

NUMERICAL ANALYSIS OF A NEURAL NETWORK WITH HIERARCHICALLY ORGANIZED PATTERNS

S. Bacci[#], J. Alfaro^{*}, G. Wiecko⁺ and N. Parga

Centro Atómico Bariloche

8400 Bariloche, Rio Negro, Argentina

Abstract

A numerical analysis of the retrieval behavior of an associative memory model where the memorized patterns are stored hierarchically is performed. It is found that the model is able to categorize errors. For a finite number of categories these are retrieved correctly even when the stored patterns are not. Instead when they are allowed to increase with the number of neurons their retrieval quality deteriorates above a critical category capacity.

PACS No.: 87.30G - 64.60C - 75.10H - 8970

[#]Fellowship granted by the Comisión Nacional de Energía Atómica, Argentina.

^{*}Permanent address: Facultad de Física, Universidad Católica de Chile, Santiago, Chile.

⁺Member of the Carrera del Investigador, CONICET, Argentina.

1. Introduction

Recent studies of neuronal networks have centered around simple models which attempt to explain how these systems work as an associative memory. According to McCulloch and Pitts [1] each neuron i can be in one of two possible states, firing or quiescent. It can then be described by an Ising-like variable $S_i = \pm 1$ ($i = 1, \dots, N$). The interaction between neurons i and j is mediated by the synaptic strengths T_{ij} which are used to store the information. The dynamics of the N neuron system is fully deterministic and, in the case an energy E can be defined, it relaxes to states which are its local minima.

For the simplest neural network model [2] E is given by

$$E = - \sum_{(i,j)} T_{ij} S_i S_j \quad (1.1)$$

where the sum extends over all pairs of neurons and the T_{ij} are given by the Hebb's rule [3]. Amit et al. [4] have shown that in the saturation regime, where the number p of memories stored in the T_{ij} 's grows linearly with N ($p = \alpha N$), the network retrieves information properly for $\alpha < \alpha_c \simeq 0.145$. At α_c the quality of the retrieval undergoes a discontinuous transition to a regime where the network is always confused. This means that even when the input state is one of the memorized patterns the system evolves to spurious states which have small overlaps with it.

As noticed elsewhere [5] an unrealistic feature of the Hopfield model is that the patterns have to be encoded in orthogonal N bit words. However, this is not the way human memory

stores new information; in general, we classify it by looking for correlations with the already memorized patterns. As a result of the classification process the patterns will appear categorized according to their common features. Instead of being orthogonal as in the Hopfield model (Fig.1a) they will be organized following a hierarchical structure given, for instance, by the tree of Fig. 1b.

Another motivation to study models with patterns classified according to categories is the expectation that also the errors in retrieval will exhibit the hierarchy. In the Hopfield model, when α exceeds its critical value α_c the system is attracted to configurations with small overlaps with the stored patterns, a fact which indicates the deterioration of the network as an associative memory. What we want is a model able to retrieve well the average features of the set of patterns which belong to the same category. In this way, even when the attractors are spurious states with finite overlaps with several memorized patterns, these will belong to the same class and there will not be errors in category retrieval.

In this paper we perform a numerical analysis of the retrieval properties of both stored patterns and categories in the plane (α, γ) , with γ the ratio between the number of categories and the number of neurons. This study reveals the existence of a region in that plane where the network confuses the patterns but it is still able to retrieve categories correctly. This occurs on the axis $\gamma = 0$ but also for slightly larger values of γ .

Classification errors, i.e. a misclassification of the

patterns in categories, become important as the number of categories increases. Since it is not reasonable to ask the system to improve its input information we will not consider these errors as a bad performance of the network. Nevertheless, there also exists a region in the plane (α, γ) where both patterns and categories are confused, i.e. there are errors which go beyond the classification ones. This occurs for large α and γ .

This work is organized as follows: in Section 2 we present the model. In Section 3 we define several probability distributions which will be useful to analyse the pattern and category retrieval. The numerical analysis and the discussion of the different regions in the plane (α, γ) are given in Section 4. Finally in Section 5 we summarize our main conclusions.

2. The model

The first stage of the definition of the model concerns the generation of patterns with a hierarchical organization. Assuming that categorization is previous to learning the memories to be stored in the system are generated according to a branched stochastic process defined on the associated hierarchical tree. The state of each neuron is determined simultaneously for the whole set of the p patterns and independently of the state of the other neurons. At the upper level of categorization all of them constitute a single category. As the stochastic process proceeds further the coarsest differences are put into evidence first and a few categories appear.

From that point on, the processes for the different categories are uncorrelated. At the lowest level of categorization the patterns are distinguished only by their finest details.

For simplicity we shall restrict the discussion to regular trees such as the one of Fig. 1b. The generalization to more complex ones is straightforward.

In particular we shall consider the simplest case of two level regular trees. This is enough for our purpose of analysing the performance of a neural network when categories are present. The parameters of the tree are the number of categories N_a and the number of patterns r stored in each of them. The total number of states in the tree is $p = N_a r$. We shall frequently name the categories as ancestors and the patterns inside them as their descendents.

The stochastic process begins at the first level where N_a variables $y_1^{(\alpha)}$ ($\alpha = 1, \dots, N_a$) are chosen by using a probability distribution $P_1(y_1 | y_0)$ with mean value y_0 . At the second level r variables $y_2^{(\alpha, \beta)}$ ($\beta = 1, \dots, r$) are selected using $P_2(y_2 | y_1^{(\alpha)})$ for each ancestor α . The state of neuron i in the pattern $\{S^{(\alpha, \beta)}\}$ is obtained by

$$S_i^{(\alpha, \beta)} = \text{sign } y_2^{(\alpha, \beta)} \quad \begin{array}{l} \alpha = 1, \dots, N_a \\ \beta = 1, \dots, r \end{array} \quad (2.1)$$

where α denotes the category and β the memory within it.

