EUROPHYSICS LETTERS 15 May 1986

Europhys. Lett., 1 (10), pp. 535-542 (1986)

Networks of Formal Neurons and Memory Palimpsests.

J. P. NADAL

Groupe de Physique des Solides, E.N.S.
24 *rue Lhomond*, 75231 *Paris Cedex* 05, *France*

G. TOULOUSE

Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris
10 *rue Vauquelin*, 75005 *Paris, France*

J. P. CHANGEUX and S. DEHAENE

Unité de Neurobiologie Moléculaire and Laboratoire Associé au C.N.R.S. n° 270
Interactions Moléculaires et Cellulaires, Institut Pasteur
25 *rue du Docteur Roux*, 75724 *Paris Cedex* 15, *France*

(received 26 February 1986; accepted 10 March 1986)

PACS. 87.30. – Biophysics of neurophysiological processes.
PACS. 75.10H. – Ising and other classical spin models.
PACS. 64.60. – General studies of phase transitions.

Abstract. – One characteristic behaviour of the Hopfield model of neural networks, namely the catastrophic deterioration of the memory due to overloading, is interpreted in simple physical terms. A general formulation allows for an exploration of some basic issues in learning theory. Two learning schemes are constructed, which avoid the overloading deterioration and keep learning and forgetting, with a stationary capacity.

Neural networks with symmetric interactions provide a simple model for a distributed, content-addressable memory [1]. A formal neuron is represented by a spin variable S_i, which can take two values, $S_i = +1$ (neuron firing), $S_i = -1$ (neuron quiescent). The synaptic efficacies T_{ij} can be positive (excitatory) or negative (inhibitory). With the assumption of symmetric interactions, an energy function can be defined

$$E = -\left(\frac{1}{2}\right) \sum_{i \neq j} T_{ij} S_i S_j \tag{1}$$

and the neuron dynamics [1] leads to downhill motion on the energy landscape in the configuration space defined by (1).

The standard procedure for the storage of p input patterns follows the generalized Hebb

rule, and is expressed by

$$T_{ij} = \sum_{s=1,p} \Delta T_{ij}(s)\,, \tag{2}$$

$$\Delta T_{ij}(s) = \left(\frac{1}{N}\right) \xi_i^s \xi_j^s\,, \tag{3}$$

where $S_i = \xi_i^s$ defines pattern s. Within such a process, optimal storage for randomly independent patterns is obtained for $p \sim 0.15N$, where N is the number of neurons in a fully interconnected network. Above this value, confusion sets in and retrieval quality sharply deteriorates [1-3]. A simple way to quantify this statement is to plot the number of memories which are retrieved with better than 97% accuracy, as a function of the total number p of printed patterns. This is shown in fig. 1*a*) for $N = 100$.

This procedure is typically instructive in character [4]. Recently, a selective version of the model has been proposed [5]. Some of the distinctive issues under discussion concerned the following aspects: initial state before learning, synaptic sign changes, categorization properties, choice of learning rule. In the present paper the selective *vs.* instructive issue is not addressed directly. The focus is put on the prevention of the afore-mentioned memory deterioration, due to overloading. It is shown that the hypothesis of uniform amplitude for the storage of any memory, and/or the absence of synaptic constraints, are responsible for the breakdown. With the help of a simple unifying treatment, it is possible to go beyond these restrictive hypotheses and to obtain a network that keeps a permanent capacity for learning. New patterns are stored on top of previous ones, which get progressively erased. For this reason, such a memory may be figuratively termed a palimpsest. Two such schemes, named «marginalist learning» and «learning within bounds», are defined and studied.

1. Basic formulation.

Two constitutive equations are derived. One relevant variable is the average of the squared synaptic efficacy modifications, over all synapses (i, j), for one pattern s:

$$k(s) = \langle \Delta T_{ij}^2(s) \rangle - \langle \Delta T_{ij}(s) \rangle^2, \tag{4}$$

which defines the pattern acquisition intensity (a measure of its memory trace). Notice that in the special case of the generalized Hebb rule (3), $k(s)$ is independent of s,

$$k(s) = 1/N^2.$$

The other basic variable is the cumulated synaptic intensity after learning p patterns:

$$K(p) = \langle T_{ij}^2 \rangle - \langle T_{ij} \rangle^2. \tag{5}$$

If the inputs are randomly independent, it is possible here to ignore the correlations between synapses and to treat the $\Delta T_{ij}(s)$ themselves as independent random variables. Hence the central limit theorem goes

$$K(p) - K(p-1) = k(p)\,. \tag{6}$$

(2)

(3)

for randomly
rons in a fully
uality sharply
ıe number of
n of the total

ive version of
ion concerned
categorization
uctive issue is
oned memory
amplitude for
esponsible for
to go beyond
ıt capacity for
sively erased.
st. Two such
ɜ defined and

verage of the
:tern s:

(4)

ɜ). Notice that
s,

p patterns:

(5)

e correlations
lom variables.

