Methods for encoding in multilayer feed-forward neural networks

E. Elizalde, S. Gómez and A. Romeo

Dept. of Structure and Constituents of Matter,
Faculty of Physics, University of Barcelona,
Diagonal 647, 08028 Barcelona.

Abstract

Neural network techniques for encoding-decoding processes have been developed. The net we have devised can work like a memory retrieval system in the sense of Hopfield, Feinstein and Palmer. Its behaviour for 2^R ($R \in \mathrm{N}$) input units has some special interesting features. In particular, the accessibilities for each initial symbol may be explicitly computed. Although thermal noise may muddle the code, we show how it can statistically rid the result of unwanted sequences while maintaining the network accuracy within a given bound.

Introduction

The idea of using layered neural networks for multiple tasks has become more and more appealing since Rosenblatt's original perceptron model was object of the first serious criticism which led to far-reaching developments [1]. Among this type of structures, the most interesting group are the multilayer feed-forward nets containing intermediate —or *hidden*— layers.

The working of any net is determined by the relative strength of the links among units (neurons), usually given by a weight or connection exchange matrix. Typically, in a feed-forward multilayer network each unit computes a nonlinear function of the weighted sum of incoming signals from the previous layer reaching its own site, and sends the outcome on to the following layer. This process ends when the emerging signal arrives at the output units, where the result is read off.

Multilayer feed-forward neural networks are specially adequate for encoding [2], understood as the turning of p possible input patterns described by N digital units into a determined set of p output patterns on M units. In the minimal set-up, there is just one hidden layer of R hidden units forming a binary representation of the N inputs, *i.e.* with $R = \log_2 N$ neurons (Fig. 1). We will take $M = N = p$ in order to have the same number of neurons at the input and output layers. The first case to be considered is that in which we have sets —or alphabets— of *unary* patterns, *i.e.* binary sequences in which one unit is on and the rest are off:

$$\xi^\mu \equiv (\overset{1}{-}, \ldots, \overset{\mu-1}{-}, \overset{\mu}{+}, \overset{\mu+1}{-} \ \ldots, \overset{N}{-}), \quad \mu = 1, \ldots, N. \tag{1}$$

Afterwards, we shall consider arbitrary input and output patterns. When the number of output units has the form 2^R, with $R \in \mathrm{N}$, our model exhibits remarkable characteristics. In particular, no thresholds will be needed in this case. A no-go theorem on the possibility of general 3-step encoding will be proven. At the end, advantage will be taken of the introduction of a moderate amount of thermal noise.

For the intermediate and output layers, the state of each unit at a given moment will be a (generally nonlinear) function of the weighted sum of the signals feeding into it. Since we use binary units, we take the

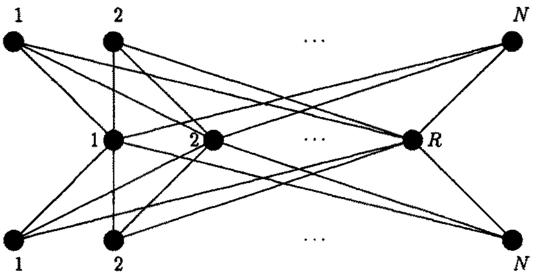

Figure 1: A 3-layer feed-forwrad.network consisting of input, intermediate and output layers.

activation function to be a sign. If the intermediate neurons are denoted by σ_j's, their excitation state will be given by

$$\left\{ \begin{array}{lcl} \sigma_j & = & \mathrm{sign}\ (h_j) \\ h_j & = & \sum\limits_{k=1}^{N} \omega_{jk}\zeta_k - \theta_j\,,\ j = 1, \ldots, R, \end{array} \right. \tag{2}$$

where ω_{jk} is the relative strength of the signal sent by the input unit k, whose state is ζ_k, into the intermediate jth site. The θ_j's are site-dependent thresholds.

