INFORME TECNICO CNEA-CAB 89/004

Título : SPIN GLASSES AND NEURAL NETWORKS

Autor N. PARGA

Enviado a : Proceedings of the XIX International Seminar on Theoretical Physics, Jaca, Spain, June 1988

Fecha : 30 de enero de 1989

COMISION NACIONAL DE ENERGIA ATOMICA
CENTRO ATOMICO BARILOCHE
8400 SAN CARLOS DE BARILOCHE (R.N.)
ARGENTINA

SPIN GLASSES AND NEURAL NETWORKS

Néstor PARGA

Centro Atómico Bariloche and Instituto Balseiro,
8400, Bariloche, Rio Negro, Argentina

The mean-field theory of spin glass models has been used as a prototype of systems with frustration and disorder. One of the most interesting related systems are models of associative memories. In these lectures we review the main concepts developed to solve the Sherrington-Kirkpatrick model and its application to neural networks.

1. INTRODUCTION

Originally proposed to explain the behaviour of magnetic impurities in a non-magnetic material, spin glass models have become interesting by themselves[1,2]. This is not only because of their applications to Statistical Mechanics but also to other areas of Science.

This situation was motivated by the progress done in the understanding of the mean-field theory of spin glasses[2] and, in particular, of the multi-valley structure of its energy landscape.

In the next section the concept of frustration is given and the spin glass models are defined. It is also discussed how two relevant features of these models, disorder and frustration, also appear in optimization problems. Section 3 is devoted to the mean-field theory of spin glasses; the mean-field equations, the replica method, order parameters, the hierarchical organization of pure states and several ultrametricity tests are discussed. An application of some of these spin-glass concepts to the graph colouring problem is given at the end of the section.

In Section 4 models of associative memories are reviewed. After a discussion of the Hopfield model some of its unrealistic features are mentioned. The modifications introduced in the standard model to get rid of some of them are described. The last section contains the conclusions.

2. FRUSTRATED SYSTEMS. SPIN GLASSES AND OPTIMIZATION PROBLEMS

The mean - field theory of spin glasses, the Sherrington Kirkpatrik (SK) model[1], has been taken during the last few years as a prototype of systems which have complex cooperative behaviour as a consequence of the presence of disorder and frustration. Because of this, recent developments in the area have had influence in other fields[2], these include:

Statistical Mechanics, here we can mention problems with quenched disorder, diffusion in hierarchical spaces, etc.

Computer Science, in this field the influence has been both in the design of special algorithms and in a possible connection between the theory of complexity and properties of disordered systems.

Brain Modelling, specially with respect to the brain ability to work as an associative memory[3].

Biology, it was suggested that the model could be applied to problems such as protein function and structure or to the immune system. A possible connection between cellular automata used to describe cell differentation[4] (the Kauffman model) and the SK model could give clues to understand disordered automata networks[5].

The common feature of these systems, which is responsible of its main properties, is that they are governed by a random function of a large number of variables acting in a conflictive way to achieve a specific goal, typically the minimization of that function.

Spin glass models are the best understood of these systems. They have been originally proposed to explain the behaviour of random magnetic impurities in a non-magnetic material[6]. The impurities are described by quenched variables, i.e. variables which do not fluctuate in any relevant time scale. In a real system the randomness is in their position, the simplest model instead assumes the impurities are on the sites of a regular lattice and that the effect of the disorder can be accounted for their mutual couplings. What the physical system and the theoretical models still have in common is the existence of conflict between the interactions, an effect known as frustration[7].

The Edwards-Anderson model is defined by the hamiltonian[6]

$$H_{EA} = -\sum_{i<j}^{N} J_{ij} S_i S_j \qquad (2.1)$$

where the N magnetic moments S_i are on a regular lattice but the nearest-neighbour couplings are obtained with a distribution, usually the gaussian or the ±J distributions, allowing both ferromagnetic and antiferromagnetic values. For these cases

$$P(J_{ij}) = 1/2\ [\delta(J_{ij} - J) + \delta(J_{ij} + J)] \qquad (2.2a)$$

or

$$P(J_{ij}) = \frac{1}{\sqrt{2\pi J^2}} \exp\left[- J_{ij}^2/2\ J^2\right] \qquad (2.2b)$$

respectively.

Since the couplings can be either positive or negative a given spin may receive conflicting demands from its neighbours in order to lower the energy function H_{EA}. This is clearly appreciated in figure 1 where S_3 has to decide whether to align with S_1 or to adopt a direction opposite to S_2. A loop such as this is said to be frustrated.

The existence of frustration makes it hard to find the ground state of the system since in general the problem cannot be solved by a local analysis. On the other hand many solutions of similar energy exist because there are many equivalent ways to frustrate bonds.

Frustration also occurs in problems of interest in Computer Sciences which consist in the optimization of a cost function[8]. This is again a function of a large number of random variables. Well-known examples are the travelling salesman problem (TSP) and the graph colouring problem (GCP). In the first a set of N cities is given together with its associated distance matrix. The number of cities is usually large and the distance is not necessarily euclidean. The TSP consists in finding the shortest tour such that after visiting once every city it returns to the initial one. Similary to the spin glass model discussed previously, the TSP is a disordered system where either the positions of the cities or the distance matrix itself are random. Again a local strategy does not work, if starting from any of the cities a tour is completed by the simple procedure of choosing systematically the nearest

one frustration will show up when closing the loop.

In the GCP a random graph has to be painted with q colours minimizing the number of neighbouring sites painted with the same colour. The disorder appears because the graph is random. The cost function associated to a given painting of the graph is

$$H_{GCP} = \sum_{(ij)} \delta_{C_i, C_j} \qquad (2.3)$$

where the sum is over all the bonds of the graph and C_i denotes the colour of vertex i. Let us notice that in the language of Statistical Mechanics the GCP is equivalent to find the ground state of an antiferromagnetic Potts model defined on the random graph. Frustration appears for q=2 and q=3. Figure 2 shows a graph which cannot be painted with three colours without frustrated bonds. It has been proved that for q=4 planar graphs are not frustrated. The same is true for q>4.

The TSP and the GCP for q=3 are examples of very difficult problems called NP-complete[9]. These are characterized by the fact that no one could find an algorithm able to solve them in polynomial time in the number of variables N. Only exponential time algorithms are known. But no one could show that such polynomial algorithms do not exist. Besides it is known that if one of them could be solved in polynomial time then the whole class has the same property.

An interesting NP-complete problem for us is the mean-field approximation of the spin glass model defined previously. It is the SK model we mentioned before. After many years of continuous effort it could be solved using an ansatz[9] and the solution interpreted physically[10]. Of course no one can say that there is not a better solution (one with less free energy) but in spite of many attempts nobody was able to find it.

It is well-known that for the ferromagnetic case the mean-field approximation is exact for a model where the interactions are extended to all pairs of magnetic moments. The same trick was used to define the SK model.

Instead of H_{EA} we now have

$$H_{SK} = -\sum_{(ij)} J_{ij} S_i S_j \qquad (2.4)$$

with the sum running over all pairs of sites. The couplings are again chosen at random with distributions such as those in equations (2.2) but now the variance has to scale with N as

$$J^2 = \tilde{J}^2 / N \qquad (2.5)$$

in order to have a well-defined model in the thermodynamic limit.

As we said this model was solved by guessing its solution. Difficulties arise when one tries to find algorithms to obtain the ground state of a fixed sample. The simplest method consists in decreasing systematically the energy of an initial configuration (chosen randomly). Defining the internal field at site i as

$$h_i = \sum_{j=1}^{N} J_{ij} S_j , \qquad (2.6)$$

then

$$H_{SK} = - 1/2 \sum_{i=1}^{N} h_i S_i \qquad (2.7)$$

can be decreased by systematically visiting all the spins (either randomly or in an orderly way), flipping them whenever they point in the direction opposite to the associated internal field:

$$S_i(t+1) = \mathrm{Sign}\,[h_i(t)] \qquad (2.8)$$

With this procedure one easily detects the origin of the difficulties which prevent us from finding the true ground state. Under the sequential dynamics of equation (2.8) the system reaches a fixed point where it remains stuck for ever because it is stable under one spin-flip. Only if the attraction basin of the lowest energy states is big they can be found. Unfortunately a numerical simulation shows that this is not the case. In figure 3 the attraction basin B, defined as the fraction of events which reached a metastable state of energy E is represented[11]. As N increases B tends to be peaked at a value of E which is less than the one found with the ansatz: $E_0 = - 0.763$[9].

The obvious way to circumvent this problem in Statistical Mechanics is to consider that algorithm as a zero temperature quenching and to realize that the barriers between different

states can be jumped by increasing the temperature. Annealing is better than quenching for problems with metastable states. In the annealing one starts from a high temperature and decreases it slowly until an energy minimum is reached at T=0. The procedure however does not guarantee that the ground state will be found at the end. Probably it will require an extremely slow procedure and exponential times to obtain a true ground state.

Simulated annealing can be applied to the other two problems we mentioned before. For the TSP and the GCP the cost function is interpreted as an energy function and temperature is introduced by weighing configurations with the corresponding Boltzmann factor.

3. THE MEAN-FIELD THEORY OF SPIN GLASSES

3.1. The techniques

The SK model was already defined in the previous section. Here we shall review the different techniques used to understand its properties. These are

Numerical: let us notice that there are no real experiments to check the results of the theory. Numerical simulations replace experiments providing data to confront with the theoretical predictions. We shall not emphasize this approach but we shall mention the relevant results when necessary.

Mean field equations: these are the equations, first derived by D.J. Thouless, P.W. Anderson and R. Palmer (TAP), which solution gives the local magnetizations[12].

The replica method: most analytical results have been found with this technique[9].

3.2. The mean field equations. The many-valley picture.

Because of the disordered character of the model, the total magnetization is zero. For a fixed sample $\{J_{ij}\}$ the thermal averages

$$m_i = \langle S_i \rangle \quad , \qquad i = 1 \ldots N \qquad (3.1)$$

are non-zero at low temperature. The TAP equations give their values, they are[12]

$$m_i = \mathrm{th}\left[\beta \sum_j J_{ij} m_j - \beta^2 \sum_j J_{ij}^2 (1-m_j^2) m_i\right] . \qquad (3.2)$$

The first term in the argument of the hyperbolic tangent is the

same which appears for the ferromagnetic case. The second one is new , for a ferromagnet it does not contribute because the coupling scales as $J \sim O(1/N)$. For the spin glass the variance is $O(1/N)$ and the internal field has to be corrected with the Onsager term, i.e. the reaction of the other spins due to the presence of S_i. The whole expression inside the brackets is the cavity field, i.e. the field at site i in the absence of spin i.

The number of solutions of these equations can be evaluated [19]. While for $T>T_c = J$ the only solution corresponds to a paramagnetic phase, i.e. $m_i = 0$, for lower temperatures there are many non-trivial solutions with $m_i \neq 0$. The order parameter

$$q_{EA} = \frac{1}{N} \sum_{i=1}^{N} \overline{m_i^2} \qquad (3.3)$$

drops to zero above T_c signalling the existence of a phase transition to a spin glass phase where the total magnetization is zero but the spins are frozen in non-zero values.

