right)\right] + \mathcal{O}\left(\frac{1}{N}\right)$$

with:

$$\rho'\left(\tilde{h}\right) = \frac{1}{\sqrt{(2\pi)^{M} \det(B')}} \ \exp\left[-\frac{1}{2} \sum_{m,n=1}^{M} \tilde{h}_m \left(B'^{-1}\right)_{mn} \tilde{h}_n\right] \qquad (8)$$

in this case we denoted by $\tilde{s}$ the vector whose M components are given by $\tilde{s}_l = s_l^{(1)}$ for $1 \leq l \leq K^{(1)}$ and $\tilde{s}_{K^{(1)}+l} = s_l^{(2)}$ for $1 \leq l \leq M - K^{(1)}$; the same convention is used for $\tilde{h}$. The elements of the matrix B' are given by $B'_{lm} = \tilde{s}_l A_{lm} \tilde{s}_m$, for $1 \leq l, m \leq M$.

Therefore any assignment of M local fields, $\left(h_1^{(1)},\ldots,h_{K^{(1)}}^{(1)};h_1^{(2)},\ldots,h_{M-K^{(1)}}^{(2)}\right)$ will appear for a finite fraction of the input space. For each choice of these M local fields, the remaining ones are uniquely determined by

$$h_l^{(2)} = \sum_{m=1}^{K^{(1)}} a_{lm}\, h_m^{(1)} + \sum_{m=1}^{M-K^{(1)}} b_{lm}\, h_m^{(2)} \qquad \text{for } M-K^{(1)} < l \leq K^{(2)}\,. \tag{9}$$

Equations (2) and (4) provide, in principle, a constructive method to evaluate the generalization error for any pair of 2LP. It is not always possible to compute eq. (4), but there are some cases (like that of *non-overlapping receptive fields*, when the correlation matrix is of the form $R_{lm}^{ab} = \delta_{lm}\delta_{ab} + r_l\,\delta_{lm}(1-\delta_{ab})$) when it is possible to explicitly integrate eq. (4). Examples of the calculation of ϵ_g for some pairs of two layered machines are presented by Priel *et al.* (1994).
Another situation in which the integral can be computed is when the first layer weights of both 2LP are linearly independent; Kendall's expansion [Kendall 1987, Saad 1994] can be used to evaluate equation (4).

4 Main result

Assume we have an regular irreducible 2LP, denoted by $\mathcal{N}^{(1)}$, and another 2LP, denoted by $\mathcal{N}^{(2)}$. Both 2LP's have N binary input units and a single binary output unit, and they have, respectively, $K^{(1)}$ and $K^{(2)}$ binary hidden units. The weight vectors that connect the inputs to hidden unit l are denoted by $\vec{W}_l^{(i)}$ for the two networks $i=1,2$. We consider the case when the generalization error vanishes, $\epsilon_g(\mathcal{N}^{(1)},\mathcal{N}^{(2)}) = 0$. The following discussion establishes the possible relations between $K^{(1)}$ and $K^{(2)}$, and between $\vec{W}_l^{(1)}$ and $\vec{W}_l^{(2)}$, that hold[b] in the limit $N \gg K^{(i)} > 1$.

We claim that under these conditions only one of two possibilities can occur:

i) if $\mathcal{N}^{(2)}$ is regular and irreducible, its first layer weighs $\vec{W}_l^{(2)}$ are copies of those of $\mathcal{N}^{(1)}$, up to some symmetries mentioned above, and $\vec{w}^{(1)}$ and $\vec{w}^{(2)}$ are the weights of two equivalent perceptrons, and hence $K^{(1)} = K^{(2)}$.

ii) if $\mathcal{N}^{(2)}$ is not regular and irreducible, a subset of its hidden units form a regular irreducible 2LP, for which the previous statement holds; *i.e.* $K^{(1)} < K^{(2)}$.

4.1 Particular case

We show that if the first layer weights of two regular irreducible networks constitute a set of linearly independent vectors, they cannot implement the same $\{-1,1\}^N \rightarrow \{-1,1\}$ function, irrespectively of what the second layer weights are.

Let us consider the case of two irreducible networks, $\mathcal{N}^{(1)}$ and $\mathcal{N}^{(2)}$ where the set $\left\{\vec{W}_l^{(1)}, \vec{W}_m^{(2)}\right\}$ $(1 \le l \le K^{(1)};\ 1 \le m \le K^{(2)})$ is linearly independent.

