The ultrametric organization of memories in a neural network

N. Parga, M.A. Virasoro

Tome 47 N° 11 NOVEMBRE 1986

LE JOURNAL DE PHYSIQUE

J. Physique **47** (1986) 1857-1864 NOVEMBRE **1986**, PAGE 1857

Classification
Physics Abstracts
87.30G — 64.60C — 75.10H — 89.70

The ultrametric organization of memories in a neural network

N. Parga (*) and M. A. Virasoro

International Centre for Theoretical Physics, Trieste
and, Dipartimento di Fisica, Universita' di Roma I « La Sapienza », Roma, Italy

(*Reçu le 7 octobre 1985, accepté sous forme définitive le 21 mai 1986*)

Résumé. — Dans le modèle de mémoire humaine proposé par Hofpield, les mots à emmagasiner doivent être orthogonaux. Du point de vue de la catégorisation, cette condition est peu commode à moins que ces mots ne soient des prototypes appartenant à des catégories primordiales différentes. Dans ce cas, le modèle doit être complété de façon à pouvoir emmagasiner tout l'arbre hiérarchique : des sous-catégories appartenant à une même catégorie, des éléments appartenant à une sous-catégorie, et ainsi de suite. Nous utilisons des résultats récents sur la théorie du champ moyen des verres de spin pour démontrer que cette réalisation est possible avec une modification minimale de la règle de Hebb. On trouve que la catégorisation est une conséquence naturelle d'une étape de précodage structurée en couches.

Abstract. — In the original formulation of Hopfield's memory model, the learning rule setting the interaction strengths is best suited for orthogonal words. From the point of view of categorization, this feature is not convenient unless we reinterpret these words as primordial categories. But then one has to complete the model so as to be able to store a full hierarchical tree of categories embodying subcategories and so on. We use recent results on the spin glass mean field theories to show that this completion can be done in a natural way with a minimal modification of Hebb's rule for learning. Categorization emerges naturally from an encoding stage structured in layers.

1. Introduction.

Neural networks have been proposed as associative memories to model the behaviour of human long term memory [1, 2]. A neural network is essentially an amorphous aggregate of neurons that, in this context, are idealized as physical two-state devices [3] coupled through a symmetrical matrix J_{ij} that represents the synapses.

During « learning » the J_{ij} are modified by the environment in a time scale T_J larger than the time scale T_S needed by the neurons to adapt to the coupling (retrieval process). So defined the model has the following properties.

i) It stores a certain amount of information (if N is the number of neurons, one can store $< -0.15\,N$ [2], [17] words of N bits).

ii) The retrieval of the information is such that from a partial, deteriorated, knowledge of one word the system is able to reconstruct the full word (the number of bits that can be so reconstructed decreases as we try to store more words and goes rapidly to zero when one surpasses the storage capability referred to in i).

iii) After storing P words the system — requested to retrieve a particular pattern — may, generically, answer with a « spurious » word, i.e. one that has not been originally stored. The number of spurious words increases at least exponentially with P [4].

iv) The storage prescription adopted in e.g. reference [2] works best if the words are at least approximately orthogonal.

The type of organization of a neural network is natural for a biological system. The idea that the J_{ij} are modified by the environment eliminates the paradox that would follow if the information about the coupling matrix among 10^{14} neurons had to be contained in the DNA [5]. Other types of architecture would be difficult to reconcile with prevalent ideas about the type of order that can evolve from evolution.

From the neural sciences point of view, such a model has the drawback of assuming symmetric synapses. Modifying it with an asymmetric matrix J_{ij} leads us outside the well-known domain of statisti-

(*) Permanent address : Centro Atomico Bariloche, 8400 Bariloche, Argentina.

cal mechanics into the realms of cellular automata where must less is known. As one is trying to attack a problem hitherto overlooked in statistical mechanics, namely, how to constrain a system to have certain states as equilibrium states it is clearly safer to proceed by steps studying first the behaviour of a symmetric neural network.

The major testable predictions of these models fall in the realm of cognitive psychology : we should confront the behaviour of a neural network with what we know about human memory. Comparisons of this type have already been discussed in the literature. For instance the spurious words have been seen as a proof that these models are not just repetitive but have the ability to imagine new representations [6]. The storage capability [2, 7] has also been discussed in this context.