The number of solutions below T_c grows exponentially with N: $\mathcal{N} \sim \exp[\alpha(T)N]$ with $\alpha(0) \simeq 0.2$ and $\alpha(T_c)=0$.

These solutions are interpreted as the local magnetizations associated to different minima of the free energy[14]. According to this picture configuration space splits into many sectors, the free energy barriers between any pair of them being infinite in the thermodynamic limit. For a finite system it is possible to go from one sector or valley to another; the time τ to pass the highest barriers scales as[15]

$$\ln \tau = aN^{1/4} - b \qquad (3.4)$$

where a and b depend on the temperature.

For an infinite system ergodicity is broken. Each solution of the TAP equations is the set of thermal averages $m_i^{(\alpha)} = \langle S_i \rangle_\alpha$ $(i=1,\ldots,N)$ over a single valley α.

These valleys are the pure states of the SK model. If F_α is the free energy of the pure state α the connection between the Gibb's average and those restricted to a single valley is, for a physical magnitude A

$$\langle A \rangle = \sum_\alpha P_\alpha \langle A \rangle_\alpha \qquad (3.5)$$

where $P_\alpha = \exp[-\beta F_\alpha]$ and the symbol $\langle\ \rangle$ denotes the average over the Gibb's ensemble.

3.3. Order parameter

The order parameter q_{EA} can detect the existence of the spin glass transition but it does not describe properly the nature of the low temperature phase. As we said this is characterized by the existence of many valleys, a better order parameter is one able to describe their appearance below T_c.

Let us first give a measure of the distance between two states α and β

$$d^2(\alpha,\beta) = 1/N \sum_{i=1}^{N} (m_i^{(\alpha)} - m_i^{(\beta)})^2 \ . \qquad (3.6a)$$

Alternatively we can use their overlap

$$q_{\alpha\beta} = 1/N \sum_{i=1}^{N} m_i^{(\alpha)} m_i^{(\beta)} \ . \qquad (3.6b)$$

The overlap distribution averaged over all samples with the weight (2.2), is[10]

$$P(q) = \overline{P_J(q)} = \sum_{\alpha,\beta} \overline{P_\alpha^J P_\beta^J \delta(q_{\alpha\beta}-q)} \qquad (3.7)$$

where P_α^J is the Boltzmann weight of state α of the sample J. Notice that only states with the same free energy density (as $N \to \infty$) are relevant.

P(q) is the order parameter we are looking for. For the ferromagnetic case it is a set of deltas: for $T > T_c$ the only solution is the one with $m=0$ which gives $P(q) = \delta(q)$; for $T < T_c$ there are two states related by symmetry, the corresponding overlaps are $+m^2$ and $-m^2$ what gives $P(q) = 1/2\,[\delta(q+m^2) + \delta(q-m^2)]$.

For the SK model P(q) is not trivial. Apart from the self-overlap contribution giving $\delta(q - q_{EA})$ there is a finite contribution from all values of q as it is shown schematically in figure 4.

The evaluation of P(q) with equation (3.7) requires the

decomposition of the Gibb's state into pure states[10]. For a ferromagnet this can be easily done by introducing a small magnetic field which breaks the symmetry between the two possible pure states. This can not be done for a spin glass. However it is possible to obtain P(q) by analyzing of spin correlation functions in the Gibb's state.

Let us consider the correlation functions

$$q_J^{(k)} = 1/N^k \sum_{j_1=1}^{N} \cdots \sum_{j_k=1}^{N} \left[\langle S_{j_1} \cdots S_{j_k} \rangle \right]^2 \qquad (3.8)$$

It is easy to see that these are moments of the overlap distribution, i.e.

$$q_J^{(k)} = \int dq \; q^k \; P_J(q). \qquad (3.9)$$

The proof consists in expressing each of the two powers of the correlation function in equation (3.8) as sums over pure states as in equation (3.5). Using their clustering property, equation (3.9) is immediately obtained.

Now it is possible to show that $P_J(q)$ can be expressed in terms of two real replicas, σ and τ, of the same sample as

$$P_J(q) = \frac{1}{Z^2} \sum_{\sigma,\tau} e^{-\beta H_J(\sigma)} \; e^{-\beta H_J(\tau)} \delta\left(\frac{\sum_i \sigma_i \tau_i}{N} - q \right) \qquad (3.10)$$

(to see the equivalence between this expression and the definition of $P_J(q)$, expand the Dirac delta and use the fact that the correlations in equation (3.8) are the moments defined in equation (3.9)).

This expression is suitable to perform numerical simulations. The numerical experiment simulates two identical systems generating spin configurations according to their respective Boltzmann weights. Then it is enough to measure the overlap between the two configurations at intervals Δt and to evaluate the corresponding histogram. This calculation was done by A.P. Young[16] at a value of T where the coefficients in equation (3.4) for the equilibration time were known. The result shows the shape in figure 4.

3.4. The replica method

Contrary to what is done in the TAP approach, in the replica method the quenched average is performed analitically at the beginning of the evaluation[4]. Besides, thermal averages are taken in the Gibb's state. Because of this ergodicity breaking is seen in the non-triviality of the overlap distribution P(q) and, in the language of the replica formalism, it is due to the breaking of the replica symmetry.

The free energy per spin corresponding to the sample {J} is

$$f_J = -\frac{1}{\beta N} \ln Z_J \; , \qquad (3.11)$$

as usual its quenched average is

$$f = \overline{f_J}$$

$$= \int [dJ] \; P(J) \; f_J \qquad (3.12a)$$

or

$$f = -\frac{1}{\beta N} \int [dJ] \; P(J) \; \ln Z_J \; . \qquad (3.12b)$$

In order to get rid of the $\ln Z_J$ the replica method uses

$$\ln Z_J = \lim_{n \to 0} \frac{Z_J^n - 1}{n} \; . \qquad (3.13)$$

The quenched average is replaced by the annealed average of n replicas, all of them with the same set of couplings {J}. The free energy is first evaluated for integer n and the limit $n \to 0$ is taken at the end.

Now the integrals over the couplings are straighforward (e.g. a set of N(N-1)/2 uncoupled gaussian integrals, for P(J) gaussian). After they are done we are left with an effective hamiltonian for the n replicas:

$$\overline{Z_J^n} = \sum_s \exp \left[\beta^2 Nn/4 + (\beta^2 N/2) \sum_{1 \le a \le b \le n} \left(\sum_i S_i^a S_i^b / N \right)^2 \right] \qquad (3.14)$$

here a,b = 1,...,n are replica indeces. The sum over s denotes the trace over the n replica spin configurations.

Using the gaussian transformation (with variables $Q_{a,b}$):

$$\overline{Z_J^n} = \int \prod_{a<b} \left(\frac{dQ_{ab}}{\sqrt{2\pi/N\beta^2}} \right) \exp\left(- N A[Q]\right) \qquad (3.15)$$

where

$$A[Q] = -n\beta^2/4 + \beta^2/2 \sum_{1\le a\le b\le n} (Q_{a,b})^2 - \ell n\, Z[Q]$$

and

$$Z[Q] = \sum_s \exp\left\{\beta \sum_{1\le a\le b\le n} Q_{a,b}\, S_a S_b\right\} . \qquad (3.16)$$

What is relevant for us is that the free energy f is given by the saddle point dominating this integral, which is a family of nxn matrices.

Notice that since all replicas are equivalent the permutation group is a symmetry of the problem. If the saddle point does not break this symmetry we can take[4]

$$Q_{ab} = q , \qquad (3.17)$$

where the parameter q is determined by a saddle point equation. However when the equation is solved it is found that the solution is unphysical. In particular the entropy becomes negative at low temperatures; at T=0

$$S(0) = -1/2\pi \quad . \qquad (3.18)$$

Also the internal energy does not agree with the numerical simulations.

What happens is that this solution is not stable. An analysis of the Hessian matrix $\partial^2 A/\partial Q \partial Q$ shows that it has negative eigenvalues[17].

Since the saddle point is not invariant under the whole permutation group it is necessary to break this symmetry. A successful ansatz was proposed by G. Parisi[9]; according to this solution replicas are organized in a hierarchical way. The matrix element Q_{ab} can be interpreted as the overlap between replicas a and b. In the first symmetry breaking the n replicas are distributed in n/m_1 groups or clusters of size m_1. Replicas in the

same cluster have overlap q_1 while for those in different clusters it is q_0. Of course $q_1 > q_0$. In a second symmetry breaking replicas in the same cluster are again distributed in m_1/m_2 subclusters of size m_2. At each stage of the symmetry breaking the free energy, entropy, internal energy and other physical quantities are evaluated and the convergence to the correct values is checked.

A remarkable agreement with the numerical data is obtained as the number of breakings increases. At the second symmetry breaking one has[9]

$$S(0) = - 0.004$$
$$E(0) = - 0.7636 \qquad (3.19)$$

for the zero temperature entropy and internal energy respectively.

No one has shown that this is the correct solution, but no one could disprove it by finding a better one. Its stability has been analyzed[18].

In our discussion of the TAP approach we have seen that the different solutions of the mean-field equations were directly related to a many-valley picture. How is ergodicity breaking seen in the replica approach?.

As we know the existence of many pure states leads to a non-trivial overlap distribution P(q). In order to understand the effect of replica symmetry breaking on P(q) let us express it in terms of replicas. Taking the quenched average of the correlations (3.8)

$$q^{(k)} = \overline{q_J^{(k)}}$$
$$= \frac{1}{N^k} \sum_{i_1=1}^{N} \cdots \sum_{i_k=1}^{N} \overline{\left[\langle S_{i_1} \cdots S_{i_k} \rangle\right]^2} \quad . \qquad (3.20)$$

Using again the replica trick (3.13) the moments $q^{(k)}$ can be written in terms of the replica overlaps Q_{ab}

$$q^{(k)} = \lim_{n\to 0} [Q_{ab}]^k \qquad (3.21)$$

However given a stable replica symmetry breaking matrix Q permutations of this solution will also be saddle points. Averaging over all of them

$$q^{(k)} = \lim_{n\to 0} \frac{1}{n(n-1)/2} \sum_{(a,b)} [Q_{ab}]^k \qquad (3.22)$$

But since

$$q^{(k)} = \int dq\ q^k\ P(q)\ , \qquad (3.23)$$

P(q) can be written as

$$P(q) = \lim_{n\to 0} \frac{1}{n(n-1)/2} \sum_{(a,b)} \delta(Q_{ab} - q)\ . \qquad (3.24)$$

Here we see that if P(q) is not trivial (a δ-function) the matrix Q_{ab} can not be replica symmetric. The necessity of replica symmetry breaking is related to the existence of many pure states, i.e. to ergodicity breaking.