The schemes

The form of connection matrix can be the result of either theoretical weight calculation *learning* [3]. Although the ability to learn is the most popular of all neural network features, our weights and thresholds will not come out from an iterative optimization process, but will be fixed beforehand. The choice fulfill requirements on memory economy in the sense that the size of the intermediate layers should be as small as possible. When the input and output sets are unary and besides they coincide, the optimal solution, *i.e.* one internal layer forming a binary representation, is possible. Then, the size of this hidden layer is

$$R = \left\{ \begin{array}{ll} \log_2 N & \text{if } \log_2 N \in \mathbf{N}, \\ [\log_2 N] + 1 & \text{if } \log_2 N \notin \mathbf{N}. \end{array} \right. \tag{3}$$

The scheme realizing the desired translation can be put as

$$\begin{array}{ccccc} \xi_k^\mu & \longrightarrow & \sigma_j^\mu & \longrightarrow & \xi_i^\nu \\ & \omega_{jk} & & \omega_{ij} & \\ & \theta_j & & \theta_i & \end{array}$$

where the weights and thresholds indicated enter the computation in the way

$$\begin{aligned} \sigma_j^\mu &= \mathrm{sign}\left(\sum_{k=1}^{N} \omega_{jk}\xi_k^\mu - \theta_j \right),\ j = 1, \ldots, R, \\ \xi_i^\nu &= \mathrm{sign}\left(\sum_{j=1}^{R} \omega_{ij}\sigma_j^\mu - \theta_i \right),\ i = 1, \ldots, N. \end{aligned} \tag{4}$$

By referring to the number of units in each layer, this scheme is called (N, R, N). In choosing the $\omega's$ and θ's, our method has taken advantage of the condition that the sequence $(\sigma_1, \ldots, \sigma_R)$ must reproduce the digits of the number $\mu - 1$ as a binary figure whenever ξ^μ is read. Further, in our choice of the weights we have found convenient to take $\omega_{jk} = \sigma_j^k$, *i.e.* the connection strength, and σ (considered as a matrix) having equal coefficients. If we choose the output set to be a permutation τ of the input set, *i.e.* $\nu = \tau(\mu)$, the weights and thresholds found are

$$\left\{ \begin{array}{lcll} \omega_{jk} & = & \sigma_j^k = (-1)^{\left[\frac{k-1}{2^{R-j}}\right]+1}, & j = 1, \ldots, R, \ k = 1, \ldots, N, \\ \theta_j & = & \displaystyle\sum_{k=1}^{N} (-1)^{\left[\frac{k-1}{2^{R-j}}\right]}, & j = 1, \ldots, R, \end{array} \right. \tag{5}$$

$$\left\{ \begin{array}{lcll} \omega_{ij} & = & (-1)^{\left[\frac{\tau^{-1}(i)-1}{2^{R-j}}\right]+1}, & i = 1, \ldots, N, \ j = 1, \ldots, R, \\ \theta_i & = & R-1, & i = 1, \ldots, N. \end{array} \right. \tag{6}$$

Once the ω's are chosen, the thresholds are not totally determined either. The above values have been selected mainly on the basis of simplicity.

The next step is the extension to sets of arbitrary (not necessarily unary) binary input and output patterns. An appropriate approach is to encode each input sequence into a specific unary pattern and, later, decode the resulting unary sequence to obtain the corresponding element of the arbitrary output set. This may be accomplished by the following enlargement of the previous structure

$$\begin{tikzcd} \zeta_l^\mu \arrow[r, "\substack{\omega_{kl} \\ \theta_k}"'] & \xi_k^\mu \arrow[r, "\substack{\omega_{jk} \\ \theta_j}"'] & \sigma_j^\mu \arrow[r, "\substack{\omega_{ij} \\ \theta_i}"'] & \xi_i^{\tau(\mu)} \arrow[r, "\substack{\omega_{hi} \\ \theta_h}"'] & S_h^{\tau(\mu)} \end{tikzcd}$$

The new weights and thresholds introduced are

$$\left\{ \begin{array}{lcl} \omega_{kl} & = & \zeta_l^k, \\ \theta_k & = & N-1, \end{array} \right. \tag{7}$$

$$\left\{ \begin{array}{lcl} \omega_{hi} & = & S_h^i, \\ \theta_k & = & -\displaystyle\sum_{\nu=1}^{N} S_h^\nu. \end{array} \right. \tag{8}$$

As one can check, they perform the above required translations. Although satisfactory, this solution takes up too many units. In fact, we have found it possible to work out the equivalent to the composition of the last three transformations, thus obtaining the (N, N, N) scheme:

$$\begin{tikzcd} \zeta_l^\mu \arrow[r, "\substack{\omega_{kl} \\ \theta_k}"'] & \xi_k^\mu \arrow[r, "\substack{\omega_{hk} \\ \theta_h}"'] & S_h^{\tau(\mu)} \end{tikzcd}$$

where the new elements are

$$\omega_{hk} = S_h^{\tau(k)}, \qquad \theta_h = -\sum_{\nu=1}^{N} S_h^\nu. \tag{9}$$

