Hierarchical associative networks

1987 J. Phys. A: Math. Gen. 20 4449

(http://iopscience.iop.org/0305-4470/20/13/044)

J. Phys. A: Math. Gen. **20** (1987) 4449-4455. Printed in the UK

Hierarchical associative networks

C Cortes†, A Krogh† and J A Hertz‡

† Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark
‡ Nordita, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

Received 18 November 1986

Abstract. We study a modified Hopfield model of associative memory with a learning rule proposed by Personnaz *et al*, for the special case of ultrametrically correlated patterns. The formula for the synaptic strength tells the 'teacher' how much stress to put on details compared to averages at each level. It is still a local rule if we assume that the ultrametric correlation structure of the patterns is given *a priori*. The result is also given in terms of the Parisi function $q(x)$. In a special limit we get the same result as Parga and Virasoro.

1. Introduction

A popular model for investigating associative memory in neural networks is a system of N two-state spins (neurons) $S_i = \pm 1$. The spins are highly internally connected with synaptic strengths J_{ij} between spins i and j. The connections J_{ij} are constructed in such a way that certain spin configurations (patterns), representing the information to be stored, are dynamically stable states. If one wants to store uncorrelated patterns, the prescription for the J_{ij} is quite simple (Hopfield 1982), but this algorithm fails in more general cases. In this paper we study how to store patterns which are hierarchically correlated.

In the Hopfield model, as in many other models, a biologically motivated Hebb-type rule (Hebb 1949) is used. If the system has to learn a new pattern ξ_i^μ, $i = 1, \ldots, N$, the J_{ij} are modified by $\Delta J_{ij} = \xi_i^\mu \xi_j^\mu$. This simple rule is called local because it only involves ξ_i^μ and ξ_j^μ. Based on the pseudo-inverse method (Kohonen 1977, 1984) Personnaz *et al* (1985) have recently proposed a non-local learning rule, which has been studied and further developed by Kanter and Sompolinsky (1987). This model is capable of storing any set of linearly independent patterns, but when adding a new pattern one has to know all the previous stored patterns to calculate the new synaptic strengths.

In this paper we consider a class of problems which lies intermediate in generality between the artificial uncorrelated limit considered by Hopfield and the completely general case discussed by Personnaz *et al* and Kanter and Sompolinsky. This is the case of hierarchically (ultrametrically (Rammal *et al* 1986)) correlated patterns. This is of special interest for two reasons. First, hierarchical organisation is a widespread feature of data structures in general, and most people have at least the subjective impression that their own memory in particular is hierarchical. Second, the discovery of ultrametric structure in the states of the SK spin glass (Mézard *et al* 1984a, b) makes it interesting to look for other systems with this structure.

0305-4470/87/134449+07$02.50

We will see that these hierarchically ordered memories can be induced by a local learning rule quite similar to the Hopfield-Hebb one; the difference is simply that the patterns must be given specific weights which depend on the parameters characterising the hierarchical structure, e.g. on the Parisi function $q(x)$ (Parisi 1979, 1983).

Finally we establish the connection to a result of Virasoro and Parga (Virasoro 1986, Parga and Virasoro 1986). They have constructed a learning rule for hierarchical patterns by another method, based directly on the known ultrametric structure of the SK spin glass (Mézard and Virasoro 1985).

We formulate the problem in the following way. The correlation between the p learned patterns ξ_i^μ are characterised by their $(p \times p)$ mutual overlap matrix

$$Q_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu \tag{1}$$

and the learning rule of Personnaz *et al* and Kanter and Sompolinsky is

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu (Q^{-1})_{\mu\nu} \xi_i^\nu. \tag{2}$$

For Q^{-1} to exist, the patterns have to be linearly independent. The storage capacity clearly cannot exceed the number N of spins in the system. This learning rule is in general non-local because the calculation of Q^{-1} requires knowledge of all the ξ_i^ν. However, it is local if the overlap matrix of the system is given *a priori*. We expect that this kind of situation is rather common—we have an ensemble of many different sets of patterns, which share certain statistical properties, described by the Q matrix. The question is how to imprint patterns with the particular kind of correlations specified by a given Q. The ensemble we study here is one in which Q has a hierarchical (ultrametric) structure, but the general idea could obviously be applied to other ensembles, specified by other kinds of structure in Q.

We suppose that our ultrametric tree of patterns has n levels. An example is shown in figure 1. At the mth level each group of patterns consists of l_m subgroups from the $(m-1)$th level, and two patterns from the same group but different subgroups have overlap q_m. The overlap matrix Q has the Parisi form

$$Q_n = \begin{bmatrix} Q_{n-1} & & & q_n \\ & Q_{n-1} & & \\ q_n & & \cdots & \\ & & & Q_{n-1} \end{bmatrix} \tag{3}$$

where the $k_{n-1} \times k_{n-1}$ matrix Q_{n-1} has the same form as Q_n. (In terms of the l_m, $k_m = l_1 l_2 \ldots l_m$.) Our central task is to invert Q_n.