3.5. Three overlap distribution. Ultrametricity.

P(q) measures two state correlations. To look for correlations between three states we define, similarly to equation (3.7)

$$P^J(q_1,q_2,q_3) = \sum_{\alpha\beta\gamma} P^J_\alpha P^J_\beta P^J_\gamma\ \delta(q_1 - q_{\alpha\beta})\ \delta(q_2 - q_{\beta\gamma})\ \delta(q_3 - q_{\gamma\alpha}) \qquad (3.25)$$

and its quenched average $P(q_1,q_2,q_3)$. Similarly to P(q) this distribution can be obtained in three different ways

a) using its definition, equation (3.25),

b) by Monte Carlo simulations. The analogous of equation (3.10) involves three real replicas:

$$P^J(q_1,q_2,q_3) = \frac{1}{Z^3} \sum_{\sigma,\tau,\xi} e^{-\beta\,[H_J(\sigma)+H_J(\tau)+H_J(\xi)]}\ .$$

$$\cdot\ \delta\left(\frac{\sum_i \tau_i\sigma_i}{N} - q_1\right)\delta\left(\frac{\sum_i \sigma_i\xi_i}{N} - q_2\right)\delta\left(\frac{\sum_i \xi_i\tau_i}{N} - q_3\right)\ , \qquad (3.26)$$

c) using the replica method. This is the technique which has been originally used to evaluate it[19].

The replica technique allows an analytical evaluation of the three overlap distribution. The result is

$$P(q_1,q_2,q_3) = 1/2\ P(q_1)x(q_1)\delta(q_1 - q_2)\delta(q_1 - q_3) +$$

$$+ 1/2\ \left\{P(q_1)P(q_2)\ \theta(q_1-q_2)\delta(q_2-q_3) + \text{permutations}\right\} \qquad (3.27)$$

where

$$x(q) = \int_{-1}^{q} dq'\ P(q') \qquad (3.28)$$

This expression reveals a remarkable structure in the space of the pure states of the SK model: given three of them they form a triangle which is either equilateral or isosceles with the shortest side (largest overlap) at its base. A space like this is known as an ultrametric space[20].

While for a metric space three points A,B,C obey the triangular inequality

$$d(A,C) \le d(A,B) + d(B,C) \qquad (3.29)$$

for an ultrametric space a stronger inequality holds

$$d(A,C) \le \text{Max}\ [d(A,B),\ d(B,C)]\ . \qquad (3.30)$$

A way to see pictorially this property is to put the points of the space in correspondence with the leaves of a tree, as in figure 5. The proximity between any two points is given by the value on the vertical axis where they meet, i.e. where they have their nearest common ancestor.

It is evident from this figure that ultrametricity implies a hierarchical organization of the points of the space. Cutting the tree at a scale $\bar{q}$ we see that the set of points or states splits into clusters such that the overlap between any pair of them inside the same cluster is greater than $\bar{q}$ while for those in different ones it is less than this value. Choosing $\bar{q}' > \bar{q}$ the clusters split again into smaller subclusters with the same property as before but now with respect to the new scale $\bar{q}'$.

3.6. Ultrametricity tests

As we said the ultrametric nature of the space of pure states of the SK model was first found using the replica method[19]. However this calculation relies entirely on an ansatz which assumes that replica space itself is ultrametric[9]. Ultrametricity of the pure states is a direct consequence of ultrametricity in replica space. It is important then to have alternative ways to check its existence. The other two techniques we have at our disposal are the TAP equations and numerical simulations.

a) *TAP equations*. In this case the program to follow is: first one should find solutions of equation (3.2), once these are known their free energies are calculated. Only those with the lowest free energy should be kept and the length of all possible triangles calculated. Using the Boltzmann weights P_α and the triangles the corresponding histogram is evaluated. Alternatively one should define a quantity such that ultrametricity could be detected easily and use it instead of $P(q_1,q_2,q_3)$.

The test is repeated for several values of N to see how the property scales with the system size.

This program is difficult to implement. As we said the SK model is an NP-complete problem; because of this to be able to find the lowest energy states the system size should be kept small. At finite temperature equations (3.2) are not easy to solve, as a consequence one cannot obtain a set of solutions large enough to construct a good histogram for $P(q_1,q_2,q_3)$ [21].

Another problem, common to both techniques, is that because of finite size effects it is hard to distinguish ultrametricity from the triangular inequality[20]. As an example, for an isosceles triangle such that its sides d_1,d_2 and d_3 satisfy

$$d_1 = d_2 \quad , \quad d_3 < d_1 \tag{3.31}$$

at finite N one would see

$$d_1 = d_2 \pm 0\ (1/N^{\lambda}). \tag{3.32}$$

On the other hand the triangular inequalities are

$$|d_2-d_3| \le d_1 \le d_2+ d_3 \ , \tag{3.33}$$

from where we see that a small d_3 could be confused with a finite size correction.

A possible ultrametricity test consists in the following procedure[22]: first fix q_3 and choose $q_2<q_3$, then consider all the triangles with any q_1 and evaluate the quantity

$$M = \frac{\int_0^{q_2} dq_1\ P(q_1,q_2,q_3)(q_1-q_2)^2}{\int_0^{q_2} dq_1\ P(q_1,q_2,q_3)} \tag{3.34}$$

Finally compare the ultrametric hypothesis with a null or "no structure" hypothesis defining the index

$$U = M/\left[\frac{\int_0^{q_2} dq_1 (q_1-q_2)^2}{\int_0^{q_2} dq_1}\right] \tag{3.35}$$

For an ultrametric set of states U should tend to zero as N increases while for a flat distribution it should vanish.

The results obtained for U are shown in table 1 for two system sizes (N= 32,64), q_3= 0.625 and several values of q_2. A general convergence to zero is observed although it occurs at a rate which depends on q_2.

TABLE 1

The ultrametricity index U for q_3= 0.625.

q_2	N = 32	N = 64
0.375	0.87	0.71
0.4375	0.47	0.24
0.5	0.33	0.23
0.5625	0.34	0.093
0.625	0.23	0.068

Actually this calculation was done[22] for a model slightly different from the SK model, which requires only zero temperature solutions of the TAP equations. Once we have the local magnetizations we assign to the solution the Boltzmann weight

$$P = \exp\ (\beta \sum_{(i,j)} J_{ij}\ m_i\ m_j) \tag{3.36}$$

However the P(q) for this model is very similar, although not identical, to the SK model one overlap distribution. In figures 6a and 6b we compare q(x) for both models.

b) *Numerical simulations*. This method consists in the simulation of three identical replicas. At regular intervals the three overlaps between them are evaluated and ordered according to their values: $q_{max} \ge q_{mid} \ge q_{min}$. Then, for a fixed q_{max} the

distribution of $\delta q = q_{mid} - q_{min}$ is constructed. R.N. Bhatt and A.P. Young[29] showed that this distribution is peaked at $\delta q = 0$, with a width which increases as $N^{-0.9}$.

3.7. Properties of the ultrametric structure of the SK model

Defining the weight of the I'th cluster as

$$\omega_I = \sum_{\alpha \in I} P_\alpha \tag{3.37}$$

the distribution of cluster weights

$$f_J(\omega,y) = \sum_I \delta(\omega-\omega_I) \tag{3.38}$$

can be easily obtained[10]. The function $y(q)$ is given by

$$y(q) = \overline{Y_J(q)} \tag{3.39a}$$

where

$$Y_J(q) = \int_q^1 P_J(q')\, dq'. \tag{3.39b}$$

The moments of the distribution are defined as

$$M_k = \int_0^1 d\omega\, \omega^k\, \overline{f_J(\omega,y)}\ , \tag{3.40}$$

these give the probability that k states are in the same cluster, as can be verified using equations (3.38) and (3.37). A simple calculation gives

$$M_k = \frac{\Gamma(y+k-1)}{\Gamma(y)\,\Gamma(k)} \tag{3.41}$$

which depends on temperature and external magnetic field only through the function $y(q)$.

The first moment, M_1, is the average weight. All of them are polynomials in $y(q)$, for instance

$$M_2 = y$$

$$M_3 = 1/2\ y\ (1+y) \tag{3.42}$$

The cluster weight distribution $\overline{f_J}$, obtained from its moments, is

$$\overline{f_J(\omega,y)} = \frac{\omega^{y-2}(1-\omega)^{-y}}{\Gamma(y)\,\Gamma(1-y)}\ , \tag{3.43}$$

integrating it over ω it can be seen that there is an infinite number of clusters, these are concentrated at $\omega=0$ but their weight is negligible.

3.8. Taxonomic analysis of ultrametric systems. Relevance of frustration.

We shall describe here how to study ultrametricity by explicit construction of the hierarchical tree[25,26]. Since in general this property becomes exact only in the large N limit it is not possible to assign a unique tree to a set of states of a finite size system. What can be done is to verify that two extreme procedures to determine a tree converge to the same answer as N increases.

These procedures have been used in biological taxonomy[27] but its application to any set of points suspected to be ultrametric is straightforward.

Let us first see which is the problem that arises when trying to built the tree associated with a system which is not exactly hierarchical. At the scale corresponding to the maximum overlap each state is a cluster by itself and the overlap between any pair of them is a well-defined quantity. As the scale gets smaller we have to decide how to define the overlap between a state and a cluster of states. For exact ultrametricity the problem does not appear because the state has the same overlap with any of the states in the cluster. But when this property holds only approximately the two extreme procedures which consist in defining this overlap systematically either as the smallest or as the largest one between the state and those already in the cluster will give different answers. These two procedures are called complete-linkage and single-linkage clustering respectively. In the large N limit they should converge to the same tree if ultrametricity becomes exact.

To quantify the difference between the two trees we need to describe their structure by convenient probability distributions[23]. This can be done using the distribution of cluster weights defined in equation (3.38), that is the probability $f(\omega,q)$ to have a cluster of weight ω when the tree is

cut at a scale q. If the distributions $f^{>}(\omega,q)$ and $f^{<}(\omega,q)$, obtained with complete-linkage and single-linkage clustering repectively, converge to a common distribution for large N then we shall say that the system becomes ultrametric in that limit.

We shall use the technique in systems where the N_s states are given the same weight. We consider three different examples:

a) a simple model which is explicitly ultrametric in the large N limit;

b) the SK model where N_s of the lowest energy states are given the same weight;

c) the GCP we already discussed in Section 2. The N_s states are quasi-optimal solutions.

As we already said a feature in common in these examples is that all the states included in the test are given the same weight. For the SK model this means to consider it outside its natural framework which is the Statistical Mechanics.

The purpose of this is that it might be interesting to investigate properties of the energy landscape such as correlations between the minima, influence of the Boltzmann weights in ultrametricity or sample to sample fluctuations, attraction basins, etc. The experience gained in this way will give not only a deeper insight into the properties of the SK model, but it might also help to understand how to built models with specific properties.

Since each of the N_s states has the same weight $1/N_s$, the weight of the I'th cluster containing n_I states is

$$\omega_I = n_I / N_s \qquad (3.44)$$

These have to be used in the equations of subsection 3.7 to evaluate the moments $M_k^{>}$ and $M_k^{<}$.