The next step is to store the p patterns in the synaptic connections. It is clear that the Hebb's rule does not work for correlated patterns. For hierarchically correlated patterns those in the same category have a finite overlap $\bar{q}$. This spoils the performance of the standard Hopfield model. To see the effect of this overlap let us consider the expression

$$\mathfrak{Z}_{ij}^{(H)} = \frac{1}{N} \sum_{\alpha=1}^{N_a} \sum_{\beta=1}^r S_i^{(\alpha, \beta)} S_j^{(\alpha, \beta)}; \quad (2.2)$$

then the action potential at neuron i

$$h_i = \sum_{j=1}^N \mathfrak{Z}_{ij}^{(H)} S_j \quad (2.3)$$

evaluated for the state $\{S^{(\alpha, \beta)}\}$ is

$$\begin{aligned}
h_i^{(\alpha_o \beta_o)} &= S_i^{(\alpha_o \beta_o)} + \bar{q} \sum_{\beta \neq \beta_o}^r S_i^{(\alpha_o, \beta)} \\
&= (1-\bar{q}) S_i^{(\alpha_o \beta_o)} + \bar{q} r S_i^{(\alpha_o)}, \quad (2.4)
\end{aligned}$$

up to the contribution of categories other than $\{S^{(\alpha_o)}\}$. Here we have defined the state of the ancestor α , $\{S^{(\alpha)}\}$, as

$$S_i^{(\alpha)} = \frac{1}{r} \sum_{\beta=1}^r S_i^{(\alpha, \beta)} \quad (2.5)$$

Even when the first term of eq. (2.4) is of the right form the second one spoils the performance of the system. But eq. (2.4) suggests the way to modify Hebb's rule; instead of (2.2) we propose

$$\mathfrak{J}_{ij} = \frac{1}{N} \left[\sum_{\alpha=1}^{N_a} \sum_{\beta=1}^r S_i^{(\alpha, \beta)} S_j^{(\alpha, \beta)} - \lambda \sum_{\alpha=1}^{N_a} S_i^{(\alpha)} S_j^{(\alpha)} \right] \quad (2.6)$$

where an ancestor contribution was added to the standard form. The parameter λ is chosen in such a way that the action potential is proportional to the memorized pattern itself. A trivial algebra yields

$$\lambda = r^2 \bar{q} / (1 + \bar{q} (r-1)) \quad (2.7)$$

For $r \gg 1$ eq. (2.6) is equivalent to the learning rule proposed in ref. [5] inspired on the microstructure of

ultrametricity in the mean field theory of spin glasses [6]. It is also contained in a more general model based on the pseudoinverse rule [7,8].

3. Analysis of pattern and ancestor retrieval

To study the retrieval properties of the model described in the previous section we shall use a set of probability distributions of the overlaps between the attractors and the stored patterns and their ancestors.

Starting from one of the memorized patterns, say $\{S^{(\alpha_0 \beta_0)}\}$, the system will evolve to an attractor $\{S\}$. Its overlap with the stored states is

$$m^{(\alpha/\beta)} = \frac{1}{N} \sum_{j=1}^N S_j^{(\alpha/\beta)} S_j; \quad (3.1)$$

in particular we will need $m = m^{(\alpha_0 \beta_0)}$ and its distribution $P(m)$. Similarly the overlap between the output and the ancestor is

$$m^{(\alpha)} = \frac{1}{N} \sum_{j=1}^N S_j^{(\alpha)} S_j \quad \alpha = 1, \dots, N_a \quad (3.2)$$

For the case that α is the initial category, we will consider the distribution $P_0(m^{(\alpha_0)})$.

We will be interested in the largest overlap M between the attractor and the categories other than the initial one,

$$M = \max (m^{(\alpha)}) \quad \alpha \neq \alpha_0 \quad (3.3)$$

and its associated distribution $P_a(M)$.

When the number of categories increases the patterns generated with the process described in Section 2 tend to have

small overlaps with categories different from their own. In this case even when the model works well we expect that the attractors will also have small overlaps with those categories. Since the network cannot learn more than what it was taught we will not consider errors due to misclassification as true ones.

In order to identify classification errors, we consider the distribution of the overlaps between the memorized pattern taken as the initial state and the categories

$$m^{(\alpha, \alpha_o \beta_o)} = \frac{1}{N} \sum_{j=1}^N S_i^{(\alpha)} S_i^{(\alpha_o \beta_o)} \quad (3.4)$$

The state $\{S^{(\alpha_o \beta_o)}\}$ has an average initial overlap with its own ancestor given by

$$\begin{aligned} m_1^{(\alpha_o)} &= m^{(\alpha_o, \alpha_o \beta_o)} \\ &= \frac{1}{r} [1 + \bar{q} (r-1)] \end{aligned} \quad (3.5)$$

We will also need the quantity

$$M^* = \max (m^{(\alpha, \alpha_o \beta_o)}) \quad \alpha \neq \alpha_o \quad (3.6)$$

This will be useful to separate classification errors from true errors by comparing the distributions for M and M^* .