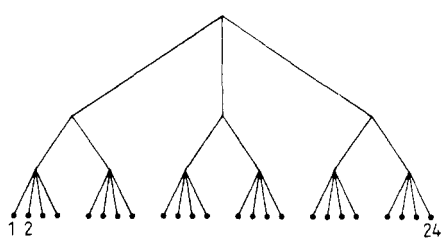

Figure 1. Ultrametric tree of $p = 24$ patterns with $n = 3$ levels. In this example $l_1 = 4$, $l_2 = 2$ and $l_3 = 3$.

2. The inverse matrix

The first row of the overlap matrix has the form

$$q_0 q_1 \ldots q_1 q_2 \ldots q_2 \ldots q_m \ldots q_m \ldots q_n \ldots q_n \tag{4}$$

where element q_m is repeated $l_1 \ldots l_{m-1}(l_m - 1)$ times. (Of course $q_0 = 1$.) The other rows contain exactly the same number of each element, the blocks being just permuted.

We find the (unnormalised) eigenvectors in the way Ogielski and Stein (1985) do for the case $l_m = 2$ for all m. All but one are constructed in the following way: for $0 \leqslant m < n$ the vector components are partitioned into groups of k_m elements. In one of the groups all the elements are $+1$, in another, -1, the rest being zero. This gives us $l_n \ldots l_{m+2}(l_{m+1} - 1)$ linearly independent eigenvectors each having the same eigenvalue

$$\eta_m = q_0 + (l_1 - l)q_1 + \ldots + l_1 \ldots l_{m-1}(l_m - 1)q_m - l_1 \ldots l_m q_{m+1} \qquad 0 \leqslant m < n. \tag{5}$$

For $m = n$ we find the eigenvector $(1, 1, \ldots, 1)$ with the non-degenerate eigenvalue

$$\eta_n = q_0 + (l_1 - 1)q_1 + \ldots + l_1 \ldots l_{n-1}(l_n - 1)q_n. \tag{6}$$

By formally putting $q_{n+1} = 0$ we can write all the eigenvectors in the following way:

$$\eta_m = q_0 + (k_1 - 1)q_1 + \ldots + (k_m - k_{m-1})q_m - k_m q_{m+1} \qquad 0 \leqslant m \leqslant n. \tag{7}$$

We will assume that all the q are non-negative, and that they are decreasing: $q_0 > q_1 > \ldots > q_n$, giving positive increasing eigenvalues

$$\eta_m - \eta_{m-1} = k_m(q_m - q_{m+1}) > 0 \qquad 0 < m \leqslant n \tag{8}$$

$$\eta_0 = q_0 - q_1 > 0. \tag{9}$$

Therefore the overlap matrix Q has an inverse matrix P. By using the well known inversion formula from linear algebra for the elements of P

$$P_{ij} = \frac{(-1)^{i+j} \det Q_{ij}}{\det Q} \tag{10}$$

(where Q_{ij} is the matrix Q without row i and column j) it is easily seen that P has the same structure as Q but with elements p_m, $m = 0, \ldots n$ instead of q_m. The eigenvalues of P are then

$$\zeta_m = p_0 + (k_1 - 1)p_1 + \ldots + (k_m - k_{m-1})p_m - k_m p_{m+1} \qquad 0 \leqslant m \leqslant n,\ p_{n+1} = 0. \tag{11}$$

The ζ_m will have the same degeneracy as the η_m and they will just be equal to $1/\eta_m$. Thus from (8) we obtain the formula

$$k_m(p_m - p_{m+1}) = \zeta_m - \zeta_{m-1} = \frac{1}{\eta_m} - \frac{1}{\eta_{m-1}} \tag{12}$$

and $p_0 - p_1 = 1/\eta_0$, corresponding formally to $1/\eta_{-1} = 0$. It is now easy to calculate the p_m recursively.

3. The connection strength J

From the matrix elements of $P = Q^{-1}$ we are able to calculate the connection strength J_{ij} from (2). Let E_m be a $k_m \times k_m$ matrix with all elements equal to 1. Then we can write P as a sum of $n+1$ $p \times p$ matrices in the following way:

$$P = p_n E_n + (p_{n-1} - p_n)\begin{bmatrix} E_{n-1} & & 0 \\ & \cdots & \\ 0 & & E_{n-1} \end{bmatrix} + \ldots + (p_0 - p_1)\mathbf{I} \tag{13}$$

where $\mathbf{I}$ is the unit matrix. On each level m of the hierarchical tree we average spin j over the patterns belonging to the same group. This gives us the mean values

$$\xi_j^{m,k} \qquad 1 \leqslant k \leqslant k_n/k_m \tag{14}$$

and formulae (2) and (12) lead to

$$J_{ij} = \frac{1}{N}\sum_{m=0}^{n} k_m \left(\frac{1}{\eta_m} - \frac{1}{\eta_{m-1}}\right) \sum_{k=1}^{k_n/k_m} \xi_i^{m,k}\xi_j^{m,k}. \tag{15}$$