From this point of view, we notice as a serious flaw in the model that words or patterns to be stored have to be encoded in approximately orthogonal vectors just exactly the opposite of the way human memory works. It is apparent that when we try to memorize new information we look for all the possible relationships with previously stored words [8]. If we can classify it, that is place it in a tree of categories we do it with so much eagerness that sometimes we just censor the data so as to cancel any exceptional anomalous features. However, if the word is really orthogonal to all the previously stored ones we have reluctantly to initiate a new category.

We therefore propose a reinterpretation of the patterns discussed in [1, 2, 4, 7] as primordial categories. Then the problem of storage capability is renormalized : how many totally uncorrelated patterns are we able to memorize ?

Another problem of the model, closely related to the previous one, is that when errors in retrieval occur, either because of spurious states or because of the limits in capacity, there is no way to control the quality of these errors. On the other hand, biosystems cannot afford certain errors more than once in a lifetime. A hierarchy among errors is therefore mandatory. This can be automatically implemented if we have classified the patterns in a hierarchial tree of categories. An unimportant error will be to confuse individuals inside a category while a more serious one will be to confuse categories (incidentally it follows that this type of categorization must occur very early in the evolutionary tree : distinguishing between prey and predators is more vital than distinguishing among varieties of predators).

Both problems above will be solved if we could modify the model in such a way that the patterns to be memorized instead of being orthogonal fall into a hierarchical tree. This type of organization (ultrametricity) appears spontaneously in a Sherrington-Kirkpatrick spin-glass [9]. If we assume that the synaptic connections J_{ij} are chosen independently at random with zero average and standard deviation $\sigma = 1/\sqrt{N}$, and look for the solutions of the equation (labelled by α)

$$S_i^\alpha = \text{sign}\left(\sum_j J_{ij} S_j^\alpha\right) \quad \alpha = 1, 2, \ldots \qquad (1.1)$$

whose energies

$$E^\alpha = \sum_{i>j} J_{ij} S_i^\alpha S_j^\alpha \qquad E^1 \leqslant E^2 \leqslant E^3 \ldots \quad (1.2)$$

do not differ too much from the ground state energy

$$\lim_{N\to\infty} (E^\alpha - E^1) = \text{finite}, \qquad (1.3)$$

then one can prove that chosen three states at random the two of them that are nearer to each other lie exactly at the same distance of the third, The distance is the natural one : $(d(\alpha, \beta))^2 = \frac{1}{N}\sum_{i=1}^{N} (S_i^\alpha - S_i^\beta)^2$. This property, called ultrametricity, implies that the states can be located in the higher ends of the branches of a tree. The distance between two states is then measured by how much one has to go down along the tree before the two branches converge.

Due to the similarity between the SK spin glass and a neural network there is an exciting possibility that the type of architecture of the network leads spontaneously to categorization.

In this paper we build a detailed model to check whether something like that could happen.

Human memory is an information processing system that can be conveniently analysed in three subprocesses : encoding, storage and retrieval. In the first stage stimuli are encoded into a form compatible with storage. In section 2 we show how such an encoding processing unit should be architectured to generate words that are ultrametric. We show that a system organized in layers so that part of the stimuli enter at different levels of the encoding process naturally leads to a hierarchical organization of the words to be stored.

In section 3 we study the storage stage. We show that a minimal modification of the Hebb rule is required. The J_{ij} that come out have the same structure as those of the SK spin glass [10].

In section 4 we consider the retrieval process. We restrict our analysis to the thermodynamic limit where, N, the number of neurons, goes to ∞, while P/N goes to zero. We do not discuss the problem of storage capability but compare our results on retrieval of spurious state with the ones derived in the Hopfield model [4]. We find that there are two types of spurious states : whether they mix patterns belonging to a same category or mix categories. A hierarchy of errors emerges. In section 5 we discuss our results.

2. Encoding. Generation of ultrametric words.

Approximately orthogonal words can be generated in the following way : one chooses every bit in a word + 1 or − 1 with equal probability. Then two such

words $\{ S_i^\alpha \}$ and $\{ S_i^\beta \}$ satisfy

$$\frac{1}{N}\sum_{i=1}^{N} S_i^\alpha S_i^\beta = O\left(\frac{1}{\sqrt{N}}\right), \qquad \alpha \neq \beta . \quad (2.1)$$

If we want to generate ultrametric words we have to proceed differently. The example of spin glasses [10] suggests the following generic procedure. We consider an inhomogeneous Markov process with K time steps (in the figure, $K = 3$). At each step we choose a value for a certain random variable with a probability distribution with parameters depending on the result of the previous step, i.e. we define

$$P_k(y_k \mid y_{k-1})\, dy_k = \text{probability distribution of } y_k \text{ conditioned to the value } y_{k-1},\ k = 1, \ldots, K . \quad (2.2)$$

The random variables y_k could in general be real continuous but for simplicity we will choose them discrete.