We proceed now to examine the three examples.

a) In this simple model states are generated with a branched stochastic process defined on the hierarchical tree[20,25]. Since for any pair of states this process is the same until their nearest common ancestor is reached, this guarantees that the states are organized according to the proposed hierarchy. As N increases fluctuations in the value of the overlaps at each node decrease as $1/\sqrt{N}$ and ultrametricity becomes exact.

For a regular tree such as the one of figure 7 the moments M_k can be obtained as the probability that k states are in the same cluster, for $q<\bar{q}$ they are

$$M_k(q) = 1/2^{k-1} \quad . \qquad (3.45)$$

The test was done[25] for regular trees with $\bar{q}_0=0$ and $\bar{q}_1=0.6$.

In figure 8 we show the moments $M_k^{<}$ and $M_k^{>}$ for two values of N. It is clear that as N increases their values change as expected, although the convergence seems to be slow.

In order to analyse the behaviour with N we compare in figure 9 the moments $M_2^{<}$, $M_2^{>}$ with the direct evaluation based on equations (3.42) and (3.39). The latter always stays between the single and complete clustering results. It can be seen that for small N (N=32, 64) the differences are large, the system size has to increase in about one order of magnitude to verify that all of them are approaching a common value.

A similar behaviour can be observed in Figure 10 for other moments, there we exhibit the ratio

$$d_k = \left[M_k^{<} - M_k^{>}\right] / 2^{k-1} \qquad (3.46)$$

at the scale q=0.5, as a function of N.

b) The moments $M_k^{<}$ and $M_k^{>}$ for the SK model are shown in figure 11. Again N=32 and 64. This figure is to be compared with figure 8. As in the first example the difference $M_k^{<} - M_k^{>}$ decreases as N increases indicating the existence of some degree of ultrametricity. However for the SK model it is much harder to evaluate these moments for N as large as 300 because of the problems we already mentioned in Section 2.

c) The GCP was already described in Section 2. We chose the following definition of distance between the configurations α and β

$$d(\alpha,\beta) = \frac{1}{N} \sum_{i=1}^{N} (1-\delta_{\alpha_i,\beta_i}) \qquad (3.47)$$

where the sum runs over all the vertices of the graph and α_i denotes the colour of node i in the configuration α.

This example allows us to verify that frustration is a necessary ingredient in ultrametric systems[28,29]. As we said in

Section 2 the GCP with three colours has frustration whereas with q=4 it is not frustrated. With the technique we are discussing it is possible to check that ultrametricity holds only for q=3[20].

Figure 12 shows the moments $M_k^<$ and $M_k^>$ for q=3. Again it must be compared with figures 8 and 11 corresponding to the other two examples. All of them have a similar behaviour with N, indicating that the 'differences between the single-linkage and complete-linkage clustering decrease.

The case q=4 is very different. This can be seen by comparing the distance distributions $P_q(d)$ for both values of q. $P_3(d)$ and $P_4(d)$ are shown in figures 13 and 14 respectively. While $P_3(d)$ extends over all values of d and does not change with N, $P_4(d)$ appears concentrated in a region about 0.77 (for N=70) and it changes with N becoming narrower as the system size increases. For q=3 the set of quasi-optimal solutions in very restricted and they are states with special properties. For q=4 there are many optimal (zero cost function) solutions and they are uncorrelated behaving as random four colour configurations. The overlaps are then expected to be 0.25, what is in agreement with the value $d \simeq 0.77$ corresponding to the peak in figure 13 for N=70. One can easily compute solutions for higher values of q and verify that $P_q(d)$ has a behaviour similar to $P_4(q)$ with a peak at d=1-1/q.

N=70.N_s=14

4 BRAIN MODELLING. MODELS OF ASSOCIATIVE MEMORY

4.1. The Hopfield model

Once the properties of the SK model were fully understood the application to other problems of what had been learnt after several years of research was almost immediate[2]. Brain modelling, in particular models of associative memories, is perhaps the most interesting example of this[3].

By the time the SK model was being understood J. Hopfield[30] had studied numerically a model of associative memory which made just the right hypothesis to allow the application of statistical mechanical techniques.

The basic elements of the models of associative memory which have been considered by physicists are two state nodes S_i (neurons) and synaptic connections J_{ij} which are the couplings between pairs of neurons. The number of nodes is usually large.

A neural network is a set of these N interconnected neurons. It is endowed with a dynamics which specifies how they change their state with time.

Although there has been some work done on neural networks as computing devices (in particular they have been applied to treat NP-complete problems[31]) we shall concentrate here on models of associative memories.

In general these systems work as follows: first a few patterns are memorized; then an input pattern, or initial neuron configuration, is given as a stimulus to the system. Under a given dynamics this state is successively modified until it reaches a fixed point. This new state is one of the previously memorized patterns or at least very close to it. In this way the stored information is retrieved well if the stimulus is not too different from a memorized state. It is said that the system works by association. Since contrary to what happens with standard computer memory here the information is not retrieved by giving a physical address but by its content itself they are also called content addressable memories.

The Hopfield model may be considered as a standard model. It was the first to be solved and many of the later networks were proposed by modifying it usually with the purpose to obtain more realistic systems. Let us then begin with a brief description of this model.

First we have to encode the information to be stored. This will be done by selecting P patterns $\{\xi_i^{(\alpha)}\}$, $\alpha=1,\ldots,P$, at random in such a way that each component $\xi_i^{(\alpha)}$, $i=1,\ldots,N$, is +1 or -1 with the same probability. In this case their overlaps

$$m^{(\alpha\beta)} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^{(\alpha)}\,\xi_i^{(\beta)} \qquad (4.1)$$

are $O(1/\sqrt{N})$.

Then we have to determine the way the information will be stored. Given the set of patterns $\{\xi_i^{(\alpha)}\}$, a learning rule establishes the relation between them and the synaptic connections. We shall take

$$J_{ij} = \frac{1}{N}\sum_{\alpha=1}^{P} \xi_i^{(\alpha)}\,\xi_j^{(\alpha)} \qquad (4.2)$$

which is known as the Hebb's rule. It is supplemented with the condition $J_{ii}=0$. Notice that in this model the synaptic matrix is symmetric. Besides, the independent quenched random variables are not the J's but the patterns (ξ^{α}).

We still have to define the dynamics, i.e. the way the memorized patterns are retrieved. Denoting with (S(t)) the neuron configuration at time t the dynamics of the Hopfield model coincides with the one defined in equation (2.8) in connection with the SK model.

This dynamics implies, as we already know, that the energy function

$$H = -\sum_{(ij)} J_{ij} S_i S_j \qquad (4.3)$$

decreases systematically until it reaches a fixed point. If the network works properly this final state should be close to one of the memorized patterns if the input is not too far from it. Equation (4.3) is identical to the hamiltonian of the SK model, let us remark however that while for the spin glass case the long range interactions are not realistic in the brain these connections are frequent. However a more realistic model has to introduce dilution: the order of magnitude of the number of neurons in the brain is 10^{11}, but each of them is connected only to about 10^{4} other neurons.

Let us see under which conditions the model is able to work as an associative memory. The memorized patterns have to be stable under the dynamics, if the network is in state $(\xi^{(\alpha)})$ the activation field defined as in equation (2.6) is

$$h_i^{(\alpha)} = \sum_{j=1}^{N} J_{ij}\, \xi_j^{(\alpha)} \qquad (4.4)$$

$$= \sum_{\beta=1}^{N} m^{(\alpha\beta)}\, \xi_i^{(\beta)}$$

where we used the Hebb's rule and the definition (4.1). Separating the term $\beta = \alpha$

$$h_i^{(\alpha)} = \xi_i^{(\alpha)} + \text{noise}, \qquad (4.5)$$

the noise term is the contribution from the other (P-1) patterns.

It is easy to estimate, since it can be thought of as a random walk with (P-1)N steps it is $O(\sqrt{(P-1)N})$. Only when this term becomes large enough to compete with the signal contribution the network will destabilize. This occurs when

$$P = \alpha N . \qquad (4.6)$$

There is a value of α where the memory deteriorates. This value, together with a description of the way this deterioration occurs was determined analitically by Amit et al. in a series of very interesting papers[22-24].

They considered the Hopfield model at finite temperature, the zero temperature dynamics defined by equation (2.8) is now generalized to a stochastic process such that neurons are updated with the probability

$$P(\tilde{S}_i) \alpha \frac{\exp(-\beta h_i \tilde{S}_i)}{\exp(-\beta h_i \tilde{S}_i) + \exp(\beta h_i \tilde{S}_i)} \qquad (4.7)$$

where β is the inverse of the temperature.

Retrieval is now measured by the order parameter

$$m^{(\alpha)} = \frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle \xi_i^{(\alpha)} \quad , \quad \alpha = 1, \ldots, P \qquad (4.8)$$

which detects the existence of a thermodynamic phase characterized by good retrieval of the state $(\xi^{(\alpha)})$.

Since we are not interested in the properties of a single sample but in those which are common to most of them the disorder is quenched, this means that we have to average the free energy over the stochastic process used to select the patterns. Denoting this average with a bar

$$\bar{f} = -\frac{k T}{N} \overline{\ln Z} \qquad (4.9)$$

The difference with the SK model lies in the way the couplings J_{ij} are chosen. This is important, while the number of disordered variables in the spin glass model is $O(N^2)$, it is only PN in the neural network. Only in the saturation regime, defined by equation (4.6), the two may become equal. As we shall see, the spin glass phase is the global minimum of $\bar{f}$ only for α large enough.

The techniques we described in the previous section can be applied to solve the Hopfield model. The phase diagram in the

plane (T,α) has an interesting structure. Apart from the P order parameters (4.8) we have

$$r = \alpha^{-1} \sum_{\beta>s}^{P} \overline{(m^{(\beta)})^2} \qquad (4.10a)$$

and

$$q = \frac{1}{N} \sum_{i}^{N} \overline{\langle S_i \rangle^2} \qquad (4.10b)$$

In equation (4.10a) s is the number of condensed patterns. Their values are determined by saddle point equations giving the exact solution of the model[33].

Above the line $T_g = 1 + \sqrt{\alpha}$ the system is in a paramagnetic phase where both q and $m^{(\alpha)}$ are zero. Below it the spin glass phase ($q \neq 0$) is always stable. It is also the global minimum in the region between T_g and another critical line $T_c(\alpha)$. Below $T_c(\alpha)$ the retrieval states are the global minima, characterized by $m^{(\alpha)} \neq 0$. However these are already stable in a portion of the spin glass phase. Although stability occurs at a temperature $T_M(\alpha)$ which is not associated with a thermodynamic transition it is important for the retrieval properties because its presence alters the dynamic behaviour. If initially the neuron configuration is close enough to one of the memorized patterns its retrieval will be good even when the global minimum is the spin glass state.

These results were obtained using the replica method under the assumption of a replica symmetric saddle point[33]. However at low temperature the entropy becomes negative, what indicates that replica symmetry has to be broken. The effect of the breaking is smaller than in the SK model. For the Hopfield model the zero temperature entropy in the spin glass phase is[34] $S(0) \simeq -0.07$ to be compared with $S(0) = -0.16$ (equation (3.18)) for the SK mean-field solution.