The distributions P , P_o and P_a associated with the overlaps m , $m^{(\alpha_o)}$ and M respectively allow us to distinguish the possible attractors. We are interested in determining the behavior of the

network in the plane (α, γ) where

$$\alpha = p/N \quad (3.7)$$

$$\gamma = N_a/N$$

We can expect three regions:

- A) a region for α and γ small where both ancestors and their descendants work well. The output either coincides with the input or eventually it is a state very close to it. Here

$$m \sim 1$$

$$m^{(\alpha_0)} \simeq \frac{1}{r} [1 + \bar{q} (r-1)] \quad (3.8a)$$

$$M \sim 0$$

or, if the classification errors are present

$$M \sim M^* \quad (3.8b)$$

- B) Here, the attractor still belongs to the right category but it is a spurious state which has overlaps with many memorized patterns. These, however, belong to the same category as the initial state. This means

$$m \ll 1$$

$$m^{(\alpha_0)} < m_I^{(\alpha_0)} \quad (3.9)$$

$$M \sim 0$$

besides $m^{(\alpha_0)} \gg M$.

Again, if classification errors are appreciable we should ask $M \sim M^*$.

C) This is a regime where both memories and categories are confused. In this case

$$\begin{aligned} m &\ll 1 \\ m^{(\alpha_o)} &\ll m_I^{(\alpha_o)} \\ M &\sim m^{(\alpha_o)} \end{aligned} \quad (3.10a)$$

Contrary to what happens in regions A and B, in this case the category retrieval errors are always greater than the classification errors

$$M > M^* \quad (3.10b)$$

4. Numerical analysis

In this section we perform a numerical analysis in order to determine in which regions of the plane (α, γ) the three possible retrieval behaviors of the network appear. In particular we want to check that for a finite number of ancestors (i.e. $\gamma = 0$) the categories are correctly retrieved, even when the patterns themselves work badly.

The typical calculation was done by taking one of the memorized patterns as the initial state. Once an attractor was reached we evaluated the different overlaps defined in Section 3. Repeating the process for several networks we obtained the corresponding distributions. Since we were interested in the large N limit the simulation was done for several values of N . Looking at what type of attractors tend to dominate for increasing N at fixed (α, γ) , we determined to which of the three regions A, B or C that point belongs.

Once α, γ and N are given the geometric structure of the tree is determined. Apart from that we also need to define the distributions used in the branched stochastic process: $P_1(y_1 | y_0)$ and $P_2(y_2 | y_1)$. We took them as Gaussians, with $y_0 = 0$ and we chose their dispersions such that $\bar{q} \approx 0.5$.

We show schematically in Fig. 2 where the regions A, B and C are located. Let us notice that since there must be at least two states in each category, the line $\alpha = 2\gamma$ sets a limit to the relevant region for this model.

We found a line $\alpha_c(\gamma)$ which separates the region where the stored patterns (and consequently the ancestors) are retrieved

well (region A) from the other two regions (B and C). For $\alpha > \alpha_c$ the attractors have a small overlap m with the initial pattern. The change from $m \sim 1$ to a smaller value is discontinuous. For $\gamma = 0$ the system destabilizes at $\alpha_c(0) \simeq 0.11$ while for larger values of γ , $\alpha_c(\gamma)$ increases rapidly to about 0.15.

The line $\alpha_c(\gamma)$ was determined by observation of the behavior of the distribution $P(m)$ as N increases. Not too above $\alpha_c(\gamma)$ this distribution appears peaked at $m \sim 1$ for small N . As N increases this peak tends to decrease while another one at a smaller value of m appears. This is similar to the result reported in Ref. [4] for the Hopfield model. The same behavior is seen for all values of γ . In Fig. 3 we show $P(m)$ for the two largest values of N we have used and $\gamma = 0$, $\alpha = 0.16$. At this point the peak at $m \sim 1$ has disappeared and the peak at smaller values of m grows with N .

On the line $\gamma = 0$ the number of categories is kept fixed as the network is studied for different values of N . In order to determine whether there is a transition to a region with bad ancestor retrieval we evaluated the distributions $P_o(m^{(\alpha_o)})$ and $P_a(M)$ corresponding to the overlaps with the initial category and with the others respectively. We found good performance of the categories up to the largest value of $\alpha = 0.8$ we considered on the axis $\gamma = 0$. For $\alpha < \alpha_c(0)$ the stored patterns are retrieved correctly and clearly the same is true for the ancestors. The distribution P_o yields an average overlap $\bar{m}^{(\alpha_o)}$ in agreement with eq (3.5). For $\alpha > \alpha_c(0)$ the attractor has finite overlaps with all the memorized patterns in the initial category. This modifies the average value of $m^{(\alpha_o)}$ which becomes smaller than $m_1^{(\alpha_o)}$.

Nevertheless, the average overlap $\bar{M}$ of the output with the other ancestors remains very small and of the same order as the classification errors (about 5×10^{-2}). This indicates a good category retrieval. The results for the average overlaps are shown in Fig. 4. The data for $\bar{m}^{(\alpha_0)}$ still show a small dependence with N but they are consistent with a continuous transition to a region with bad ancestor retrieval at large, perhaps infinite, α . This is in agreement with the expectation that for a finite number of ancestors there is categorization of errors. Although not relevant for category retrieval, we observed that for large α the attractor has finite overlap with a small fraction (about 5%) of patterns inside the other category (for $\gamma = 0$ we took $N_a = 2$).

On the other side of the diagram, close to the line $\alpha = 2\gamma$ the categories behave quite differently. Similarly to what occurs for memory retrieval also the overlap between the attractor and the initial category jumps discontinuously to a smaller value. Simultaneously $\bar{M}$ becomes greater than the classification errors and acquires a value comparable to that of $\bar{m}^{(\alpha_0)}$. This is shown in Fig. 5 for the line $\alpha = 2\gamma$, but a similar behavior occurs for smaller values of γ defining a line $\gamma_c(\alpha)$. We cannot determine however, whether these two lines, α_c and γ_c , coincide close to $\alpha = 2\gamma$. Here the number of memories in each category is small and a change in memory retrieval might induce a change in ancestor retrieval as well. Besides, categories containing only two patterns are very close to being simple memories in the standard Hopfield model. This is in agreement with our rough estimation of γ_c on the line $\alpha = 2\gamma$ which is compatible with $\gamma_c \sim 0.14-0.16$.