The draw of a y_k value is repeated M times, where M is the number of branchings of the tree we are trying to generate. At the end we will have R choices of y_K (see Fig. 1). Then the value of the spin at site i will be

$$S_i^\alpha = \text{sign } y_K^\alpha \qquad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K) . \quad (2.3)$$

The whole procedure is repeated, independently, at every site of the lattice. In this way we generate works which are ultrametric to order $1/\sqrt{N}$.

Instead of distances it is more convenient to work in terms of overlaps i.e.

$$q^{\alpha\beta} = \frac{1}{N}\sum_i S_i^\alpha S_i^\beta \quad (2.4)$$

in the limit $N \to \infty$ the overlaps among the R states are solely dependent on the level where the branches originating in the different states converge. In the example of figure 1

$$q_3 = 1 = \int P_1(y_1)\, dy_1 \int P_2(y_2 \mid y_1)\, dy_2 \times \times \int P_3(y_3 \mid y_2)\, dy_3 (\text{sign } y_3)^2$$

$$q_2 = \int P_1(y_1)\, dy_1 \int P(y_2 \mid y_1)\, dy_2 \times \times \left[\int P(y_3 \mid y_2) \text{ sign } y_3\, dy_3\right]^2$$

$$q_1 = \int P_1(y_1) \left[\int P_2(y_2 \mid y_1)\, dy_2 \times \times \int P(y_3 \mid y_2) \text{ sign } y_3\, dy_3\right]^2 dy_1 . \quad (2.5)$$

There are many different Markov processes that generate the same tree with the same overlaps. The differences manifest themselves in the value of the multiple correlations

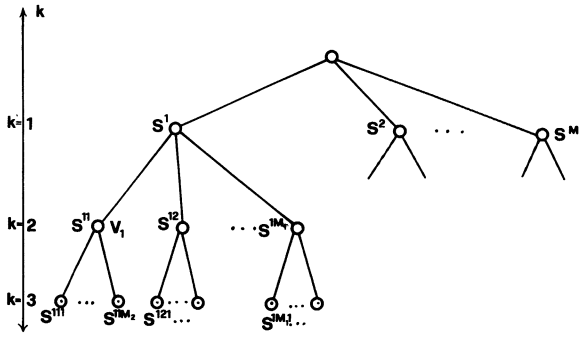

Fig. 1. — An ultrametric tree with three levels. The states lie at the lower ends of the branches. They are parametrized by three superindices that indicate their genealogy. At the other branchings we place the ancestors.

$$\frac{1}{N}\sum_i S_i^\alpha S_i^\beta \ldots S_i^\omega . \quad (2.6)$$

We have investigated several different examples, albeit in a non-systematic way. In the following we will after refer to the following two particular cases.

A) At level k, y_k can take only two values $\pm r_k$ with

$$P_k(+) = (\text{probability that } y_k = + r_k) = \frac{1}{2}\left(1 + \frac{y_{k-1}}{r_k}\right)$$

$$P_k(-) = (\text{probability that } y_k = - r_k) = \frac{1}{2}\left(1 - \frac{y_{k-1}}{r_k}\right). \quad (2.7)$$

Furthermore $r_K = 1$ so that sign $y_K = y_K$. Obviously

$$\bar{y}_k = (\text{average value of } y_k/\text{conditioned to } y_{k-1}) = y_{k-1}$$

$$\bar{y}_k^2 = r_k^2 \quad (2.8)$$

Therefore $q_k = r_k^2$.

B) At level k, y_k can take three values $(-1, 0, 1)$ with the following conditional probabilities :

a) if $y_{k-1} = \pm 1$ then $y_k = \pm 1$ with probability one,

b) if $y_{k-1} = 0$ then

$$P_k(y_k = 1) = P_k(y_k = -1) = r_k/2$$

$$P_k(y_k = 0) = 1 - r_k . \quad (2.9)$$

The r_k increase with k and $r_K = 1$. In this case again

$$\bar{y}_k = y_{k-1}; \qquad \bar{y}_k^2 = r_k \tag{2.10}$$

and therefore $q_k = r_k(1 - q_{k-1}) + q_{k-1}$.