This solution is less wasteful as far as memory occupation is concerned, but is still away from the ideal minimum of just R intermediate units. Actually, we have been able to prove the following

Theorem: It is not possible to encode through the scheme

$$\begin{tikzcd} \zeta_l^\mu \arrow[r, "\substack{\omega_{jl} \\ \theta_j}"'] & \sigma_j^\mu \arrow[r, "\substack{\omega_{hj} \\ \theta_h}"'] & S_h^{\tau(\mu)} \end{tikzcd}$$

for arbitrary sets $\{\zeta_l^\mu\}$ and $\{S_h^{\tau(\mu)}\}$.

The proof proceeds by just showing cases where the scheme cannot work. Choosing a set containing linearly dependent ζ's, the weighted sums that constitute the inequations for the ω_{jl}'s lead to obvious contradictions.

Further, if one takes the output set such that a given component –say j– of the N output patterns forms the Boolean XOR —or generalized XOR (oddness)— function on the σ_j's, a similar contradiction comes up.

Therefore, the reasonable minimal way out will be to demand linear independence for the input set and to add an intermediate layer between the σ's and the output, as one has to do when trying to obtain the smallest implementation of the XOR function. Thus, we will have the structure

$$\begin{tikzcd} \zeta_l^\mu \arrow[r, "\substack{\omega_{jl} \\ \theta_j}"'] & \sigma_j^\mu \arrow[r, "\substack{\omega_{ij} \\ \theta_i}"'] & \xi_i^{\tau(\mu)} \arrow[r, "\substack{\omega_{hi} \\ \theta_h}"'] & S_h^{\tau(\mu)} \end{tikzcd},$$

where the only new elements are the ω_{jl}'s and the θ_j's. Being a bit demanding, we will require these thresholds to vanish. In order to find the ω_{jl}'s, we will temporarily reintroduce the unary-pattern layer between the input and the σ's, and figure out the composition of two successive affine transformations:

$$\begin{tikzcd} \zeta^\mu \arrow[r, "\xi = A\zeta + B"'] & \xi^\mu \arrow[r, "\sigma = C\xi + D"'] & \sigma^\mu \end{tikzcd}$$

$$\xi_k = \sum_l A_{kl}\zeta_l + B_k, \qquad \sigma_j = \sum_k C_{jk}\xi_k + D_j.$$

Then we find a solution for their coefficients, that reads

$$\left\{ \begin{array}{lcll} A_{kl} & = & 2(\zeta)^{-1}{}_{kl}, & k, l = 1, \ldots, N, \\ B_k & = & -1, & k = 1, \ldots, N, \end{array} \right. \tag{10}$$

where $(\zeta)^{-1}$ is the inverse of the matrix $(\zeta)_{l\mu} \equiv \zeta_l^\mu$, and

$$\left\{ \begin{array}{lcll} C_{j\mu} & = & \frac{1}{2}\sigma_j^\mu, & j = 1, \ldots, R,\ \mu = 1, \ldots, N, \\ D_j & = & \dfrac{1}{2}\displaystyle\sum_{\nu=1}^{N} \sigma_j^\nu, & j = 1, \ldots, R \end{array} \right. \tag{11}$$

When composing both by the rule

$$\sigma = C\xi + D = C(A\zeta + B) + D = CA\zeta + CB + D, \tag{12}$$

we find that $CB + D = 0$, and therefore it reduces to a linear tranformation as we wished. The weights that appear follow from

$$\sigma_j = \sum_l \omega_{jl}\zeta_l, \qquad \omega_{jl} = \sum_\nu \sigma_j^\nu (\zeta)^{-1}{}_{\nu l} = \sum_{\nu=1}^{N} (-1)^{[\frac{\nu-1}{2^{R-j}}]+1} (\zeta)^{-1}{}_{\nu l}. \tag{13}$$

Apart from the preceding, we have devised several schemes as variations of those already described. For instance, taking the five layer network (N, N, R, N, N), we have composed the two intermediate transformations, thus eliminating the σ layer at the expense of using more complicated weights and thresholds, the outcome being an (N, N, N, N) net and so on.

Accessibilities

The subject of the accessibilities comes up when one looks back at the initial unary-pattern three-layer permutator system and wonders what happens when the input pattern ξ is not any of the unary ξ^μ's. Now the *fields*

$$h_j = \sum_{k=1}^{N} \omega_{jk}\xi_k - \theta_j \tag{14}$$

may in fact vanish, as a result of which $\sigma_j = \text{sign}\ h_j$ is no longer well defined. Retaining the same sort of logistic function, *i.e.* a sign function, we can get out of this problem by redefining it so as to avoid the unwanted zeros.