In Figs. 6(a,b) we have included the distributions $P_o(m^{(\alpha_o)})$ and $P_a(M)$ for $N = 1000$ and two values of α which are at different sides of the critical lines ($\alpha = 0.14$ and 0.18 respectively). In Fig. 7(a,b) we present the data for $P(m)$ for the same values of α and r . From these figures it is clear that $\alpha = 0.14$ is in region A, while $\alpha = 0.18$ is in C. As we said before we cannot rule out completely the existence of a region like B in between; this has been indicated by the dashed line in Fig. 2.

From our previous discussion we see that as r increases at fixed but large α the network goes from a region of good category retrieval to the region C just described. We determined the way this occurs by observing the distributions $P_o(m^{(\alpha_o)})$ and $P_a(M)$. While for $r = 0$ these two distributions do not overlap appreciably, as r increases they shift to a common average. This is shown in Figs. 8(a,b) for $\alpha = 0.32$ and the two extreme values of r .

The numerical simulation in region A was done analysing 20 different points on the plane (α, r) . For each of them we took about 150 statistical independent measurements corresponding to different samples. In region B we considered 15 points each one evaluated over 50-150 different samples. Finally in region C we simulated 20 points each of them averaged over a number of networks ranging from 20 to 50.

The numerical simulations have been done with a Micro VAX II at Centro Atómico Bariloche and a VAX 8650 at the University of Rome.

5. Conclusions

It is important for a living system to be able to recognize at least a few concepts without errors, even when the finest details of the information are not retrieved well. We have shown that this is the case for models of associative memories where the stored patterns are organized hierarchically.

We performed our analysis for the simplest regular hierarchical tree where only two levels of categorization are present. It would be interesting however to see how the system work when more levels are included. In this case we think that when the number of categories and subcategories are kept fixed the their retrieval will still remain good even for large α .

Acknowledgments

Two of us (CW and NP) would like to thank the International Centre for Theoretical Physics (Trieste) for hospitality. NP wishes to thank the Dipartimento di Fisica of the Universita di Roma for the use of their computer in the first stages of this work. JA was partially supported by DIUC 201/86.

References

- 1) McCulloch, W.S. and Pitts, W., Bull. Math. Biophys. 5 (1943), 115.
- 2) Hopfield, J.J., Proc. Natl. Acad. Sci., U.S.A. 79 (1982), 2554; Proc. Natl. Acad. Sci., U.S.A. 81 (1984), 3088.
- 3) Hebb, D.O., The Organization of Behavior (New York: Wiley, 1949).
- 4) Amit, D.J., Grutfreund, H., and Sompolinsky, H., Phys. Rev. A 32 (1985), 1007; Phys. Rev. Lett. 55 (1985), 1530; Ann. Phys. 173 (1987), 30.
- 5) Parga, N. and Virasoro, M.A., J. Physique 47 (1986), 1857.
- 6) Mezard, M. and Virasoro, M.A., J. Physique 46 (1985), 1293.
- 7) Personnaz, L., Guyon, I. and Dreyfus, G., J. Physique Lett. 46 (1985) L359.
- 8) Cortes, C. Krogh, A., and Hertz, J.A., J. Phys. A.

Figure Captions

Fig. 1a) A tree with orthogonal patterns.

1b) A regular tree with two levels. In this example $N_a = 3$ and $r = 2$.

Fig. 2) The retrieval behavior in the plane (α, r) : we indicate the three regions, A, B, and C, described in the text. Within our numerical accuracy we cannot determine if the dashed region exists or not.

Fig. 3) The distribution $P(m)$ for $\alpha = 0.16$, $r = 0$ and $N = 500$, 1000, the empty histogram corresponds to $N = 500$, the dashed one to $N = 1000$. The double dashed region shows where the two distributions overlap.

Fig. 4) The average overlap $\bar{m}^{(\alpha_0)}$ of the distribution $P_o(m^{(\alpha_0)})$ at $r = 0$ with $N = 500$ (o) and $N = 1000$ (●).

Fig. 5) The average overlaps of the distributions $P_o(m^{(\alpha_0)})$ (○) and $P_a(M)$ (●) for $N = 1000$ at the line $\alpha = 2r$. The dashed line is the average overlap of the distribution of M^* (i.e. the effect of classification errors).

Fig. 6) The distributions $P_o(m^{(\alpha_0)})$ and $P_a(M)$ with $N = 1000$; and $r = \frac{\alpha}{2}$. $P_o(m^{(\alpha_0)})$ is indicated with the empty histogram, $P_a(M)$ with the dashed one. These distributions overlap in the double dashed region. a) corresponds to $\alpha = 0.14$, b) $\alpha = 0.18$.

Fig. 7a) The distribution $P(m)$ for $\alpha = 0.14$ and $r = \alpha/2$. $N = 1000$.

7b) The same distribution for $\alpha = 0.18$, $r = \alpha/2$.

Fig. 8) The distributions $P_o(m_o^{(\alpha_o)})$ and $P_a(M)$ with $N = 1000$. The empty histogram corresponds to the distribution $P_o(m_o^{(\alpha_o)})$, the dashed one to $P_a(M)$, and the double dashed region indicates where they overlap. a) $\alpha = 0.32$, $r = 0$. b) $\alpha = 0.32$, $r = 0.16$.