In the context of an ultrametric organization it is useful to define ancestors of a word [10]. If there are M words that have all the same overlap q among them we can define their common ancestors as the configuration which obtains when we average the local value of the spin of all the descendants. The ancestors lie at every branching point and embody the amount of information common to the descendants. It is the prototype of a category and may be different from all the states belonging to the category. In the example of figure 1 the ancestor located at the vertex V_1 has local magnetization

$$m(y_2) = \int P(y_3 \mid y_2)\, dy_3 \text{ sign } y_3 . \tag{2.11}$$

We can define a natural cell structure on the ancestor by grouping all sites with the same value of y_2. Then each descendant knows about this cell structure. In fact, if for one particular word we average the spin value among all sites belonging to a cell then we obtain again $m(y_2)$. The cell structure and the average cell magnetization is all the information that is common to the category.

In the context of neural networks the words or patterns to be stored must be the result of an encoding stage in the biosystem. The question then arises whether the general mechanism that we have discussed to generate ultrametric works is in any sense natural.

For this purpose we remember that any Markov process can be seen as following a discretized Langevin equation so that

$$a(y_k) = b(y_{k-1}) + \varepsilon , \tag{2.12}$$

where ε is a random noise. The latter represents the input stimuli. Equation (2.12) suggests an encoding system structured in ordered layers. Each layer receives from the previous one an amount of information (partially elaborated) that determines a set (presumably finite and not too large) of possible roads which the Markov process can take. The new stimuli arriving at that level then chooses among these roads. Interestingly enough the first layer determines the primordial category into which the pattern will be classified. The details are received and coded later.

There is some partial evidence that the perception system has such a layered structure [11]. This suggests the fascinating hypothesis that the most primitive categorization follows directly from a layered perception system without any elaboration by the central nervous system. In more evolved species perhaps the same layered structure has been reproduced for the encoding stage.

3. The storage stage. The Hebb rule.

We discuss in this section how the pattern is stored in the synapses J_{ij}. We will show that our prescription is very similar to the relation that exists in the spin glass between the equilibrium states and the couplings. The prescription is, however, extremely simple and natural so that a reader, not interested in this comparison, may simply jump to equation (3.6).

In reference [10] it was proved that each state (with all its ancestors) defines a clusterization of the sites in cells. Each cell (of a particular state) is labelled by an K-tuple of real parameters where K is the total number of generations. Each cell of an ancestor, K' generations older than the state, is characterized by $K - K'$ real parameters. If we call

$$\mathcal{C}_{y_0 y_1 \ldots y_K} = \{ \text{ set of sites belonging to the cell } y_0\, y_1 \ldots y_K \text{ of the state } \alpha \} \tag{3.1}$$

then the cell of its ancestor

$$\mathcal{C}_{y_0 y_1 \ldots y_{K-K'}} = \bigcup_{y_{K-K'+1} \ldots y_K} \mathcal{C}_{y_0 y_1 \ldots y_K} . \tag{3.2}$$

We now define

$$J_{y_0 \ldots y_l, y'_0 \ldots y'_l} = \sum_{\substack{i \in \mathcal{C}_{y_0 \ldots y_l} \\ j \in \mathcal{C}_{y'_0 \ldots y'_l}}} J_{ij} . \tag{3.3}$$

Then it was proved that [10]

$$\begin{aligned} J_{y_0 \ldots y_l, y'_0 \ldots y'_l} = {} & J_{y_0 \ldots y_{l-1}, y'_0 \ldots y'_{l-1}} + \\ & + \left(\frac{y'_l - y'_{l-1}}{q_l - q_{l-1}} - \frac{\rho_l}{2} \left(m_{q_l}(y'_l) - m_{q_{l-1}}(y'_{l-1}) \right) \right) \left(m_{q_l}(y_l) - m_{q_{l-1}}(y_{l-1}) \right) \\ & + \left(m_{q_l}(y'_l) - m_{q_{l-1}}(y'_{l-1}) \right) \left(\frac{y_l - y_{l-1}}{q_l - q_{l-1}} - \frac{\rho_l}{2} \left(m_{q_l}(y_l) - m_{q_{l-1}}(y_{l-1}) \right) \right) \end{aligned} \tag{3.4}$$

where q_l, q_{l-1} are the overlaps at the considered level, $m_{q_l}(y_l)$ is the average magnetization on the cell and ρ is related to the limit of Parisi's order parameter at $T = 0$.