Since this would cause a clumsy asymmetry, it is more reasonable to make room for zeros, with the additional difficulty that the σ's will now be three valued. A different option, that will be later discussed, is the use of strictly binary stochastic units at some finite temperature.

When zeros are allowed, 3^R possible sequences for the σ units exist. It is therefore appropriate to find what is the accessibility of each. The concept of accessibility of a configuration has been taken from Hopfield et al [5] and means the fraction of initial random states leading to the retrieval of that pattern. The difference is that now this is not an associative memory device but an encoding system, and we are interested in the rate of occurrence of *good* translations as well as in the extent of the proportion of code muddled by zeros. Let $A(\sigma)$ denote the accessibility of the intermediate pattern σ. By this we mean

$$A(\sigma) = \frac{\#\text{ possible } \xi' s \text{ leading to } \sigma}{\#\text{ possible } \xi' s}.$$

An advantage of our scheme is the special properties we have when $N = 2^R$, being R an integer. In all these cases one can check, from our expressions, that $\theta_j = 0$ for any j, *i.e.* no thresholds are needed. Furthermore, the connection strength matrix ω has a very interesting property: all its rows are mutually orthogonal, as are the R vectors formed with each binary digit of the numbers $0, 1, 2, \ldots, 2^R - 1$.

The calculation of the accessibilities starts by considering the possible values of h_j, which turn out to be $N, N-2, \ldots, 0, \ldots, -N+2, -N$. For any h_j within this range, let $f(h_j)$ be the number of possibilities that the jth component of h takes on this precise value, *i.e.* h_j. After some combinatorial brain-racking, we arrive at

$$f(h_j) = \binom{N}{\frac{N-h_j}{2}}, \tag{15}$$

and therefore

$$f(h_j = 0) = \binom{N}{\frac{N}{2}}, \qquad f(h_j \neq 0) = 2^N - \binom{N}{\frac{N}{2}}. \tag{16}$$

As we shall see, each accessibility will be determinable once the joint frequencies:

$$f(h_i = 0 \wedge h_j = 0),\ i \neq j,$$

$$f(h_i = 0 \wedge h_j = 0 \wedge h_k = 0), i, j, k \text{ different}$$

$$f(h_i = 0 \wedge h_j = 0 \wedge h_k = 0 \wedge h_l = 0), i, j, k, l \text{ different}$$

$$\vdots$$

are known. All these numbers can be computed, after some work, by consideration of the above mentioned properties of othogonality of the ω's, by turning sets of inequations in indetermined linear equation systems, looking at their general solutions and applying further combinatorial thinking, in which the form of the individual frequency(15) plays a crucial role. Following the process described elsewhere [4] we obtain

$$f_N(h_i = 0 \wedge h_j = 0) = \binom{\frac{N}{2}}{\frac{N}{4}}^2,$$

$$f(h_i = 0 \wedge h_j = 0 \wedge h_k = 0) = \sum_{k=0}^{N/4} \binom{\frac{N}{4}}{k}^4,$$

$$f(h_i = 0 \wedge h_j = 0 \wedge h_k = 0 \wedge h_l = 0) = \sum_{a=0}^{N/8} \sum_{b=0}^{N/8} \sum_{c=0}^{N/8} \sum_{d=0}^{N/8} \binom{\frac{N}{8}}{a} \binom{\frac{N}{8}}{b} \binom{\frac{N}{8}}{c} \binom{\frac{N}{8}}{d}$$

$$\cdot \binom{\frac{N}{8}}{2a + b + c + d - \frac{N}{4}} \binom{\frac{N}{8}}{\frac{N}{4} - (a + b + c)} \binom{\frac{N}{8}}{\frac{N}{4} - (a + b + d)} \binom{\frac{N}{8}}{\frac{N}{4} - (a + c + d)}.$$

$$\vdots \tag{17}$$

The binomial coefficients are to be understood in a generalized sense, *i.e.* when the number downstairs is negative or when the difference between upstairs and downstairs is a negative integer, they vanish. Otherwise we would have to explicitly state that the above multiple sum is restricted to a, b, c and d between the bounds and also satisfying

$$\left\{\begin{array}{lcccl} 0 & \leq & 2a+b+c+d-\frac{N}{4} & \leq & \frac{N}{8} \\ 0 & \leq & \frac{N}{4}-(a+b+c) & \leq & \frac{N}{8} \\ 0 & \leq & \frac{N}{4}-(a+b+d) & \leq & \frac{N}{8} \\ 0 & \leq & \frac{N}{4}-(a+c+d) & \leq & \frac{N}{8}. \end{array}\right. \tag{18}$$