Apart from correcting the entropy the breaking of the replica symmetry also solves another problem. The critical value of α at $T=0$ obtained from numerical simulations is $\alpha_c \simeq 0.145$ while the symmetric solution yields $\alpha_c = 0.138$. The value obtained with replica symmetry breaking is in agreement with the numerical one[35].

4.2. Patterns with hierarchical correlations

Here we discuss the interesting case where the memorized patterns are organized hierarchically in categories[36,37,38]. Members of the same category have overlaps greater than a scale $\bar{q}$ while those in different ones have overlaps which are less than this value. This means that the patterns to be learnt form an ultrametric space.

As we did for the Hopfield model, we have to define[36]:

a) the way the set of ultrametric patterns is generated,

b) the learning rule,

c) the dynamics

a) The structure of the ultrametric tree has to be given first. We shall restrict the discussion to regular trees such as the one of figure 7, in this case their structure is completely defined by the position of the levels of branching and the number of categories at each of these levels. For a two-level regular tree the parameters are $\bar{q}_0$, $\bar{q}_1$, the number of categories N_a and the number r of patterns in each category. In the example given in figure 7, $N_a=3$ and $r=4$.

The patterns are generated following a branched stochastic process defined on the tree. Different stochastic processes can be chosen. One possibility is to use real variables; at $q=\bar{q}_0$ N_a variables $Y_1^{(\alpha)}$ ($\alpha = 1,\ldots,N_a$) are selected using a distribution $P_1(Y_1|Y_0)$ with mean value Y_0 and dispersion σ. At $q=\bar{q}_1$ another rN_a variables $Y_2^{(\alpha,\beta)}$ ($\beta = 1,\ldots,r$) are obtained with N_a probability distributions $P_2(Y_1|Y_1^{(\alpha)})$, such that their mean values are the N_a real numbers obtained at the previous level. At the last level, the second in our example in figure 7, the state of neuron i in the pattern $(S^{(\alpha\beta)})$ is given by

$$S_i^{(\alpha\beta)} = \text{sign}\,[Y_2^{(\alpha\beta)}] \qquad \begin{matrix}\alpha = 1,\ldots,N_a \\ \beta = 1,\ldots,r\end{matrix} \qquad (4.11)$$

here α and β denote the category and the pattern within it respectively.

b) Since patterns in the same category have a non-zero average overlap the learning rule given in equation (4.2) does not work. This is easy to see, in the present notation the Hebb's rule reads

$$J_{ij} = 1/N \sum_{\alpha=1}^{N_a} \sum_{\beta=1}^{r} S_i^{(\alpha\beta)} S_j^{(\alpha\beta)}, \qquad (4.12)$$

in order to check the stability of pattern $(S^{(\alpha_o\beta_o)})$ under the dynamics we evaluate the associated activation field

$$h_i^{(\alpha_o\beta_o)} = (1-\bar{q})\, S_i^{(\alpha_o\beta_o)} + \bar{q}\, r\, S_i^{(\alpha_o)} \qquad (4.13)$$

where we defined the category state as

$$S_i^{(\alpha)} = \frac{1}{r} \sum_{\beta=1}^{r} S_i^{(\alpha\beta)} \qquad (4.14)$$

Here we took $\bar{q}_o \sim 0$ and $\bar{q}_i = \bar{q}$ and neglected the contribution from categories other than α_o. The second term can be cancelled exactly by the following modification of the learning rule[36]

$$NJ_{ij} = \sum_{\alpha=1}^{N_a} \sum_{\beta=1}^{r} S_i^{(\alpha\beta)} S_j^{(\alpha\beta)} - \lambda \sum_{\alpha=1}^{N_a} S_i^{(\alpha)} S_j^{(\alpha)} \qquad (4.15)$$

with

$$\lambda = r^2\bar{q}/(1+\bar{q}\,(\lambda-1)), \qquad (4.16)$$

alternatively it could be chosen as a new parameter restricted to take values such that the retrieval properties[38] are not spoiled.

An analysis of the thermodynamics of the model indicates that the retrieval properties undergo a transition at the same value that the Hopfield model[38]. More precisely the order parameters

$$\mu^{(\alpha\beta)} = 1/N \sum_{i=1}^{N} \left[\xi_i^{(\alpha\beta)} - \xi_i^{(\alpha)}\right] \langle S_i \rangle \qquad (4.17a)$$

$$m^{(\alpha)} = 1/N \sum_{i=1}^{N} \xi_i^{(\alpha)} \langle S_i \rangle \qquad (4.17b)$$

are both zero above $\alpha_c = 0.145$. This means that according to this analysis the network cannot recognize neither the memorized pattern (because $\mu^{(\alpha\beta)}=0$) nor the category (because $m^{(\alpha)} = 0$) for $\alpha>\alpha_c$.

However the mean field techniques used in reference (38) can detect only free energies O(N), while for $\alpha>\alpha_c$ the system is in a spin glass phase characterized by metastable states separated by

barriers which are smaller than O(N). Following the dynamics, equation (2.8), the system will be trapped in states which are stable at least under one spin-flip. This produces remanence effects which are well-known in spin glass models[39]. They are also present in the Hopfield model[34].

If the effects were such that the correct category is distinguished by the metastable fixed points, the network could recognize it even when the memorized patterns are confused.

In order to describe the results of a numerical simulation of the dynamics of this model[40] we define the following quantities. The numerical experiment takes the state $(S^{(\alpha_o\beta_o)})$ as the initial pattern, then under the dynamics it will evolve to a fixed point (S). The overlap between this state and one of the stored patterns is

$$m^{(\alpha\beta)} = 1/N \sum_{j=1}^{N} S_j^{(\alpha\beta)} S_j \,. \qquad (4.18)$$

Its overlap with the category $(S^{(\alpha)})$ is

$$m^{(\alpha)} = 1/N \sum_{j=1}^{N} S_j^{(\alpha)} S_j \,. \qquad (4.19)$$

In particular we are interested in $m^{(\alpha_o)}$ and

$$M = \max\,[m^{(\alpha)}], \qquad \alpha \neq \alpha_o \,. \qquad (4.20)$$

Similary to the parameter $\alpha = P/N$ we define for the categories

$$\gamma = Na/N \,. \qquad (4.21)$$

In reference 40 it was found that for $\gamma=0$ there is a critical value of α_c where a change in the retrieval properties occurs. For $\alpha<\alpha_c$ both memorized patterns and categories are not confused. For $\alpha>\alpha_c$ most of the $m^{(\alpha_o\beta)}$ $(\beta=1,\ldots,r)$ are non-zero while the category overlaps $m^{(\alpha)}$ are zero with the exception of $m^{(\alpha_o)}$. In figure 15 we show its average value. We see that the correct category is not confused with the others; while $m^{(\alpha_o)}$ drops to a finite value M is always zero.

We must remark however that the output state is not the category itself or a state close to it; in order to recognize it we need a list containing all the categories and their overlaps with the fixed point have to be computed. Since the number of

categories is finite this can be easily done.

Since each category must contain at least two patterns the relevant region in the plane (α,γ) is bounded by the line $\alpha=2\gamma$. On this line $m^{(\alpha_0)}$ and M have a behaviour very different from the one on $\gamma=0$. In figure 16 we see that both averages are non-zero and converge to a common limit as α increases. The network loses its capability to recognize the category. Even if we argue that there is a finite difference between the two averages the effect of fluctuations or numerical errors will prevent us from distinguishing the correct category. Besides, the fact that γ is finite (i.e. N_a is O(N)) implies an infinite list of categories, making recognition impractical.

Good category retrieval for $\gamma=0$ suggests that the metastable states have an organization which respects the hierarchy introduced in the memorized patterns. This can be checked by evaluating the number $\mathcal{N}(m, m_1, m_2)$ of one spin-flip stable states with overlap m with a pattern in category 1 and overlaps m_1 and m_2 with categories 1 and 2 respectively[41].

The technique to count metastable states was first developed for the SK model[13]. Application to neural networks have also been done[42,69]. $\mathcal{N}(m, m_1, m_2)$ was obtained in reference 41. The total number of metastable states is

$$\overline{\mathcal{N}} = \int_0^\infty \prod_{i=1} d\lambda_i \; \mathrm{tr}_S \, \delta(h_i - \lambda_i S_i)$$

$$= \int dm \, dm_1 \, dm_2 \, \overline{\mathcal{N}}(m, m_1, m_2) \qquad (4.22)$$

where as usual the bar denotes an average over the quenched variables which in this case are those associated with the branched stochastic process used to generate the memorized patterns. The calculation was done for the particular case $N_a=2$. The main result is that $\mathcal{N}(m, m_1, m_2)$ is denominated by a saddle point with $m_2=0$ but $m_1 \neq 0$.

This is in agreement with the suggestion we just made that the metastable states themselves respect the hierarchy. However a more detailed analysis is necessary to make a conclusive statement about this.

4.3. Models of short term memory

A serious drawback of the Hopfield model is that for $\alpha > \alpha_c$ the memory deteriorates completely. It is not only unable to store more information but it also forgets what it had learnt before. A different type of models avoid this problem, although they do it at the cost of a smaller capacity.

In these models the network can always learn new information but only the last $k = xN$ of them are retrieved well. It works as a short-term memory[44-46].

A way to achieve this is to saturate the synapsis[44]. According to Hebb's rule the synaptic strengths change linearly with the contribution of each new pattern

$$J_{ij}(t+1) = J_{ij}(t) + 1/\sqrt{N}\; \xi_i^{(t+1)} \xi_j^{(t+1)}, \qquad (4.23)$$

here the patterns are labelled according to the time when they are memorized. In general one can take a non-linear relation

$$J_{ij}(t+1) = f\,(J_{ij}(t) + 1/\sqrt{N}\; \xi_i^{(t+1)} \xi_j^{(t+1)}) \; . \qquad (4.24)$$

A short-term memory model can be obtained taking the function[44]

$$\begin{aligned} f(x) &= -A \qquad && \text{for } x < -A \\ f(x) &= x \qquad && \text{for } -A < x < A \\ f(x) &= A \qquad && \text{for } x > A \end{aligned} \qquad (4.25)$$

where A is given, for instance, by the value of the synaptic strengths where the standard Hopfield model loses its retrieval properties. Since the J_{ij} are $O(\sqrt{\alpha})$ this means $A \simeq \sqrt{\alpha_c} \simeq 0.4$. With this model good retrieval is obtained for $x < 0.04$. The capacity has a maximum for $A \sim 0.35$.

Short-term memory models have been also studied by other authors. Following reference 45 the synapsis are modified according to

$$J_{ij}(t+1) = \lambda\,(J_{ij}(t) + \varepsilon/N\; \xi_i^{(t+1)} \xi_j^{(t+1)}) \; , \qquad (4.26)$$

here ε is the acquisition amplitude of each new pattern. The model, closely related to the previous one, does not allow the synapsis to grow indefinitely. The following condition has to be fulfilled:

$$\overline{J_{ij}^2} - \overline{J_{ij}}^2 = 1/N \; , \qquad (4.27)$$

which determines $\lambda = (1 + \varepsilon^2/N)^{-1/2}$.