Based on this information we propose

$$NJ_{ij} = \sum_{\alpha_1 \ldots \alpha_l} \frac{(S_i^{\alpha_1 \ldots \alpha_l} - S_i^{\alpha_1 \ldots \alpha_{l-1}})(S_j^{\alpha_1 \ldots \alpha_l} - S_j^{\alpha_1 \ldots \alpha_{l-1}})}{q_l - q_{l-1}} + NJ_{ij}^{\text{ancestors}}$$

$$S_i^{\alpha_1 \ldots \alpha_{l-1}} = \sum_{\alpha_l} S_i^{\alpha_1 \ldots \alpha_l} \,\%\, \sum_{\alpha_l} 1 \tag{3.5}$$

where the labels $\alpha_1 \ldots \alpha_{l-1}\ \alpha_l$ indicate the descendants of the ancestor $\alpha_1 \ldots \alpha_{l-1}$; q_l is the self-overlap of the descendant while q_{l-1} is the overlap among the descendants of the same ancestor.

In the case of figure 1 we derive

$$NJ_{ij} = \sum_{\alpha\beta\gamma} \frac{(S_i^{\alpha\beta\gamma} - S_i^{\alpha\beta})(S_j^{\alpha\beta\gamma} - S_j^{\alpha\beta})}{1 - q_2} + \sum_{\alpha\beta} \frac{(S_i^{\alpha\beta} - S_i^{\alpha})(S_j^{\alpha\beta} - S_j^{\alpha})}{q_2 - q_1} + \sum_{\alpha} \frac{S_i^{\alpha} S_j^{\alpha}}{q_1}. \tag{3.6}$$

This learning mechanism generalizes the Hebb rule. The original Hebb mechanism is unique in the sense that storing a new word does not require any « thought » or « reflection » on the part of the central nervous system. This is reasonable as long as we are trying to store uncorrelated words. However, it becomes artificial as soon as there is any relational content. The generalization we are proposing, suitable for an ultrametric organization, is natural and fast.

Assume that R words have already been stored and we are trying to store the $R + 1$. If R is sufficiently large, the new word will not, in a first approximation, affect the average configuration corresponding to the ancestors. So at this point there is no need to recall previous information to store the new word. In a second moment, however, the system may have to modify the ancestor taking into account the new pattern. Only then the piece of J_{ij} depending on the ancestors will be modified. The prototype of the category has changed.

4. The retrieval process. The spurious words.

In this section we discuss the ability of the system to recall a stored pattern. For practical reasons we will refer to one particular example of a tree with two generations. There will be P number of ancestors ($\alpha = 1 \ldots P$) and P_α number of descendants of ancestor α. Both these numbers can become large but they are negligible with respect to N, the number of sites that tends to infinity. Then

$$NJ_{ij} = \sum_{\alpha=1}^{P} \frac{S_i^{\alpha} S_j^{\alpha}}{q} + \sum_{\alpha=1}^{P} \sum_{\beta=1}^{P_\alpha} \frac{(S_i^{\alpha\beta} - S_i^{\alpha})(S_j^{\alpha\beta} - S_j^{\alpha})}{1 - q}$$

$$J_{ii} = 0 \tag{4.1}$$

The system starts from an initial configuration and relaxes according to the equation [2]

$$S_i(t + \Delta t) = \operatorname{sign}\left(\sum_j J_{ij} S_j(t)\right) \tag{4.2}$$

until it reaches an asymptotic state. The updating can be done in an infinite number of different ways from the purely sequential rule, as in usual Monte Carlo algorithms, to a parallel synchronous updating in which all spins are flipped simultaneously [1]. The limit motion in these two cases satisfy the equations

$$\text{sequential} \qquad S_i = \operatorname{sign}\left(\sum_j J_{ij} S_j\right) \tag{4.3a}$$

$$\text{synchronous} \begin{cases} S_i = \operatorname{sign}\left(\sum_j J_{ij} S'_j\right) \\ S'_i = \operatorname{sign}\left(\sum_j J_{ij} S_j\right). \end{cases} \tag{4.3b}$$

Solutions of (4.3a) are included among the solutions of (4.3b). A biosystem will act somehow half way between these two extremes. There is no clock to synchronize the updating but delays in signal transmission amount to a certain degree of parallelism.. The relevant dynamics therefore corresponds to parallel updating of conveniently chosen subsets of spins. If these subsets are sufficiently random and change in each updating the general solution of (4.3b) cannot remain stable. Therefore a biosystem will eventually evolve into the fixed points of (4.3a).