After a little algebra this can be put in the form

$$J_{ij} = \frac{k_n}{N\eta_n}\xi_i^{n,1}\xi_j^{n,1} + \frac{1}{N}\sum_{m=0}^{n-1}\frac{k_m}{\eta_m}\sum_{k=1}^{k_n/k_m}(\xi_i^{m,k} - \xi_i^{m+1,[k/l_{m+1}]})(\xi_j^{m,k} - \xi_j^{m+1,[k/l_{m+1}]}) \tag{16}$$

where the notation $[x]$ means x should be rounded up to the next integer. This form shows explicitly how one can teach the system hierarchically ordered patterns. One teaches first the mean patterns $\xi_i^{n,1}$, then the next level details, then the details within each of these groups, and so on. Regarding the successive levels of detail as the basic pattern set, the procedure is very much like the Hebb rule except for the weights k_m/η_m associated with the details at level m. This factor tells the teacher how much stress to put on details, relative to gross features, at each level. We also note that the rule in this form (or in the form (15)) is local in the sense mentioned in the introduction. If we are looking for biological relevance, this is a desirable feature, as stressed by Kanter and Sompolinsky.

4. Writing the result in terms of the Parisi function $q(x)$

We know from spin glass theory (Parisi 1979, 1983) that a convenient way to parametrise an ultrametrically correlated set of configurations is by the order function $q(x)$, or, equivalently, by its inverse function $x(q)$. Our result (16) can be written very neatly in terms of $q(x)$.

The function $x(q)$ is the cumulative overlap distribution function. For our patterns, we have

$$x_m \equiv x(q_m) \equiv P(\text{overlap} \leqslant q_m) = 1 - \frac{k_{m-1}}{k_n} \qquad 1 \leqslant m \leqslant n \tag{17}$$

$$x_0 = 1 \tag{18}$$

giving

$$\frac{k_m - k_{m-1}}{k_n} = x_m - x_{m+1}. \tag{19}$$

This allows us to write η_m as an integral of $q(x)$, where $q(x)$ is defined to be equal to q_m on the interval $[x_m, x_{m+1}]$:

$$\begin{aligned}\eta_m &= q_0 + (k_1 - 1)q_1 + \ldots + (k_m - k_{m-1})q_m - k_m q_{m+1} \\ &= k_n[(1 - x_1)q_0 + (x_1 - x_2)q_1 + \ldots + (x_m - x_{m+1})q_m - (1 - x_{m+1})q_{m+1}] \\ &= k_n \int_{x_{m+1}}^{1} [q(x') - q(x_{m+1})]\, \mathrm{d}x'. \end{aligned} \tag{20}$$

Introducing this notation we can rewrite the reciprocal of the weight coefficients k_m/η_m which appear in (16) as

$$\begin{aligned}\frac{\eta_m}{k_m} = \frac{\eta_m/k_n}{k_m/k_n} &= \int_{x_{m+1}}^{1} \frac{[q(x') - q(x_{m+1})]}{1 - x_{m+1}}\, \mathrm{d}x' \\ &= \frac{\int_{x_{m+1}}^{1} [q(x') - q(x_{m+1})]\, \mathrm{d}x'}{\int_{x_{m+1}}^{1} \mathrm{d}x'} = \langle q \rangle_{[x_{m+1},1]} - q_{m+1}. \end{aligned} \tag{21}$$

We hope the meaning of the $\langle\ \rangle$ notation is clear. In words, the reciprocal of the learning strength to be given to data at correlation level q (which has cumulative overlap probability $x(q)$) is just the difference between the average of $q(x)$ over all larger cumulative overlaps and q itself.

5. The Parga and Virasoro learning rule

Virasoro (1986) investigated an ultrametric hierarchy in which there were only two levels present. His overlap matrix looks like this

$$C = \begin{bmatrix} C_1 & & & 0 \\ & C_2 & & \\ & & \ldots & \\ 0 & & & C_n \end{bmatrix} \tag{22}$$

where

$$C_\alpha = \begin{bmatrix} 1 & & & q_\alpha \\ & 1 & & \\ & & \ldots & \\ q_\alpha & & & 1 \end{bmatrix}. \tag{23}$$

That is, the overlap between the k_α patterns in group α is q_α; the different groups need not have the same size. Thus in one way this model is less general than the one we have considered in the preceding sections, while in another way it is more so.