We have called the above joint frequencies *orthogonalities*. Next let us see how the frequencies $f(h_1 \neq 0, \dots, h_j \neq 0), j = 0, \dots, R$ can be easily put in terms of the orthogonalities. This will almost immediately lead to the accessibilities. The sum of all the frequencies can be decomposed by considering each subset $\{k_1, \dots, k_j\}$ of j indices, $0 \leq j \leq R$, for which the corresponding field components vanish (the rest being nonzero):

$$\begin{aligned} 2^N &= \sum_{j=0}^{R} \sum_{\{k_1,\dots,k_j\}} f(h_1 \neq 0, \dots, h_{k_1} = 0, \dots, h_{k_j} = 0, \dots, h_R \neq 0), \\ &= \sum_{j=0}^{R} \binom{R}{j} f(h_1 = 0, \dots, h_j = 0, h_{j+1} \neq 0, \dots, h_R \neq 0), \end{aligned} \tag{19}$$

where we have observed that these frequencies depend on the number of null components, but not on their position. On separating the term $j = 0$, we get

$$f(h_1 \neq 0, \dots, h_R \neq 0) = 2^N - \sum_{j=1}^{R} \binom{R}{j} f(h_1 = 0, \dots, h_j = 0, h_{j+1} \neq 0, \dots, h_R \neq 0). \tag{20}$$

Similarly, we have

$$\begin{gathered} f(h_1 = 0, \dots, h_j = 0, h_{j+1} \neq 0, \dots, h_R \neq 0) = \\ f(h_1 = 0, \dots, h_j = 0) - \sum_{k=1}^{R-j} \binom{R-j}{k} f(h_1 = 0, \dots, h_{j+k} = 0, h_{j+k+1} \neq 0, \dots, h_R \neq 0). \end{gathered} \tag{21}$$

(20) and (21) constitute a system of interwoven recurrent equations, whose solution happens to be [4]

$$f(h_1 \neq 0, \dots, h_R \neq 0) = 2^N + \sum_{k=1}^{R} (-1)^k \binom{R}{k} f(h_1 = 0, \dots, h_k = 0). \tag{22}$$

Making the additional observation that for $N = 2^R, R \in \mathrm{N}$, all the nonspurious *—without from zeros—* σ's have the same frequency, we conclude that, for any of these sequences,

$$A(\sigma) = \frac{f(h_1 \neq 0, \dots, h_R \neq 0)}{2^N}, \tag{23}$$

which are now known.

Temperature-induced noise

The other alternative, namely the use of stochastic units, is based on the presence of thermal noise of 'temperature' $T = \frac{1}{\beta}$. The state of each σ neuron will no longer be determined by the field at its site, but will be a random binary function with probabilities

$$P(\sigma_j = \pm 1) = \frac{1}{1 + e^{\mp 2\beta h_j}}. \tag{24}$$

Therefore, no chance of $\sigma_j = 0$ exists. Input sequences that gave rise to spurious σ's will now yield good σ's. However, since the process happens to be on the whole stochastic the accessibilities do not have any meaning unless we consider them as averaged quantities over many repetitions of the encoding for all the possible inputs. Only then can we define

$$< A(\sigma^{\mu}) >= \frac{\text{cumulative \# input patterns which have given } \sigma^{\mu}}{\text{cumulative \# patterns read } (= \text{ \# repetitions } \cdot 2^{N})}$$

Simulations like the one represented in Fig. 2 show the tendency to equiaccessibility as the number of repetitions increases in the sense that

$$< A(\sigma^{\mu}) > \rightarrow \frac{1}{2^{R}}.$$

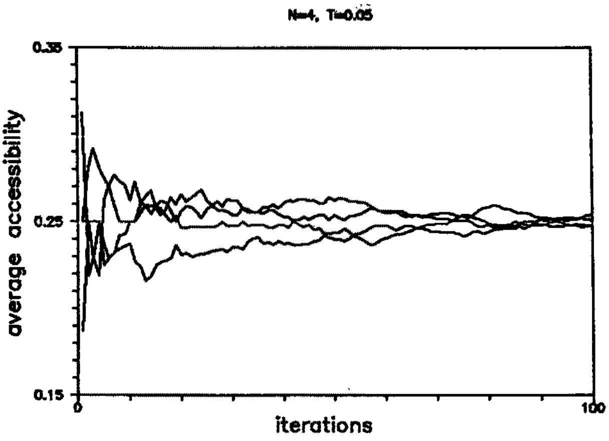

Figure 2: Result of a simulation for N=4 at finite T. The curves represent the cumulative average accessibilities of each ξ^{μ}.