(6)

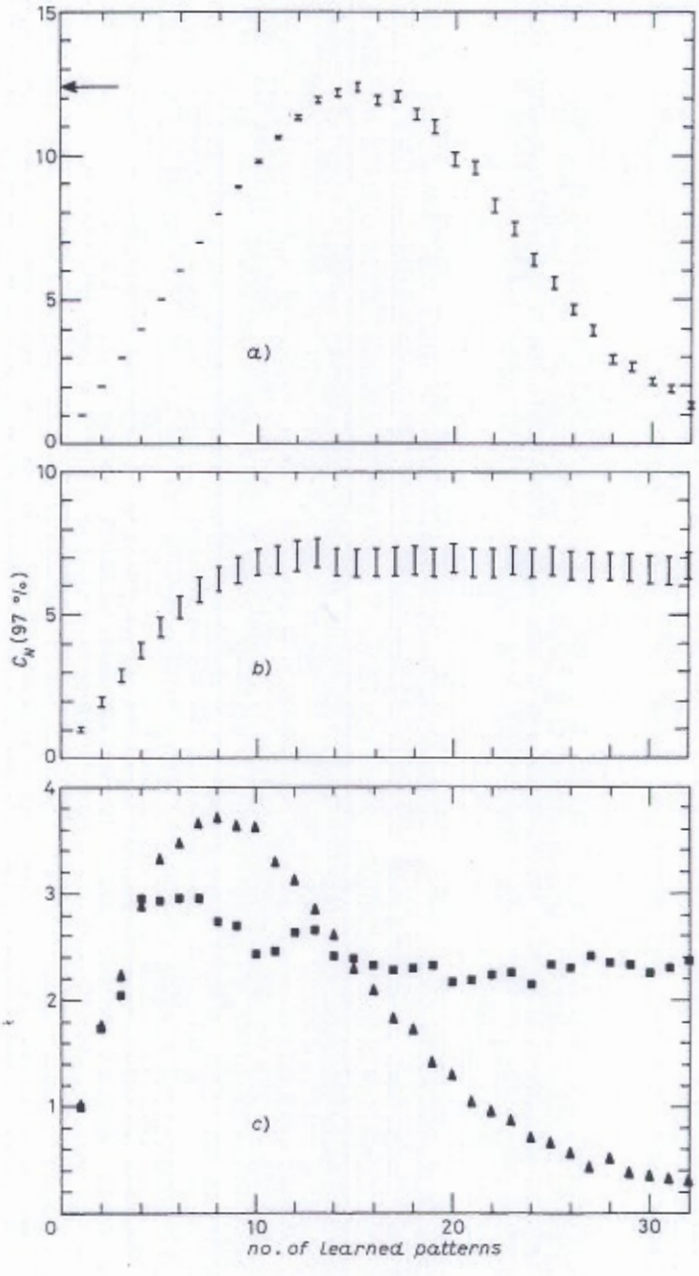

Fig. 1. – Number of patterns with a retrieval overlap q (defined as the projection of an input pattern on its attractor) better than 97%, as a function of the total number p of learned patterns, for $N = 100$. *a*) Hopfield model, the optimal capacity is reached at $p \sim 12.5$. *b*) Marginalist scheme for $\varepsilon = 2.7$. *c*) Scheme with bounded efficacies for $\varepsilon = 1$ (▲) (the catastrophic deterioration occurs), and $\varepsilon = 2.7$ (■) (the memory reaches a stationary regime).

A continuous version of this equation, in terms of a «time» variable proportional to the number of stored patterns, may be written as

$$K'(t) = k(t) . \tag{7}$$

This first constitutive equation gives the increase in total synaptic efficacy due to the last learning event.

The second equation gives the threshold intensity $\boldsymbol{k}$ for one pattern to be safely stored against a background intensity K. It is known from spin glass theory, and it is easily checked numerically [5] see also [6], that

$$\boldsymbol{k} = \varepsilon^2 K/N , \tag{8}$$

where ε is a numerical factor, which was estimated as $\varepsilon \sim 2.5$ in the spin glass limit, where strict random independence of the interactions holds by definition. This value (a kind of cross-over point) was found to guarantee almost perfect retrieval (98% accuracy), and the scaling dependence with N has been checked for $N = 64$, 100, 200.