This model is a member of a family of neural networks which can be solved analitically[45]. Its capacity has a maximum at an optimum value of ε: $\varepsilon_{op} = 4.108$; the corresponding value of α is $\alpha_{op} = 0.0489$. The acquisition amplitude cannot be smaller than $\varepsilon_c = 2.465$. These values are in agreement with previous numerical estimates obtained in reference 46.

4.4. Limit cycles

A naive way to memorize a cycle of length ℓ in the synaptic strengths is

$$J_{ij} = 1/N \sum_{\alpha=1}^{\ell} \xi_i^{(\alpha+1)} \xi_j^{(\alpha)} \qquad (4.28)$$

($\xi_i^{\ell+1} = \xi_i^{(1)}$). It was already noticed by Hopfield[30] that this does not work. The reason is that the pattern $\{\xi^{(\alpha)}\}$ cannot stabilize before the system goes to the next state in the cycle. The solution to this problem consists in introducing two different time scales[47]. In the short one the current pattern stabilizes while it is in the long time scale when the transition to $\{\xi^{(\alpha+1)}\}$ occurs. This requires two different synaptic matrices

$$J_{ij}^{(S)} = 1/N \sum_{\alpha=1}^{p} \xi_i^{(\alpha)} \xi_j^{(\alpha)} \qquad (4.29a)$$

$$J_{ij}^{(L)} = \lambda/N \sum_{\alpha=1}^{\ell} \xi_i^{(\alpha+1)} \xi_j^{(\alpha)} \qquad (4.29b)$$

The updating is done following the same rule as in equation (2.8) but now the activation field has two components

$$h_i = h_i^{(S)} + h_i^{(L)} \qquad (4.30)$$

where

$$h_i^{(S)}(t) = \sum_{j=1}^{N} J_{ij}^{(S)} S_j(t) \qquad (4.31a)$$

$$h_i^{(L)} = \sum_{j=1}^{N} J_{ij}^{(L)} \left[\int_{\infty}^{t} dt' \, \omega(t-t') S_j(t') \right] \qquad (4.31b)$$

The monotonously decreasing function ω has a characteristic time τ. In this way the network keeps memory of the previous neuron states. The parameter λ is chosen such that the transition to the pattern $\{\xi^{(\alpha+1)}\}$ in the cycle does not occur before the

network stayed at $\{\xi^{\lambda}\}$ during a time $O(\tau)$.

An interesting application of this model was proposed by D.K. Kleinfeld and H. Sompolinsky[48] who used it to generate the rhythmic swimming movements of the *Tritonia Diomedea*. Biological studies[49] of this mollusk have determined the structure of the group of neurons which generates the neural activity necessary to produce the right movements. This set of neurons is the central pattern generator of the *Tritonia Diomedea*. The firing patterns that the four premotor interneurons produce when swimming are known[49]. Using these in formulae (4.29) D.K. Kleinfeld and H. Sompolinsky could predict a synaptic matrix with short and long time scale components which are in good agreement with the experimental ones. Most of the fast synaptic components predicted by the theory are observed and have the right signs. For the slow components the theory predicts more synapsis than those which have been observed. However the authors of reference 48 argue that the experimental values can be taken as a dilute system. Since neural networks are robust under dilution both the theoretical model and the real system exhibit similar behaviours.

Probably, the reason why the theoretical model is able to describe the central pattern generator of the *Tritonia Diomedea* is the simplicity of this system. There is no reason to expect that a more complicated network could be explained by this simple model. A detailed study of the *Tritonia* as well as other organisms would be necessary to elucidate this point.

4.5. Asymmetry

The symmetry of the synaptic connections adopted by the Hopfield model is far from being realized in biological systems. In particular, if neuron i sends its axon to the dendrites of neuron j the inverse is not necessarily true, a case usually called asymmetric dilution.

The number of metastable states is easily calculated using the techniques of reference 13 for several types of asymmetry[49]. Surprisingly, the number of non-retrieval states is not appreciably altered with respect to the symmetric model. This, however, does not say everything about the dynamics since the size and shape of the attraction basins is also important. It was found in reference 50 that in spite of the existence of these metastable states the time necessary to reach them grows exponentially with

the size of the system. The dynamic behaviour looks chaotic.

Apart from the description of models with asymmetric synaptic matrices a cognitive role of asymmetry was suggested by G. Parisi[51]. When the dynamics reaches a fixed point indicating that something has been recognized there is nothing in the Hopfield model which guarantees that it is one of the memorized patterns. Spin glass states are always present and they may confuse the system since there is no way to distinguish them from the retrieval states. Then asymmetry helps: since those spurious states are replaced by chaotic trajectories the problem is automatically avoided.

5. CONCLUSIONS

Some of the most relevant aspects of the mean-field theory of spin glasses, the SK model, have been reviewed in these lectures. Neural networks have also been discussed from the point of view of Statistical Mechanics.

Frustration and disorder are responsible for the interesting properties of spin glass models. For the mean-field model ergodicity is broken in the thermodynamic limit, a fact related to the necessity of breaking the replica symmetry.

The main properties of the structure of pure states have been understood. While two state correlations give a non-trivial one overlap distribution $P(q)$, three state correlations produce a three overlap distribution $P(q_1,q_2,q_3)$ which reveals the ultrametric character of the space of pure states. Different tests to detect the existence of ultrametricity have been discussed.

The role of frustration in the validity of this property has been emphasized. This was done by analyzing the graph colouring problem for three and four colours; these two cases differ in that the first presents frustration while the second does not. Correspondingly a test based on the explicit construction of the hierarchy shows that only the first has some degree of ultrametricity.

The existence of a low temperature phase similar to the one we discussed for the SK model in finite dimensional spin glasses is under discussion[52]. This sheds some doubts on the relevance of the SK model to understand real spin glasses and the need to look for other solvable models has been pointed out[53]. In spite of this

possibility the SK model is interesting by itself and as we said it was taken as a prototype of systems with disorder and frustration and applied to many other areas of Science.

Neural networks is perhaps the most intensively studied. Once formulated by J. Hopfield, a model of fully connected neurons with symmetric synapsis was easily generalized and solved by D. Amit et al. using techniques already developed to study the SK model. The effects of non-linearities, correlated patterns, dilution and asymmetry have been introduced later.

Actually the field of neural networks is much wider than the approach described in these lectures and it did receive important contributions from many fields other than Statistical Mechanics. W.S. Mc Culloch and W.A. Pitts defined networks of formal neurons and they showed that these are general computing devices[54]. Learning was introduced by D. Hebb[55] who proposed the rule which has been the basis of our discussion in Section 4.

Although we have discussed only systems which work as associative memories, there are other neural networks which are able to solve different tasks.

The capability of the network to solve them depends crucially on its architecture, that is the way neurons are interconnected. One of the simplest machines is the perceptron[56]. It consists of two layers of neurons with feed-forward processing; the outer world determines the state of the first or input layer, this information is processed by neurons in the output layer. These are linear threashold units. The task the machine has to solve is defined by a mapping from the input to the output states. The interlayer couplings are supposed to be adjusted in such a way that the network solves the task. In the perceptron there are not couplings between neurons in the same layer. An interesting property is that there is a convergence theorem which states that whenever the perceptron is able to solve a task then the couplings can be found by a simple gradient algorithm[57]. Unfortunately there are many problems which cannot be solved by the perceptron[57]. A trivial example is given by the mapping corresponding to the exclusive-OR; the perceptron cannot make the correct classification because the problem is not linearly separable. This machine was abandoned because of these limitations.

More recently there was a reactivation in the field[58]. While

for the original perceptron a multilayer system was not interesting (since the neurons are taken as linear units it can always be reduced to two layers), a network made of several layers of non-linear units is a richer structure which may allow to go beyond the scarce capabilities of the perceptron. The existence of a non-linear mapping solving an arbitrary task in guaranteed by a theorem due to Kolmogorov[59]. It states that a function of N variables in the domain $[0,1]^N$ can be expressed in terms of a layered structure of functions of a single variable. Unfortunately it says nothing about these functions and the determination of the architecture is purely empirical.

These networks have been applied with some success to explain features of language acquisition[58], early vision[60], etc.. Most of these applications have been motivated by the ability of the non-linear networks to recognize the features of a task after a set of examples has been presented during a learning session. The network itself determines the algorithm to solve the problem.

Up to now physicists did not contribute appreciably to the analysis of these networks, perhaps because they do not seem to be tractable analytically and it is difficult to make statements valid not just to a particular network but to a generic one.

ACKNOWLEDGEMENT

I would like to thank my collaborators S. Bacci, E. Miranda and G. Mato for many comments. Special thanks are given to M.A. Virasoro for invaluable discussions.

I thank Raquel Cohen for efficient typing of the manuscript.

REFERENCES

1) D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. 35 (1975) 1792; S. Kirkpatrick and D. Sherrington, Phys. Rev. B17 (1978) 4384.

2) M. Mézard, G. Parisi and M.A. Virasoro, Spin Glass Theory and Beyond (World Scientific, Singapore, 1987).

3) D.J. Amit, The properties of models of simple neural networks, in: Heidelberg Colloquium on Spin Glasses, eds. J. van Hemmen and I. Morgenstern (Springer, Berlin, 1987).

4) S.A. Kauffman, J. Theor. Biol. 22 (1969) 437.

5) B. Derrida and H. Flyvbjerg, J. Phys. A19 (1986) L-1003; E.N. Miranda and N. Parga, Europhys. Lett. 6 (1988) 297; E.N. Miranda and N. Parga, J. Phys. A21 (1988) L-357.

6) S.F. Edwards and P.W. Anderson, J. Phys. F5 (1975) 965.

7) G. Toulouse, Comm. Phys. 2 (1977) 115.

8) M.R. Garey and D.S. Johnson, Computer and Intractability (Freeman, New York, 1979).

9) G. Parisi, Phys. Rev. Lett. 43 (1979) 1754; J. Phys. A13 (1980) L-115; J. Phys. A13 (1980) 1887.

10) G. Parisi, Phys. Rev. Lett. 50 (1983) 1946.

11) N. Parga and G. Parisi, On the locally stable states of the Sherrington-Kirkpatrick model, ICTP-preprint IC/85/133 (1985).

12) D.J. Thouless, P.W. Anderson and R. Palmer, Phil. Mag. 35 (1977) 593.

13) A.J. Bray and M.A. Moore, J. Phys. C13 (1980) L-469; C. de Dominicis, M. Gabay, T. Garel and H. Orland, J. Physique 41 (1980) 923; F. Tanaka and S.F. Edwards, J. Phys. F10 (1980) 2471.

14) A.P. Young, Phys. Rev. Lett. 51 (1983) 1206.

15) N.D. Mackenzie and A.P. Young, Phys. Rev. Lett. 49 (1982) 301.

16) A.P. Young, Phys. Rev. Lett. 51 (1983) 1206.