Fig. 1a

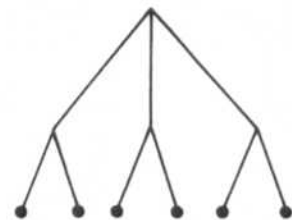


Fig. 1b

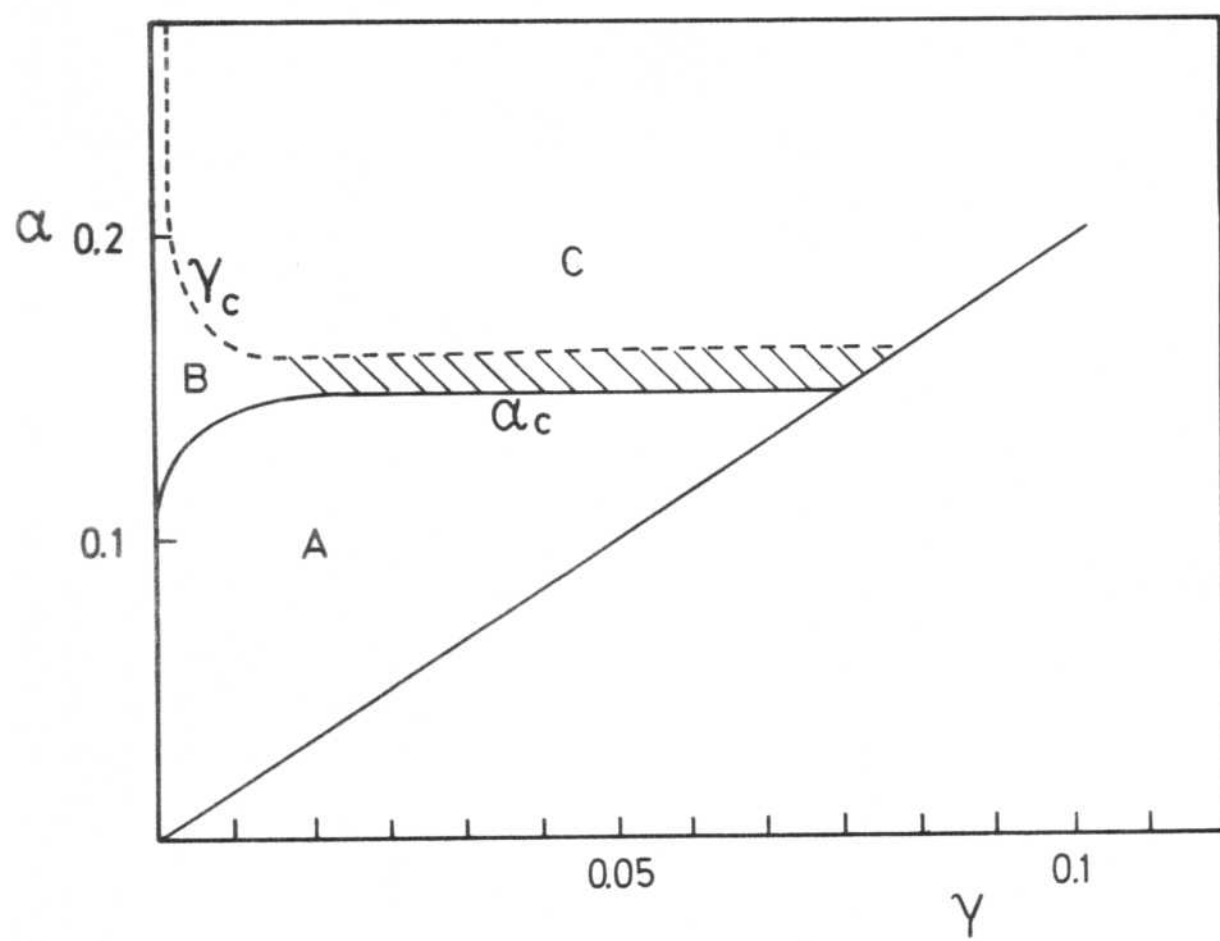
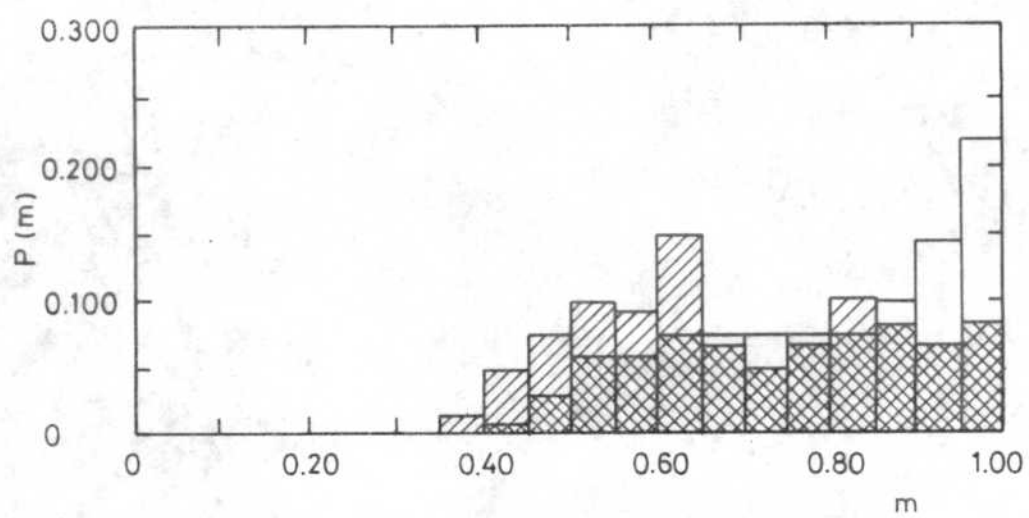