The presence of random noise is modelled as in statistical mechanics through the introduction of temperature. Equation (4.2) becomes

$$\mu_i(t + \Delta t) = \tanh\left(\beta \sum_j J_{ij} \mu_j(t)\right) \tag{4.4}$$

where μ_i is the short time average of S_i.

Equation (4.2) can be conveniently studied by contracting it will all the spin configurations corresponding to words stored in the J_{ij} [12]. Calling

$$m^{\alpha\beta} = \frac{1}{N} \sum_i S_i^{\alpha\beta} S_i, \qquad m^{\alpha} = \frac{1}{N} \sum_i S_i^{\alpha} S_i \tag{4.5}$$

we derive

$$m^{\alpha\beta}(t + \Delta t) = \left\langle\!\!\left\langle S^{\alpha\beta} \operatorname{sign}\left(\sum_\alpha \frac{S^\alpha m^\alpha(t)}{q} + \sum \frac{(S^{\alpha\beta} - S^\alpha)}{1 - q}(m^{\alpha\beta}(t) - m^\alpha(t))\right)\right\rangle\!\!\right\rangle \tag{4.6}$$

where the double brackets indicate the average value with respect to the stochastic variables $S^{\alpha\beta}$ and S^α. There is an energy functional that decreases monotonically during the sequential relaxation

$$E = \sum_{i>j} J_{ij} S_i S_j = \frac{1}{2} \sum_\alpha \frac{m_\alpha^2}{q} + \frac{1}{2} \sum_{\alpha\beta} \frac{(m_{\alpha\beta} - m_\alpha)^2}{1 - q}. \tag{4.7}$$

The solutions of equation (4.3a) are of the following type :

a) the original patterns or words stored in the system, i.e.

$$m^{\alpha_0\beta_0} = 1 \; ; \; m^{\alpha_0\beta} = q \, , \; \beta \neq \beta_0 ; \; m^{\alpha\beta} = 0 \, , \; \alpha \neq \alpha_0 \tag{4.8}$$

b) spurious solutions which mix descendants of the same ancestor

$$\begin{aligned} m^{\alpha_0\beta_i} &\neq 0 \, , \quad i = 1, ..., p \\ m^{\alpha\beta} &= 0 \, , \qquad \alpha \neq \alpha_0 \end{aligned} \tag{4.9}$$

c) spurious solutions which mix different ancestors

$$m^{\alpha_i\beta_j} \neq 0 \, , \quad j = 1 \ldots p_{\alpha_i} \, , \qquad i = 1 \ldots p \, . \tag{4.10}$$

The solutions of the Hopfield model have been analysed in great detail [4]. Unfortunately some of the more interesting properties of the solutions are sensitive to the distribution probability of the S_i^α. If we choose instead

$$NJ_{ij} = \sum_\alpha \xi_i^\alpha \, \xi_j^\alpha \tag{4.11}$$

where ξ_i^α is a random continuous variable with $\overline{\xi_i^\alpha} = 0$ and $\overline{\xi_i^{\alpha 2}} = 1$, then the number of spurious solutions and their stability properties are modified.

In our case we do not have any « natural » choice for the random process generating the $S_i^{\alpha\beta}$, S_i^α. Therefore we cannot expect to derive detailed results. We must rely much more on semiqualitative arguments. Eventually we should try to optimize storage and retrieval capabilities by choosing a particular ultrametric tree.

Solutions of the type a) and b) are similar to the ones appearing in the Hopfield model. We can argue that a category delimits a cone-like region in the space of configurations with the central axis coinciding with the ancestor. The « words »

$$\{ S_i^{\alpha_0\beta} - S_i^{\alpha_0} \}$$

are vectors which in the $N \to \infty$ limit are orthogonal as in the Hopfield model. Their probability distribution is in general different and this affects the general properties of the spurious solutions. However we can choose it so as to deplete the probability of zero and then we expect the same qualitative behaviour, i.e. :

i) solutions of type a) are stable below a certain critical temperature;

ii) the number of spurious solutions increases at least exponentially with P_α;

iii) the spurious solutions with low energy are essentially linear combinations of the original patterns;

iv) the attraction basin of all spurious solutions increases with P_α, this means that the attraction basin of the spurious solution lies in the frontier region between the original patterns.