17) J.R.L. de Almeida and D.J. Thouless, J. Phys. A11 (1978) 983.

18) D.J. Thouless, J.R.L. de Almeida and J.M. Kosterlitz, J. Phys. C13 (1980) 3271.

19) M. Mézard, G. Parisi, N. Sourlas, G. Toulouse and M.A. Virasoro, J. Physique 45 (1984) 843.

20) R. Rammal, G. Toulouse and M.A. Virasoro, Rev. Mod. Phys. 58 (1986) 765.

21) K. Nemoto and H. Takayama, J. Magn. Magn. Mat. 54 (1986) 135.

22)N. Parga, G. Parisi and M.A. Virasoro, J. Physique Lett. 45 (1984) L-1063.

23)R.N. Bhatt and A.P. Young, J. Magn. Magn. Mat. 54 (1986) 191.

24)J. Vannimenus, G. Toulouse and G. Parisi, J. Physique 42 (1981) 565.

25)N. Parga, J. Physique 48 (1987) 499.

26)R. Ramal, J.C. Angles d'Auriac and B. Doucot, J. Physique Lett. 46 (1985) L-945.

27)R.R. Sokal and P.H.A. Sneath, Principles of Numerical Taxonomy (Freeman, 1963).

28)J.P. Bouchaud and P. Le Doussal, Europhys. Lett. 1 (1986) 91.

29)S. Bacci and N. Parga, Ultrametricity, frustration and the graph colouring problem, Bariloche preprint (1988), submitted to J. Phys. A.

30)J.J. Hopfield, Proc. Natl. Acad. Sci. USA 79 (1982) 2554.

31)J.J. Hopfield and D.W. Tank, Science 233 (1986) 625.

32)D.J. Amit, H. Gutfreund and H. Sompolinsky, Phys. Rev. A32 (1985)1007.

33)D.J. Amit, H. Gutfreund and H. Sompolinsky, Phys. Rev. Lett. 55 (1985) 1530.

34)D.J. Amit, H. Gutfreund and H. Sompolinsky, Ann. Phys. 173 (1987) 30.

35)A. Crisanti, D.J. Amit and H. Gutfreund, Europhys. Lett. 2 (1986) 337.

36)N. Parga and M.A. Virasoro, J. Physique 47 (1986) 1857.

37)V. Dotsenko, Physica A140 (1986) 410.

38)M.V. Feigelman and L.B. Ioffe, International J. of Mod. Phys. B1 (1987) 51.

39)W. Kinzel, Z. Phys. B60 (1985) 205.

40)S. Bacci, J. Alfaro, C. Wiecko and N. Parga, Numerical analysis of a neural network with hierarchically organized patterns, Bariloche preprint, to appear in J. Physique (April 1989).

41)S. Bacci, G. Mato and N. Parga, On the metastable states of a neural network with hierarchically correlated patterns, Bariloche preprint (1989).

42)E. Gardner, J. Phys. A19 (1986) L-1047.

43)A. Treves and D.J. Amit, Jerusalem preprint RI/88/119 (1988).

44)G. Parisi, J. Phys. A19 (1986) L-617.

45)J.P. Nadal, G. Toulouse, J.-P. Changeux and S. Dehaene, Europhys. Lett. 1 (1986) 535.

46)M. Mézard, J.P. Nadal and G. Toulouse, J. Physique 47 (1986) 1457.

47)D. Kleinfeld, Proc. Natl. Acad. Sci. USA 83 (1986) 9469; H. Sompolinsky and I. Kanter, Phys. Rev. Lett. 57 (1986) 2861.

48)D.K. Kleinfeld and H. Sompolinsky, Associative neural network model for the generation of temporal patterns: theory and application to central pattern generators, ATT Bell Labs. preprint (1987), submitted to J. Neurosci.

49)P.A. Getting and M.S. Dekin, *Tritonia* swimming: a model system for integration within rhythmic motor systems, in: Model Neural Networks and Behavior, ed. A.I. Selverston (Plenum Press, 1985).

50)A. Crisanti and H. Sompolinsky, Jerusalem preprint RI/87/89.

51)G. Parisi, J. Phys. A19 (1986) L-675.

52)K. Binder and A.P. Young, Rev. Mod. Phys. 58 (1986) 801.

53)C. De Dominicis and P. Mottishaw, J. de Physique 47 (1986) 2021.

54)W.S. Mc Colluch and W.A. Pitts, Bull. Math. Biophys. 5 (1943) 115.

55)D.O. Hebb, The Organization of Behaviour (Wiley, New York, 1949).

56)F. Rosemblatt, Principles of Neurodynamics (Spartan Books, New York, 1962).

57)M. Minsky and S. Pappert, Perceptrons (MIT Press, Cambridge, Ma, 1969).

58)D.E. Rumelhart and J.L. Mc Clelland, Parallel Distributed Processing Vol.1 and 2 (Bradford Books, Cambridge, Ma, 1986).

59)A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR 114 (1957) 953.

60)R. Linsker, Proc. Natl. Acad. Sci. USA 83 (1986) 7508, 8779, 8390).

FIGURE CAPTIONS

Figure 1: A frustrated loop. The direction of the third spin was left unspecified, any of them will leave an unsatisfied bond.

Figure 2: A frustrated graph for the GCP. Three of its nodes have been painted with blue, red and green, the fourth node has to repeat one of these colours giving rise to an unsatisfied bond.

Figure 3: The basin of attraction for different system sizes:(x) corresponds to N=32, the dots to N=64, the triangles to N=128,(+) to N=256 and the open circles to N=512.

Figure 4: The one overlap distribution for the SK model. The figure is schematic. q_{EA} is the self-overlap at temperatureT.

Figure 5: A hierarchical tree. The states are located at the end of the branches. The position of the nearest common ancestor of two states is at the value of their overlap. The tree has been cut at two different scales indicated with the dashed line.

Figure 6a y 6b:The order parameter q(x) for a) N=32 and b) N=64 and two different values of δ for the model of reference 22. The curve corresponds to the predictions of reference 24. Notice that for a finite system only a finite number of values of the overlap are possible.

Figure 7: A regular tree. In this figure there are two levels of branching.

Figure 8: The moments of the cluster weight distributions of the trees obtained with single-linkage and complete-linkage clustering for the model ultrametric by construction. For N=32: single-linkage clustering (●), complete-linkage clustering (o); for N=64: single-linkage (+), complete- linkage (x), (at the scale q=0.5).

Figure 9: Comparison of the second moment of the cluster weight distributions of the trees obtained with single-linkage and complete-linkage and complete-linkage clustering with y(q) for q=0.5 as a function of N for the model which is known to be exactly ultrametric. The dots stand for the single-linkage result and the open circles for the complete-linkage. The crosses indicate the direct calculation of y(q).

Figure 10: The ratio of equation (3.46) (multiplied by 100)as a function of N for k=2 (●), k=3 (+), k=4 (o) and k=5 (x) (at the scale q=0.5).

Figure 11: The moments of the two extreme cluster weight distributions for the SK model. System sizes and symbols are as in figure 8.

Figure 12: The moments of the cluster weight distributions for single-linkage and complete-linkage clustering for the GCP with q = 3 at a scale d=0.4 and 14 states. For single-linkage clustering: N=25 (■),N=50 (●), N=70 (x) and for complete-linkage:N=25 (□),N=50 (o), N=70 (+).

Figure 13: P(d) for the GCP with three colours. N=25 and N_s= 50.

Figure 14: P(d) for the GCP with four colours. Crosses correspond to N=25 and circles to N=70. N_s=14.

Figure 15: The average overlap between the output state and the correct category for a tree with N_a=2, $\bar{q}$=0.5, N=500 (o) and N=1000 (●).

Figure 16: The empty circles are the average overlap between the output state and the correct category on the line $\alpha=2\gamma$. The black circles are the average maximum overlap $\bar{M}$ between the output state and the other categories on the same line. Again $\bar{q}$=0.5. N=1000. The dashed line indicates the classification errors.