Since eq. (6) holds, a finite fraction of the input space contributes to any set of locals fields. Therefore a non-vanishing fraction of the input space gives rise to

[b] By “hold in the limit $N \gg K^{(i)} > 1$” we mean that various statements, such as eq. (4), are correct to order $\frac{1}{N}$.

the simultaneous appearance of any pair of IR, $\vec{\sigma}^\eta$ and $\vec{\sigma}^\nu$, for $\mathcal{N}^{(1)}$ and $\mathcal{N}^{(2)}$;

$$\mathrm{P}(\vec{\sigma}^\nu;\ \vec{\sigma}^\eta) \ > \ 0 \qquad \text{for all} \quad \vec{\sigma}^\nu \in \{-1,1\}^{K^{(1)}} \ \text{ and } \ \vec{\sigma}^\eta \in \{-1,1\}^{K^{(2)}}$$

In particular, pairs of IR that produce different outputs on the two networks will appear. Such pairs of IR can always be found: if $\vec{\sigma}^\nu$ and $\vec{\sigma}^\eta$ produce the same outputs, clearly $\vec{\sigma}^\nu$ and $-\vec{\sigma}^\eta$ produce opposite ones. Therefore $\epsilon_g\Big(\mathcal{N}^{(1)};\mathcal{N}^{(2)}\Big) \neq 0$ and hence the two networks cannot implement the same mapping.

4.2 General case

Let us consider, without loss of generality, that $\Big\{\vec{W}_m^{(1)}, \vec{W}_l^{(2)}\Big\}$ $(1 \le m \le K^{(1)};\ 1 \le l \le M - K^{(1)})$ are linearly independent, and the remaining weight vectors of $\mathcal{N}^{(2)}$ can be expressed as a linear combination of these M, according to eq. (7). In Section 4.2.1 we prove that if $\epsilon_g\Big(\mathcal{N}^{(1)};\mathcal{N}^{(2)}\Big) = 0$, then two statements *(A)* and *(B)* hold as well:

(A) For all $m = 1, \ldots, K^{(1)}$ there exists at least one l in the range $M - K^{(1)} < l \le K^{(2)}$, such that $a_{lm} \neq 0$ (see eqs. (7) and (9)).

(B) For any m, there exists at least one l such that

$$h_l^{(2)} = a_{lm}/;h_m^{(1)}/;, \tag{10}$$

for any input vector $\vec{x}$.

From *(A)* and *(B)* we conclude that there is an l such that

$$\vec{W}_l^{(2)} = a_{lm}/;\vec{W}_m^{(1)}$$

for every m. Since the norm of the first layer weights is irrelevant, and since a negative a_{lm} can be compensated by a "gauge" operation discussed above, we have proved actually that for every $m \leq K^{(1)}$ there exist an l such that

$$\vec{W}_l^{(2)} = \vec{W}_m^{(1)}.$$

Hence the weights incident on $K^{(1)}$ (out of the $K^{(2)}$) units of $\mathcal{N}^{(2)}$ are copies of the first network's weights, $\vec{W}_m^{(1)}$. Therefore we have shown that in order to copy exactly the Boolean function implemented by a regular irreducible network we must use at least $K^{(1)}$ hidden units. Hence the *minimal* copying network must use, in order to map the hidden layer to the output, an equivalent perceptron to that of the original network. So the second layer weights of the copy must belong to the same class as the original, and we can conclude that two regular irreducible 2LP's that perform the same Boolean function must have the same number of hidden units and weights[c].

[c] Since K is finite, the weights from hidden layer to output may differ, as long as the same Boolean function is implemented.

4.2.1 Proof of statements A and B

Proof of *(A)*:

Assume that there exists an m^* such that $a_{lm^*} = 0$ for all l such that $M - K^{(1)} < l \leq K^{(2)}$. This means that $\vec{W}^{(1)}_{m^*}$ does not appear in the expansion of any of the weight vectors $W^{(2)}_l$. Hence all local fields $h^{(2)}_l$, that are generated in response to *any* input, can be expressed in the form eq. (9), so, $h^{(1)}_{m^*}$ does *not* appear on the right hand side. Since $\mathcal{N}^{(1)}$ is irreducible, there must be an IR, denoted by $\vec{s}$, such that

$$y^{(1)}(s_1, \ldots, s_{m^*}, \ldots, s_{\mathrm{K}^{(1)}}) = -y^{(1)}(s_1, \ldots, -s_{m^*}, \ldots, s_{\mathrm{K}^{(1)}}) . \tag{11}$$