This hierarchy can be looked on as n small one-level hierarchies giving independent

contributions J_{ij}^{α} to J_{ij}:

$$J_{ij} = \sum_{\alpha=1}^{n} J_{ij}^{\alpha}. \tag{24}$$

The J_{ij}^{α} can be calculated with the help of equation (16):

$$J_{ij}^{\alpha} = \frac{1}{N}\left(\frac{1}{1-q_\alpha}\sum_{\beta=1}^{k_\alpha}(\xi_i^{\alpha\beta}-\xi_i^{\alpha})(\xi_j^{\alpha\beta}-\xi_j^{\alpha}) + \frac{k_\alpha}{1+(k_\alpha-1)q_\alpha}\xi_i^{\alpha}\xi_j^{\alpha}\right) \tag{25}$$

where ξ_i^{α} means the average of $\xi_i^{\alpha\beta}$ in group α. When k_α is large, this gives exactly the result of Virasoro:

$$J_{ij} = \frac{1}{N}\sum_{\alpha=1}^{n}\left(\frac{1}{1-q_\alpha}\sum_{\beta=1}^{k_\alpha}(\xi_i^{\alpha\beta}-\xi_i^{\alpha})(\xi_j^{\alpha\beta}-\xi_j^{\alpha}) + \frac{1}{q_\alpha}\xi_i^{\alpha}\xi_j^{\alpha}\right) \tag{26}$$

although he reached his result by a different argument.

Formula (26) is a special case of a general formula proposed by Parga and Virasoro (1986) for a multilevel hierarchy. They propose

$$J_{ij} = \frac{1}{N}\sum_{m=0}^{n-1}\frac{1}{q_m - q_{m+1}}\sum_{k=1}^{k_n/k_m}(\xi_i^{m,k}-\xi_i^{m+1,[k/l_{m+1}]})(\xi_j^{m,k}-\xi_j^{m+1,[k/l_{m+1}]}). \tag{27}$$

This resembles our formula (16), but there is a difference in the weight coefficients, as one can see from (21). Equivalently we get from (8)

$$\frac{k_m}{\eta_m} = \left(\frac{\eta_{m-1}}{k_m} + q_m - q_{m+1}\right)^{-1} \tag{28}$$

so in the limit $k_{m-1}/k_m \ll 1$ (infinite branching ratio) the results are the same.

We know from Kanter and Sompolinsky (1986) that our model has a single critical temperature below which all learned patterns become stable. In general the learning rule of Parga and Virasoro would lead to different critical temperatures for the different groups of patterns. However, this question, and even that of the stability of the different memories at $T = 0$, have not been systematically studied yet.

6. Conclusion

We have shown how hierarchically correlated patterns can be embedded in a Hopfield-style network with a rather simple modification of the simple Hebb learning rule. The new rule is still local; the only change is in the relative weight given to components of the patterns at different levels of the ultrametric hierarchy. Closely related results have been obtained by Parga and Virasoro (1986) and Feigelman and Ioffe (1986), but the present method gives direct insight into the form of the result.

We note again that the maximum capacity of the model is N patterns. If one wants to store more, one must introduce three-spin or higher-order couplings. A start in this direction for hierarchical patterns has been made by Feigelman and Ioffe (1986).

Possible extensions of this work include studies of the thermodynamics of the network at finite temperature, the effects of fluctuations around the ideal ultrametric structure discussed here, and what happens in the presence of 'forgetting' as discussed for the uncorrelated-pattern case by Mézard *et al* (1984a, b) and Parisi (1986).

References

Feigelman M V and Ioffe L B 1986 *Preprint* Moscow
Hebb D O 1949 *The Organization of Behavior* (New York: Wiley)
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kohonen T 1977 *Associative Memory; A System-Theoretical Approach* (Berlin: Springer)
—— 1984 *Self Organization and Associative Memory* (Berlin: Springer)
Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457
Mézard M, Parisi G, Sourlas N, Toulouse G and Virasoro M A 1984a *Phys. Rev. Lett.* **52** 1156
—— 1984b *J. Physique* **45** 843
Mézard M and Virasoro M A 1985 *J. Physique* **46** 1293
Ogielski A T and Stein D L 1985 *Phys. Rev. Lett.* **55** 1634
Parga N and Virasoro M A 1986 *J. Physique* **47** 1857
Parisi G 1979 *Phys. Rev. Lett.* **43** 1754
—— 1983 *Phys. Rev. Lett.* **50** 1946
—— 1986 *J. Phys. A: Math. Gen.* **19** L617
Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
Rammal R, Toulouse G and Virasoro M A 1986 *Rev. Mod. Phys.* **58** 765
Virasoro M A *NATO ASI Series* vol F20 *Disordered Systems and Biological Organization* (Berlin: Springer) p 197