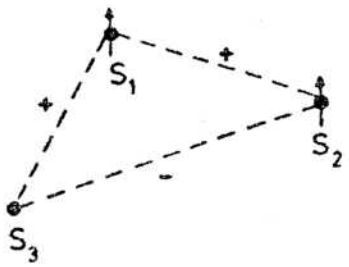

FIG 1

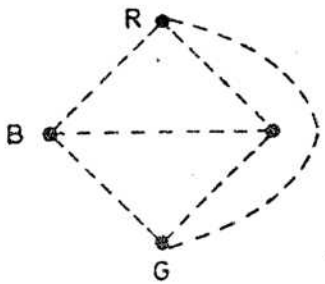

FIG 2

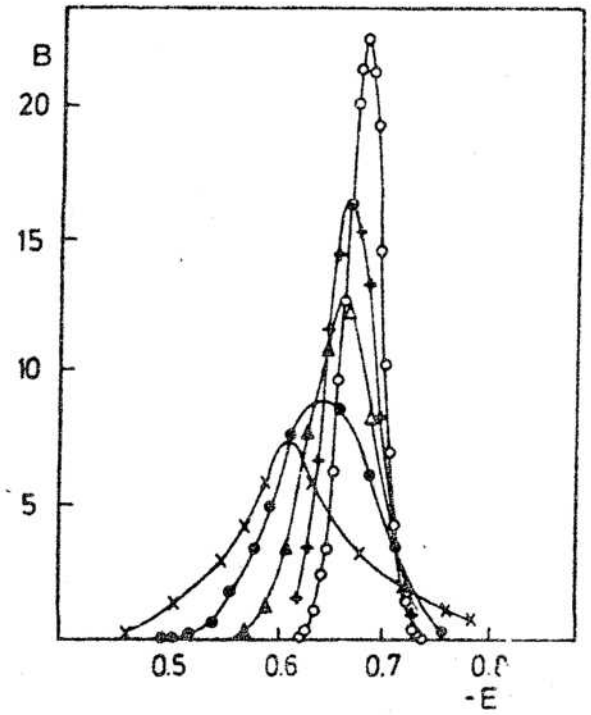

FIG 3

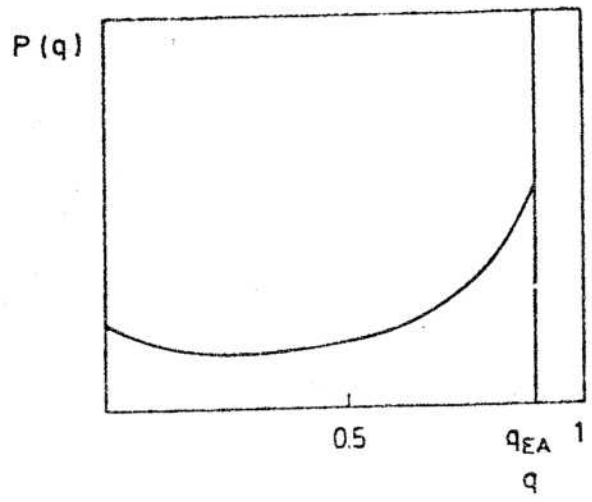

FIG 4

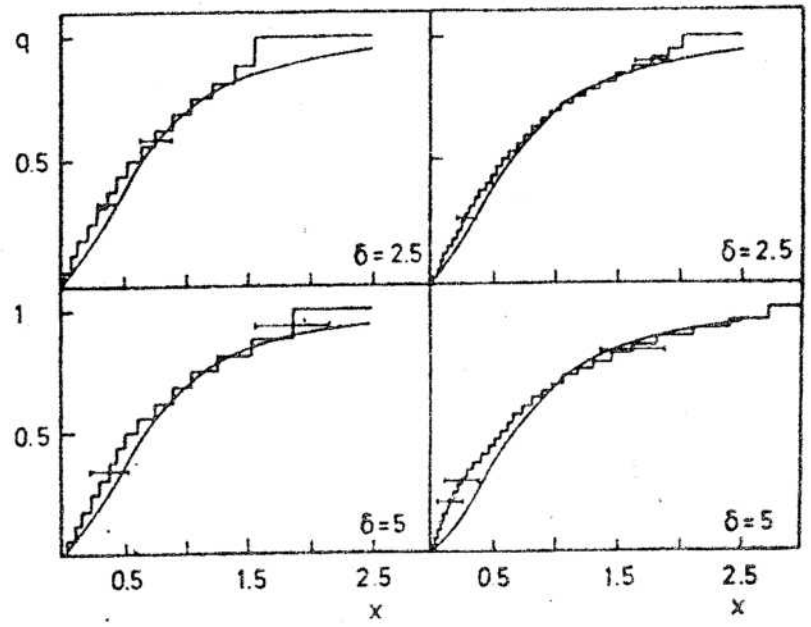

FIG 6

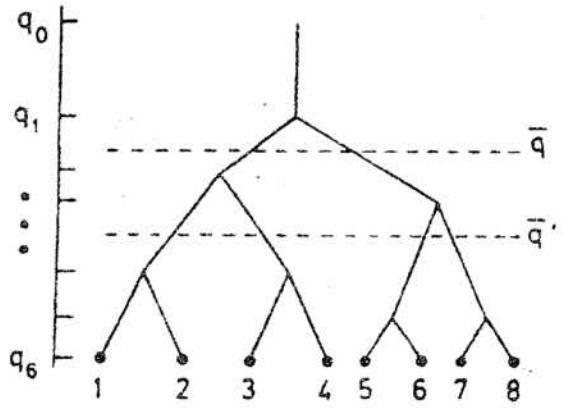

FIG 5

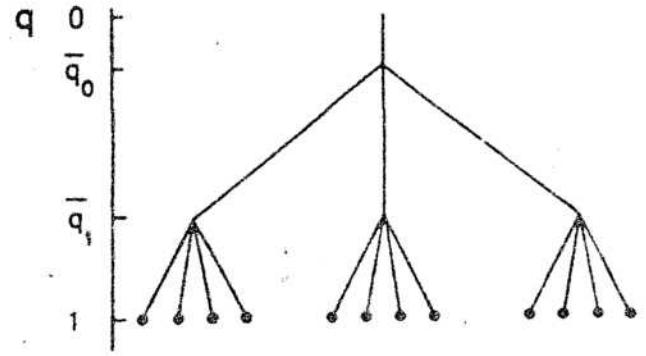

FIG 7

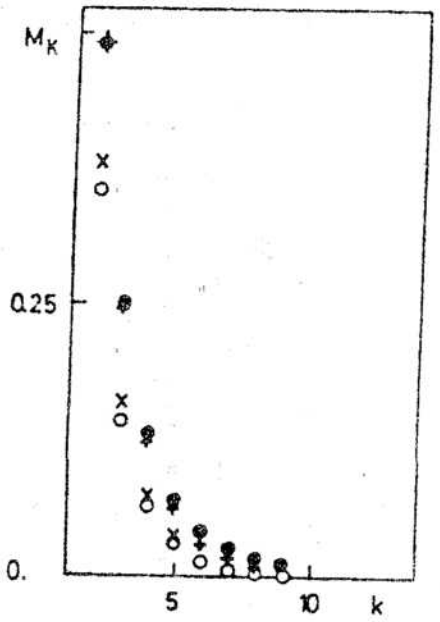

FIG 8

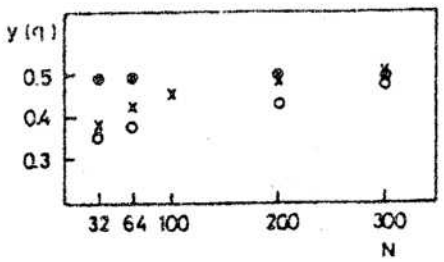

FIG 9

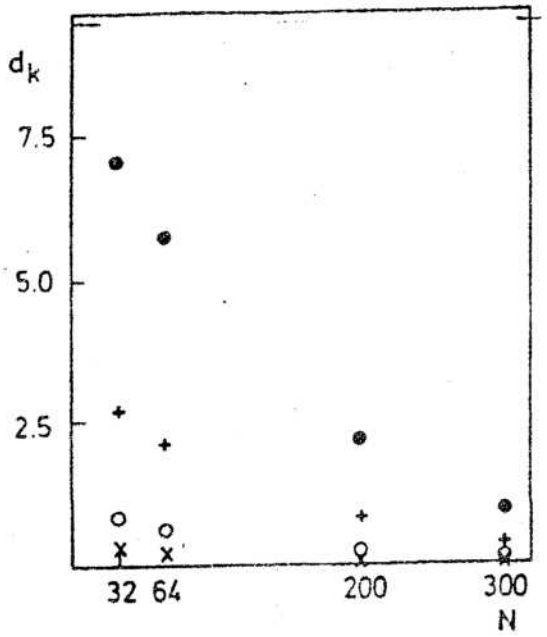

FIG 10

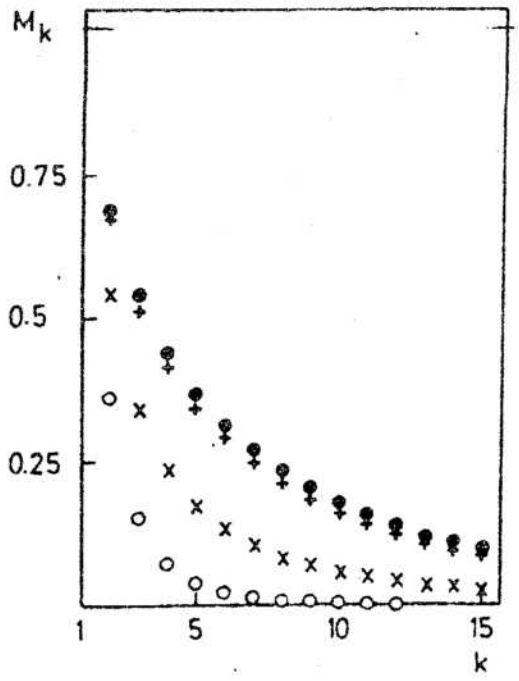

FIG 11

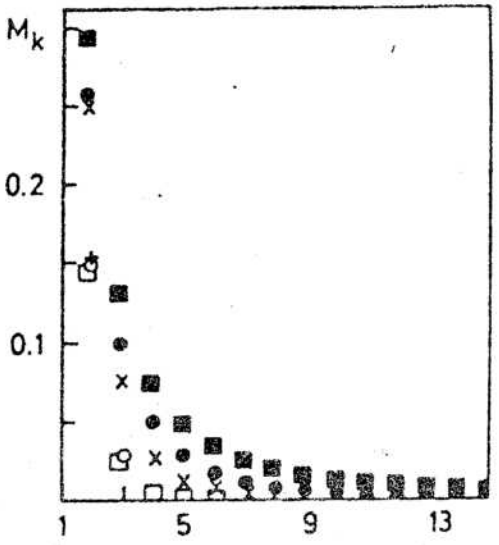

FIG 12

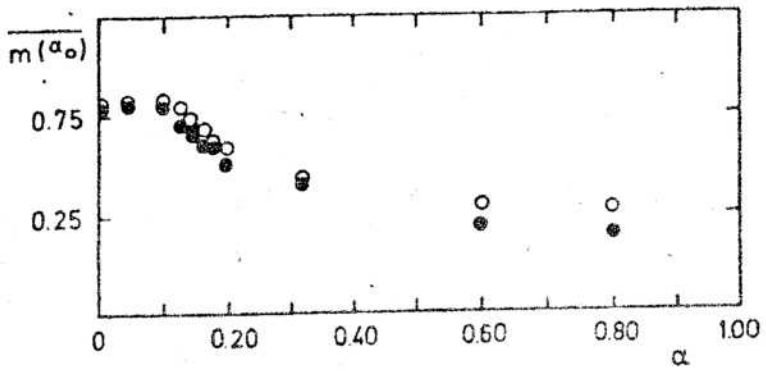

FIG 15

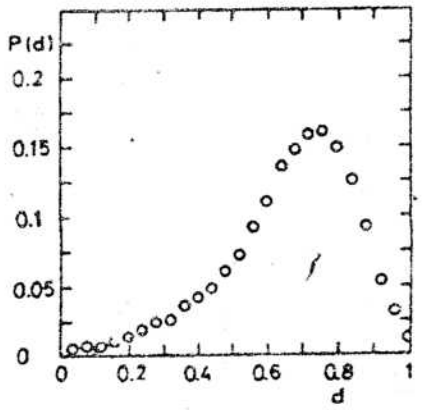

FIG 13

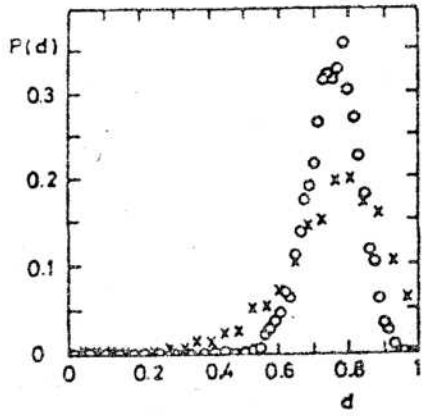

FIG 14

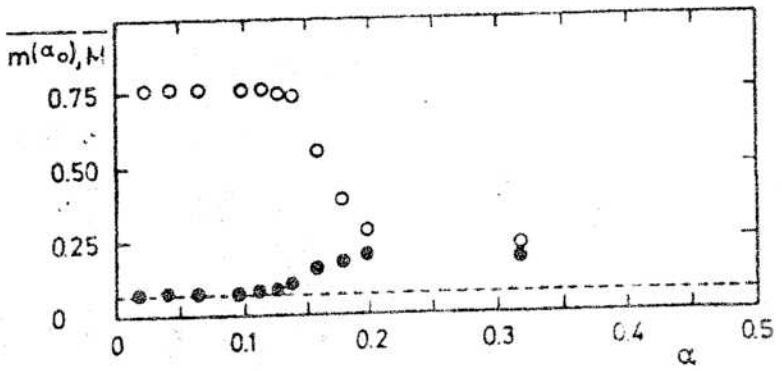

FIG 16