Since the updating of all units is stochastic, $\xi = \xi^{\mu}$ does not necessarily mean that $\sigma_j = \sigma_j^{\mu}$, and only averages can be controlled. It would be desirable to have cases in which $< \sigma_j >_{\xi=\xi^{\mu}} = \sigma_j^{\mu}$ and that a critical temperature T_c existed so that for $T < T_c$ this type of preservation could be guaranteed. However, taking into account that $< \sigma_j >_{\xi=\xi^{\mu}} = \tanh(\beta h_j^{\mu})$ and that in our system $\sigma_j^{\mu} = \omega_{j\mu}$, the above conservation relationship reads

$$\omega_{j\mu} = \tanh(2\beta\omega_{j\mu}), \tag{25}$$

which has no solution for finite β's, *i.e.* no there is no critical temperature. However, if we content ourselves with preserving the average encoding just up to a certain accuracy, the same line of thinking allows us to find error bounds. If one requies that

$$| < \sigma_j >_{\xi=\xi^{\mu}} - \sigma_j^{\mu} | \leq \varepsilon, \tag{26}$$

it will suffice to take

$$\beta \geq \frac{1}{4} \log \frac{2-\varepsilon}{\varepsilon}. \tag{27}$$

e.g. , if we want our average values to be reliable up to the third decimal digit, taking $\varepsilon = 10^{-4}$ yields $\beta \geq 2.47$ or $T \leq 0.40$, which agrees quite fairly with the behaviour observed in our simulations.

Concluding remarks

Bringing in low temperature thermal noise changes the accessibilities of the non-spurious sequences with the result that what was zero-temperature equiaccessibility is maintained in the form of average equiaccessibility. Temperature increases do not change these averages. Even if at a given moment the system is presented with sequences that give rise to zero fields, it must now decide between one of the binary values, and, when making this choice, each increase of accessibility with respect to the zero-temperature set-up is roughly the same for each ξ^{μ}.

Our work carries on with the spirit of the ideas put forward by Hopfield, Feinstein and Palmer [5] in the sense that external control of one type or another can lead a retrieval system to work with equal accessibility by more or less smooth modifications of the initial working of the net.

These results are quite likely to find application in any type of encoding-decoding questions. Moreover, even though we have not dealt with other sorts of networks, *e.g.* associative memory systems, noise considerations are called for when interpreting processes of pattern identification in the context of problems such as image or trajectory reconstruction [6]-[7].

Additional developments of these ideas would surely appear after obtaining more amenable expressions for the accessibilities, particularly of their dependence on the number of initial patterns. Incorporating these solutions into learning methods (perhaps as initial weight configurations) may offer a new opportunity for improving the performance of any learning algorithm.

Acknowledgements

This work has been partly supported by Comisión Interministerial de Ciencia y Tecnología (CICYT) contract no. AEN 90-0033. SG acknowledges an FPI fellowship from the Spanish Ministry of Education and Science.

References

[1] Minsky, M. & Papert, S. *Perceptrons*, (MIT, Cambridge 1969).

[2] Rumelhart, D.E., McClelland, J.L., and the PDP research group, ***Parallel Distributed Processing: Explorations in the Microstructure of cognition. Vols 1 and 2.***, (MIT, Cambridge, 1986).

[3] Rumelhart, D.E., Hinton, G.E. & Williams, R.J., *Nature* **323** 533-536 (1986).

[4] Elizalde, E., Gómez, S. and Romeo, A., to be published (1991).

[5] Hopfield, J.J., Feinstein, D.I. & Palmer, R.G. *Nature* **304** 158-159 (1983).

[6] Denby, B. , Campbell, M. , Bedeschi, F., Chris, N., Bowers, C.& Nesti, F., to be published.

[7] Stimpfl-Abele, G. & Garrido, L. *Computer Phys. Commun.*, to appear (1990).