For the tree generated by the Markov process B) of section 2 all of these arguments can be rigorously proved because one can derive

$$\frac{m^{\alpha_0\beta} - m^{\alpha_0}}{1 - q} = \left\langle S^{\alpha_0\beta} \operatorname{sign} \sum_\beta S^{\alpha_0\beta} \frac{(m^{\alpha_0\beta} - m^{\alpha_0})}{1 - q} \right\rangle \tag{4.12}$$

an equation identical to the one in the Hopfield model [4].

For the random choice A) there is one particular spurious solution degenerate with the original patterns, namely the ancestor itself

$$S_i = \pm \, S_i^\alpha \, | \, q \, |^{-1/2} \, . \tag{4.13}$$

The solutions of type c) are the more interesting in this model. They can be interpreted as spurious categories. In the introduction we have stressed that one of the motivations to introduce a hierarchy among patterns was the need to introduce a hierarchy among errors during retrieval. In this context we would like to analyse whether the attraction basin of the spurious categories grows exponentially with the number of categories or with the number of patterns. In the latter case something would have to be modified in the model. It is already hard to understand why in Hopfield's model the percentage of stimuli that leads to spurious patterns grows when we store more patterns ([1]). In our case spurious categories can be easily interpreted (they classify border case patterns). But it is hard to accept that the number of external stimuli that should end in these spurious categories should grow exponentially with the number of patterns correctly stored in the originally categories.

We have analysed several ultrametric trees, albeit in a non-systematic way. The tree generated by the random process A) of section 2 is more similar to Hopfield's model and therefore facilitates comparisons. We consider the most interesting case :

$$P_\alpha \gg P \tag{4.14}$$

i.e. number of patterns belonging to a category much larger than the number of categories.

In that case for a large class of input stimuli the system will retrieve a spurious pattern with appreciable overlap on a large set of original patterns. Therefore we can apply the central limit theorem to the distribution of

$$X = \sum_{\alpha\beta} \frac{(m_\alpha^{\alpha\beta} - m^\alpha)}{1 - q} (S^{\alpha\beta} - S^\alpha) \tag{4.15}$$

([1]) Unless we see it as a proof that « the more we learn the more we realize we do not know nothing ». However a spurious pattern is not recognized as such so it would be more similar to « the more we learn the more we deceive ourselves ».

with $\overline{X} = 0$ and

$$\overline{X}^2 = \sum_{\alpha\beta} \frac{(m^{\alpha\beta} - m^{\alpha})^2}{1 - q} = r(t) \qquad (4.16)$$

then the equations of motions become

$$m^{\alpha}(t + \Delta t) = \left\langle\!\!\left\langle S^{\alpha} \operatorname{sign}\left(\sum_{\alpha} S^{\alpha} m^{\alpha}(t) + X\right) \right\rangle\!\!\right\rangle \qquad (4.17)$$

$$\sum_{\alpha\beta} v^{\alpha\beta}(t + \Delta t)\, v^{\alpha\beta}(t) = \left(\sum_{\alpha\beta} (v^{\alpha\beta}(t))^2\right)^{1/2} \times \times \left\langle \sqrt{\frac{1}{\pi}} \exp - \frac{\left(\sum_{\alpha} S^{\alpha} m^{\alpha}(t)\right)^2}{2\, r(t)\, q^2} \right\rangle$$

with

$$v^{\alpha\beta}(t) = m^{\alpha\beta}(t) - m^{\alpha}(t)\,. \qquad (4.18)$$

In equation (4.17) the average is over the S^{α} and X while in (4.18) there is a simple average over S^{α}.

Equation (4.17) at fixed $X(t)$ is similar to the corresponding equation in Hopfield's model with temperature. The fixed point

$$m^{\alpha} = 0 \qquad (4.19)$$

is stable if $r(t) > \frac{2}{\pi}$. Equation (4.18) implies that if the vector $v^{\alpha\beta}(t)$ stays in the region where the Gaussian approximation holds then its norm will decrease. This reflects in equation (4.17) as a decrease of the effective temperature. Eventually $r \to 0$ or $v^{\alpha\beta}$ is zero except for a few values of α, β. In the first case the fixed points of equation (4.17) will be the ones of the Hopfield model.