The IR $\vec{s}$ appears when the local fields on the hidden units are $\left(h^{(1)}_1, ..., h^{(1)}_{m^*}, .. \right.$ $\left. .., h^{(1)}_{\mathrm{K}^{(1)}}\right)$. We can identify a finite fraction of the input space that give rise to any assignment of the local fields of $\mathcal{N}^{(1)}$ and the first $M - K^{(1)}$ of $\mathcal{N}^{(2)}$ (see eq. (8)). Hence there is a finite subset X_1 of input space that generates the local fields $\left(h^{(1)}_1, \ldots, h^{(1)}_{m^*}, \ldots, h^{(1)}_{\mathrm{K}^{(1)}}; h^{(2)}_1, \ldots, h^{(2)}_{\mathrm{M}-\mathrm{K}^{(1)}}\right)$ and another finite subset X_2, that generates $\left(h^{(1)}_1, \ldots, -h^{(1)}_{m^*}, \ldots, h^{(1)}_{\mathrm{K}^{(1)}}; h^{(2)}_1, \ldots, h^{(2)}_{\mathrm{M}-\mathrm{K}^{(1)}}\right)$. Clearly, when we switch from an input in X_1 to one in X_2, the sign of h_{m^*} changes, $\vec{s}_{m^*}$ changes and according to eq. (11) the output of $\mathcal{N}^{(1)}$ changes.

On the other hand, the IR of $\mathcal{N}^{(2)}$ is not altered; because changing of h_{m^*} to $-h_{m^*}$ does not affect the local fields on its linearly dependent hidden units, and the linearly independent ones remain fixed. This means that there is a finite subset of the input space (X_2), for which $\mathcal{N}^{(1)}$ and $\mathcal{N}^{(2)}$ give different answers, contradicting the hypotheses $\epsilon_g\Big(\mathcal{N}^{(1)};\mathcal{N}^{(2)}\Big) = 0$.

Proof of *(B)*:

Let us assume that *(B)* is wrong; *i.e.* there exists a unit $m^* \leq K^{(1)}$, such that there is no l^* for which eq. (10) holds. One possibility for this to happen would be that m^* does not appear in any of the expansions (7) or (9); this would, however, violate *(A)* and hence can be ruled out. Violation of *(B)* means that appearance of m^* in any expansion of the form eq. (9) is always accompanied by some other fields, *i.e.* for all l, $M - K^{(1)} < l \leq K^{(2)}$, if $a_{lm^*} \neq 0$ then

$$h_l^{(2)} \;=\; a_{lm^*} h_{m^*}^{(1)} + H_{lm^*} \tag{12}$$

where

$$H_{lm^*} \;=\; \sum_{m\neq m^*}^{K^{(1)}} a_{lm}\, h_m^{(1)} + \sum_{m=1}^{M-K^{(1)}} b_{lm}\, h_m^{(2)} \qquad\qquad (\text{not all } a_{lm}, b_{lm} = 0)$$

We introduce again the fact that any assignment of local fields $\Big(h_1^{(1)}, \ldots, h_{K^{(1)}}^{(1)};$ $h_1^{(2)}, \ldots, h_{M-K^{(1)}}^{(2)}\Big)$ on the M linearly independent hidden units will appear for a subset X^* of the inputs (that constitute a finite fraction of the total input space). In particular, we can chose X^* to contain those inputs for which (a) the output of $\mathcal{N}^{(1)}$ is positive, $y^{(1)} > 0$; (b) the corresponding IR is such that changing the sign of s_{m^*} causes change of the sign of $y^{(1)}$; (c) $H_{lm^*} \neq 0$ for $M - K^{(1)} < l \leq K^{(2)}$; and ($d$) the absolute value $\|h_{m^*}^{(1)}\|$ is smaller than $\min_{M\text{-}K^{(1)}<l\leq K^{(2)}} \left\{ \left\| \frac{H_{lm^*}}{a_{lm^*}} \right\| \right\}$[d]. Clearly,

[d] It follows from eq. (8) that a finite fraction of the input space satisfy this condition

there exists another set of inputs, $\widetilde{X^*}$, such that the field on unit m^* is reversed, while on all other $M-1$ units it remains the same. By our definition of X^*, we must then have $y^{(1)}\left(\vec{x} \in \widetilde{X^*}\right) = -y^{(1)}(\vec{x} \in X^*)$. On the other hand, from condition (d), we have

$$\text{sign}\left(a_{lm^*} h^{(1)}_{m^*} + H_{lm^*}\right) = \text{sign}\left(-a_{lm^*} h^{(1)}_{m^*} + H_{lm^*}\right) \qquad \text{for} \quad M - K^{(1)} < l \leq K^{(2)}$$

Therefore none of the hidden units of $\mathcal{N}^{(2)}$ is changed when we switch inputs from $\vec{x} \in X^*$ to some $\vec{x'} \in \widetilde{X^*}$, whereas $\mathcal{N}^{(1)}$ produces outputs of different sign. Thus we must have either $y^{(1)}(\vec{x}) = -y^{(2)}(\vec{x})$, or $y^{(1)}\left(\vec{x'}\right) = -y^{(2)}\left(\vec{x'}\right)$, and contradicting the assumption $\epsilon_g\left(\mathcal{N}^{(1)}; \mathcal{N}^{(2)}\right) = 0$.