Fig. 2



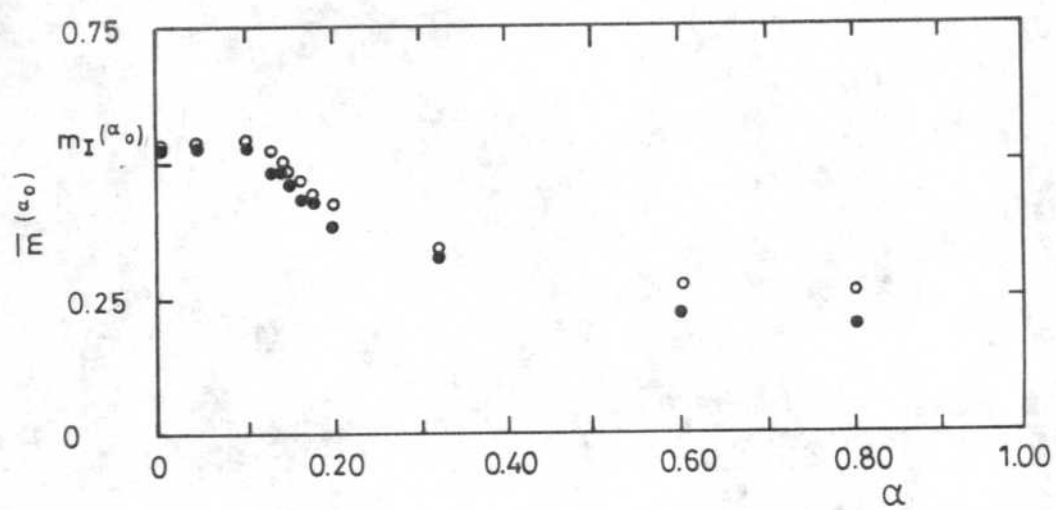


Fig. 4

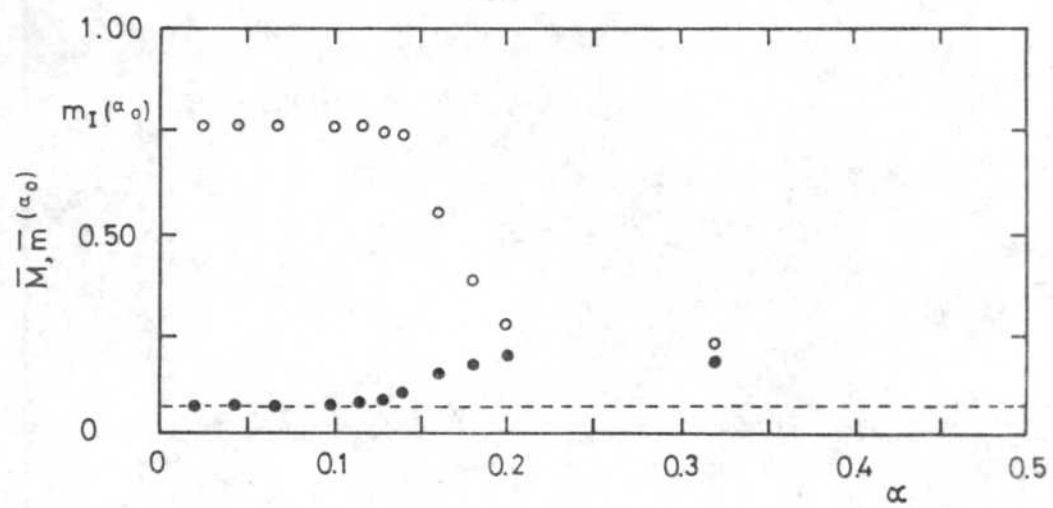


Fig. 5

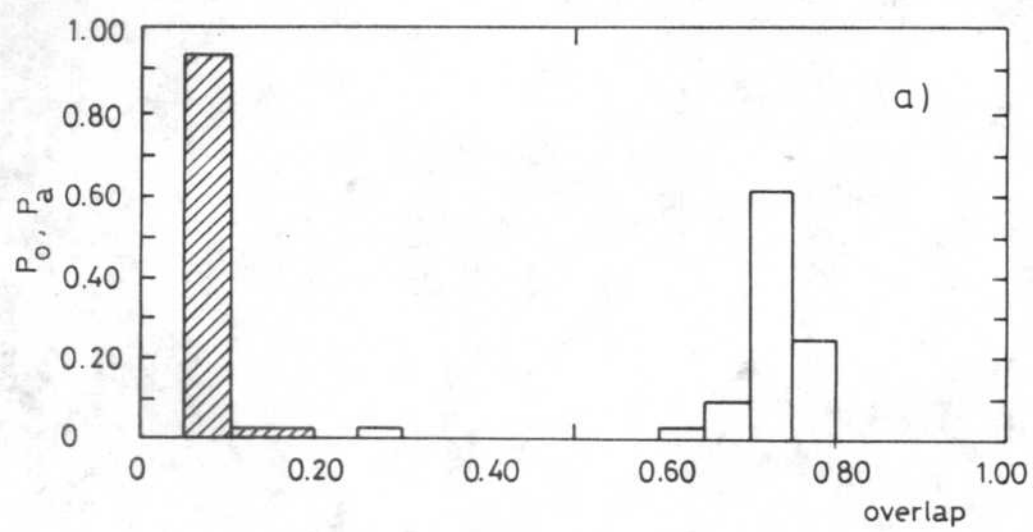


Fig. 6

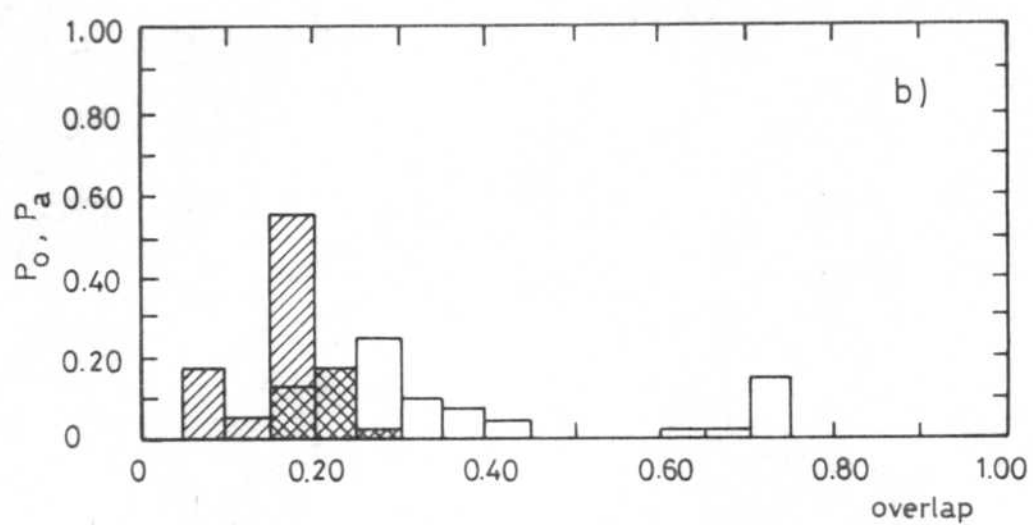


fig. b

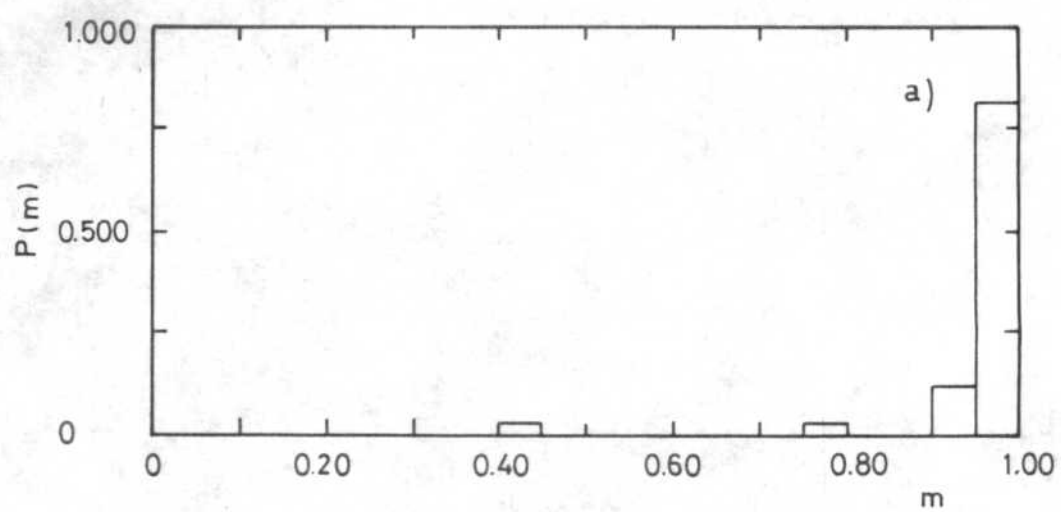


Fig 7.2

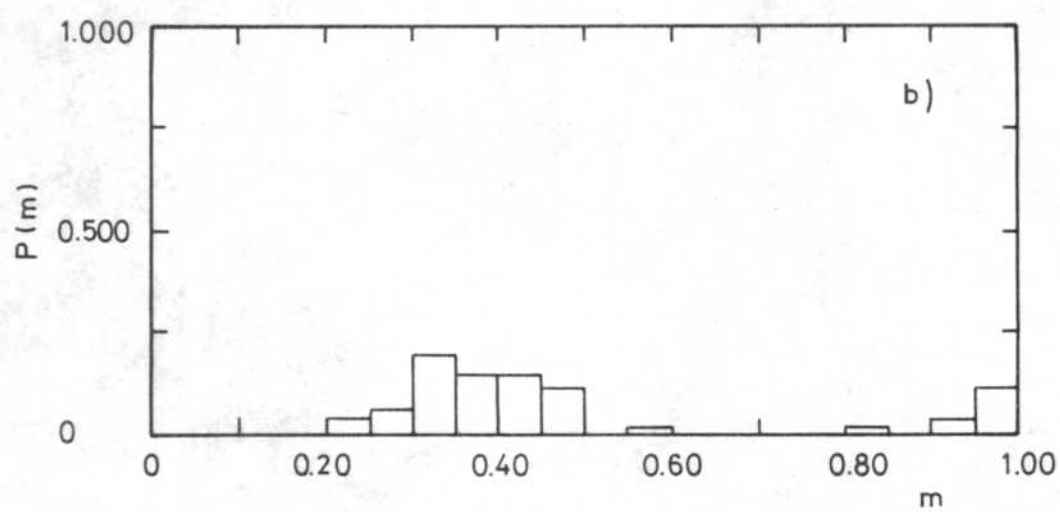


Fig. 7b

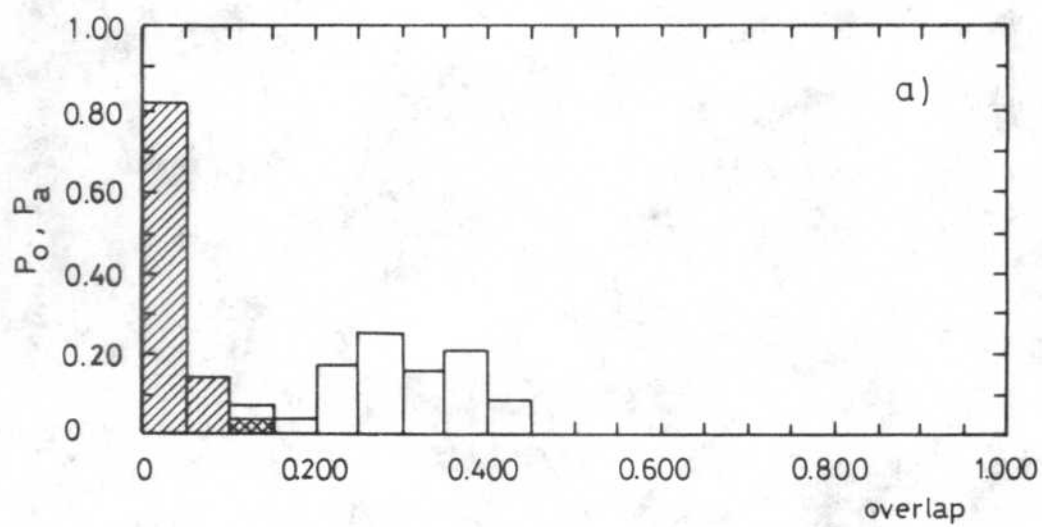
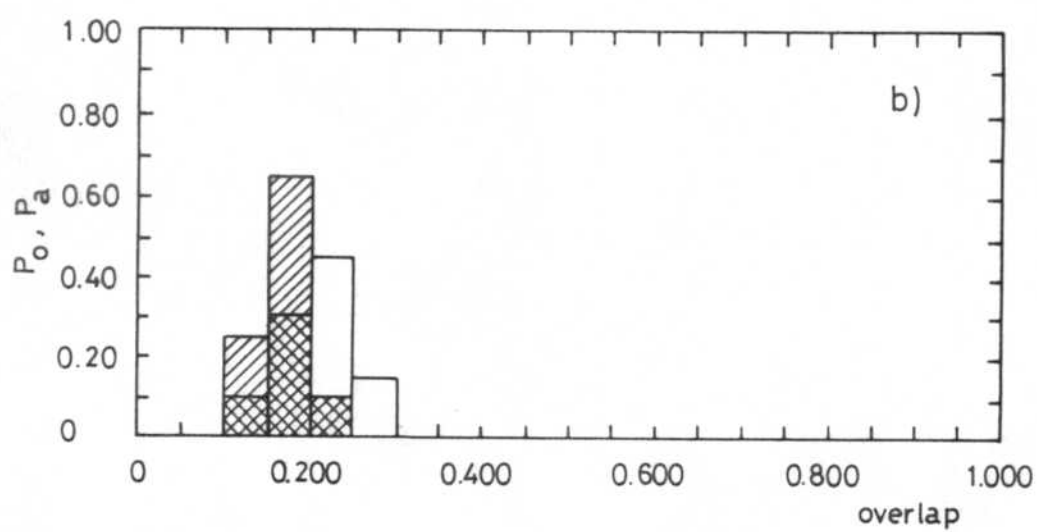


Fig. 8



1 JUN. 1988

SEÑOR DIRECTOR DEL CENTRO ATOMICO BARILOCHE:

De acuerdo con las normas establecidas para la publicación de trabajos (BAP 96/78), elevo para su consideración el trabajo titulado "Numerical analysis of a neural network with hierarchically organized patterns".

S. Bacci, J. Alfaro, C. Wiecko y N. Parga
producido por

Le Journal de Physique
el mismo será publicado en la revista

gratuito