This process is reminiscent of a spontaneous simulated annealing [13]. It may actually help in the retrieval of the lowest energy patterns.

5. Discussions and outlook.

In this paper we have shown that it is possible to modify Hopfield's model in such a way that it efficiently stores words organized in categories. However, categorization does not seem to emerge spontaneously as a consequence of the particular type of architecture of a neural network but must be rather superimposed on it. On the other hand, it seems that categorization is connected to the layered character of the encoding stage or even with the perception system itself. In this sense there is interesting evidence coming from the field of cognitive psychology and neurophysiology. Grouping colours into classes is clearly a subjective task. However, there is evidence that the usual classification is universal [14] and that it has to do with the way colours are processed by the visual system [15].

The limits of the storage capability of the model have not been investigated here. We have instead shown that the system is able to distinguish between spurious solutions that belong to well-defined categories and spurious solutions that mix categories. A hierarchy among errors can therefore be introduced.

Another question that remains to be answered is whether there is a best choice for the Markov process that generates the ultrametric tree and its possible connection with the one appearing spontaneously in the SK spin glass model. This interesting problem requires much more work.

Finally we would like to stress that we are aware that the relational context that exists in the human memory is infinitely richer than the one we have been analysing. We believe however that placing a word into a more complicated semantic context requires lots of processing by the central nervous system [16]. Our point is that placing it in an ultrametric tree is almost for free.

Acknowledgments.

We would like to thank M. Mezard for a very fruitful interaction and G. Toulouse for a crucial discussion at the very beginning of this work. Looking backwards we now realize how many of his ideas we have been using freely. We would also like to thank Professor Abdus Salam, the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

References

[1] LITTLE, W. A., *Math. Biosci.* **19** (1974) 101.
[2] HOPFIELD, J. J., *Proc. Natl. Acad. Sci. USA* **79** (1982) 2554.
[3] MCCULLOCH, W. S. and PITTS, W., *Bull. Math. Biophys.* **5** (1943) 115.
[4] AMIT, D. J., GUTFREUND, H. and SOMPOLINSKY, H., *Phys. Rev.* **A 32** (1985) 1007.
[5] See for example : CHANGEUX, J. P., in *Disordered Systems and Biological Organization*. Eds. E. Bienenstock *et al.* (Springer Verlag, Berlin-Heidelberg) 1986.
[6] ANDERSON, J. A., *IEEE Trans. Sys. Man Cybern.* **SMC-13** (1983) 799.
[7] PERETTO, P., *Biol. Cybern.* **50** (1984) 51.
[8] See for example : KLATZKY, R. L., *Human Memory* (W. H. Freeman, San Francisco) 1975.
[9] MEZARD, M., PARISI, G., SOURLAS, N., TOULOUSE, G. and VIRASORO, M. A., *Phys. Rev. Lett.* **52** (1984) 1156 ; *J. Physique* **45** (1984) 843.
[10] MÉZARD, M. and VIRASORO, M. A., *J. Physique* **46** (1985) 1253.
[11] KUFFLER, S. W., NICHOLS, J. G. and MARTIN, A. R.,

From Neuron to Brain, 2nd Ed. (Sinauer Associates Inc., Sunderland, Massachusetts) 1984.

[12] PROVOST, J. P. and VALLÉE, G., *Phys. Rev. Lett.* **50** (1983) 598.

[13] KIRKPATRICK, S., GELATT, Jr. C. D., and VECCHI, M. P., *Science* **220** (1983) 671.

[14] ROSCH, E., On the internal structure of perceptual and semantic categories, in *Cognitive Development and the Acquisition of Language*, Ed. T. E. Moore (Academic Press, New York) 1973.

[15] DE VALOIS, R. L. and JACOBS, G. H., *Science* **162** (1968) 533.

[16] See for example : HINTON, G. E., SEJNOWSKI, T. J. and ACKLEY, D. H., *Cognitive Sci.* **9** (1985) 147.

PERSONNAZ, L., GUYON, I. and DREYFUS, G., Neural network design for efficient information retrieval in *Disordered Systems and Biological Organization*, Eds. E. Bienenstock *et al.* (Springer Verlag, Berlin-Heidelberg) 1986.

[17] AMIT, D. J., GUTFREUND, H. and SOMPOLINSKY, H., *Phys. Rev. Lett.* **55** (1985) 1530.