5 Discussion

We analyzed the equivalence of two-layered perceptrons of binary units in the limit of large number of inputs. We found that if two irreducible regular 2LP's perform the same input-output mapping, they must have the same number of hidden units, and must be composed of equivalent perceptrons. In the limit of large number of inputs ($N \gg 1$) this imposes that the weights incident on the hidden units of these two networks must be pairwise equal, $\vec{W}_l^{(2)} = \vec{W}_l^{(1)}$, up to trivial symmetries (sign change and permutations). A direct consequence of this statement concerns restricted 2LP's (of fixed, non-equivalent second layer weights). If two restricted regular 2LP's employ different (not equivalent) perceptrons from hidden layer to output, they must implement completely distinct sets of Boolean functions for any assignment of $\vec{W}_l^{(2)}$ and $\vec{W}_m^{(1)}$. This result was obtained by Priel *et al.* (1994) for the case of non-overlapping receptive fields.

As an additional result, we found the joint distribution of the local fields (4) of two 2LP. In principle this expression, together with equation (1), can be used to calculate the generalization error of any two-layered machine. We also showed that if the first layer weight vectors of two machines are linearly independent, these networks should implement different Boolean functions, irrespectively of the mapping from the hidden units to the output unit.

Since equation (4) is valid for any set of perceptrons that receive simultaneously the same input, it can be used to compute the distribution of the local fields on the hidden units.

We conclude by stating two additional results implicit in our treatment. The first addresses the issue of what is meant by “different Boolean function”. We have in effect shown that the Boolean functions realized by two different 2LP's differ in the output that corresponds to a *finite fraction* of the 2^N possible inputs. The reason is that eq. (6) does not depend on N. The second point concerns finite size (N) effects. It can be shown that our results hold when N is large enough, such that $2^N >> N$.

References

F. Albertini and E.D.Sontag 1993, "For neural networks, function determines form", *Neural Networks*, **6**(7), 975–990.

M.L. Dertouzos 1964, "An approach to single-threshold-element synthesis", *IEEE Trans. on Electronic Computers*, **EC-13**, 519–528.

T. Grossman, R. Meir and E. Domany 1989, "Learning by Choice of Internal Representations", *Complex Systems*, **2**, 555–575.

G. Györgyi and N. Tishby (1990), "Statistical Theory of Learning a Rule", in *Neural Networks and Spin Glasses*, W.K. Theumann and R. Koeberle editors (Singapore: World Scientific).

R. Hecht-Nielsen 1990, "On the algebraic structure of feedforward network weight spaces", in *Advances Neural Computers*, 129–135, R. Eckmiller editor (Elsevier).

P.M. Lewis II and C.L. Coates 1967, *Threshold Logic*, John Wiley & Sons, New York.

M.G. Kendall, A. Stuart, and J.K. Ord 1987, *Kendall's advanced theory of statistics*, **1**, 484. Charles Griffin & Co., London.

V. Kurkova and P.C. Kainen 1994, "Functionally equivalent feedforward neural networks", *Neural Computation*, **6**(3), 543–558.

D. Nabutovsky, T. Grossman and E. Domany 1990, "Learning by CHIR without storing internal representations", *Complex Systems*, **4**(2), 519–541.

M. Opper, W. Kinzel, J. Kleinz and R. Nehl (1990), "On the Ability of the Optimal Perceptron to Generalize", *J. Phys. A: Math Gen* **23**, L581–L586.

A. Priel, M. Blatt, T. Grossman, E. Domany and I. Kanter 1994, "Computational capabilities of restricted two layered perceptrons", *Phys. Rev. E*, **50**(1), 577–590.

D.E. Rumelhart, G.E. Hinton and R.J. Williams 1986, "Learning representations by back-propagating errors", *Nature*, **323**, 533–536.

D. Saad 1994, "Explicit Symmetries and the capacity of multilayer neural networks", *J. Phys. A: Math Gen* **27**(8), 2719–2734.

H.S. Seung, H. Sompolinsky and N. Tishby (1990), "Statistical mechanics of learning from examples", *Phys. Rev.***A 45**, 6056–6091.

H. Sussman 1992, "Uniqueness of the weights for minimal feedforward nets with a given input-output map", *Neural Networks*, **5**(4), 589–593.