El costo de la publicación es

INDEC. DE INVEST. Y D.E. CENTRO ATOMICO BARILOCHE ADMINISTRACION	
1 JUN. 1988	
441.300-63/88	
Salto: 1 JUN. 1988	


Luis Masperi
Jefe División Teoría

I) DIRECCION

S.C. DE BARILOCHE,

6 JUN 1988

Vista la solicitud de publicación presentada por el
Sr. Jefe de la División **Teoría**
del trabajo **"Numerical analysis of a neural network with
hierarchically organized patterns"**.

producido por **S. Bacci, J. Alfaro, C. Wiecko y N. Parga**

autorizo su publicación en la revista **Le Journal de Physique**

el costo de la publicación es **gratuito**

D.I.D.
C.A.B.
rsc


ARTURO LOPEZ DAVALOS
Director
Centro Atómico Bariloche

BUENOS AIRES, 29 JUN 1988

A LA DIRECCION DEL CENTRO ATOMICO BARILOCHE:

De acuerdo a lo establecido en la Resolución de Presidencia N°1018, artículos 4° y 5°, se remite en devolución el presente expediente.

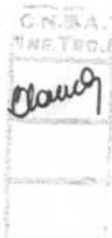
El trabajo titulado: "NUMERICAL ANALYSIS OF A NEURAL NETWORK WITH HIERARCHICALLY ORGANIZED PATTERNS".

producido por: S. Bacci, J. Alfaro, C. Wiecko y N. Parga.

ha sido visado y se ha procedido al archivo de una copia del original en esta Comisión de Publicaciones.

Queda a cargo del organismo de origen la prosecución de los trámites del pago de gastos que origine la publicación.

Solicitamos que nos envíen 3 separatas de la publicación ya editada, a fin de que podamos ingresar este material en el catálogo general de las publicaciones de la CNEA.



A handwritten signature in ink, which appears to be "Tito Suter".

Dr. TITO SUTER

JEFE

DTG. INFORMACION TECNICA

Presidente de la
Comisión de Publicaciones