2. Standard learning process.

In this scheme, by assumption [1-3], the interactions are vanishing in the initial state (*tabula rasa* hypothesis) and the acquisition intensity is uniform:

$$k(s) = k .$$

Accordingly, the cumulated intensity $K(t)$ grows linearly with «time»:

$$K(t) = kt , \tag{9}$$

in virtue of (7), and the maximal capacity $\boldsymbol{p}$ obtains for

$$k = \varepsilon^2 K(\boldsymbol{p})/N \tag{10}$$

in virtue of (8). Inserting (9) into (10), one gets

$$\boldsymbol{p} = N/\varepsilon^2 . \tag{11}$$

Two remarks. The value of the learning intensity k has dropped out from expression (11); this is not surprising, since there is no intensity scale within the *tabula rasa* hypothesis. From the theoretical estimate $\boldsymbol{p} = 0.138N$, one derives a value for ε, $\varepsilon = 2.69$, which is in striking accord with the spin glass estimate [2, 5]. This suggests that the effect of correlations between synaptic efficacies is small, at least for that matter. Continuation of this learning process for $p > \boldsymbol{p}$ is catastrophic, because all memories sink simultaneously under the threshold level defined by (8).

The case without *tabula rasa* can be treated with equal ease. For simplicity, we assume that the synapses are randomly excitatory or inhibitory, with $K = K_0$, before learning occurs. Then (9) is modified into

$$K(t) = K_0 + kt . \tag{12}$$

Inserting (12) into (8), one gets now

$$p = (N/\varepsilon^2) - (K_0/k)\,. \tag{13}$$

The asymptotic capacity is unchanged, but a threshold in N has appeared. For efficient storage to occur, the network size must be larger than a threshold:

$$N > N_c = K_0\,\varepsilon^2/k\,. \tag{14}$$

It should be reminded here that, within this model, N is also the connectivity of the network, *i.e.* the number of neurons connected to any given one. Note that the acquisition intensity k is no longer irrelevant and does govern N_c. To give some numbers, if the synaptic modification in one learning event is 10% of its initial value, the network must contain at least a thousand neurons for any efficient storage. With a network size equal to double this threshold, the optimal storage capacity is around a hundred patterns. When this number has been reached, cumulated synaptic modifications amount to about 50%.

3. Marginalist learning.

By definition, a learning process will be called *marginalist*, if the acquisition intensity for learning the last pattern is tuned to be exactly at its threshold value.

Catastrophic memory deterioration, due to overloading, is mainly due to the hypothesis of uniform acquisition intensity. Indeed, we know from (8) that is it always possible to learn a new pattern, provided the acquisition intensity is strong enough. There is of course a price to pay in order to secure this stabilization, and this comes as an increased erasure of the previously learned patterns.

As a consequence of the marginalist hypothesis, (7) and (8) combine to yield an exponential growth for both the cumulated and marginal intensities:

$$K(t)/K_0 = k(t)/k(0) = \exp\left[\varepsilon^2\, t/N\right]. \tag{15}$$

Indeed, in the original discrete «time» process, the synaptic efficacy evolution reads

$$T_{ij}(p) = T_{ij}(p-1) + \varepsilon[K(p-1)/k]^{\frac{1}{2}}\,\delta_{ij}(p)\,, \tag{16}$$

where, in the case of the generalized Hebb rule (3),

$$\delta_{ij}(p) = \sqrt{N}\,m_i^p\,\sqrt{N}\,m_j^p.$$

This leads to

$$K(p) = K_0 \prod_{s=1,p} \left\{1 + \left(\frac{\varepsilon^2\,\sigma(s)}{N}\right)\right\} \tag{17}$$

with

$$\sigma(s) = \langle \delta_{ij}(s)^2\rangle - \langle \delta_{ij}(s)\rangle^2 = 1 - \langle \delta_{ij}(s)\rangle^2.$$

For independently random patterns, and large N, $\sigma(s) = 1$. Thus asymptotically, (17) reduces to (15). Note that the generalization of the preceding arguments to learning rules, other than Hebb, is straightforward.

In the strict sense, only the last pattern emerges above threshold. However, if one counts the number of stored memories which are retrieved with better than, say, 97% accuracy, our numerical results show that, for $p \geq 0.1N$, the capacity reaches a running plateau, which is about a half of the optimal storage capacity (11). The ability to keep learning has been paid by a factor two reduction in the storage capacity (fig. 1*b*)).

It is intuitive, and in fact true, that within this marginalist-learning process, the more anciently a pattern has been printed, the more deeply it is buried and forgotten. This «time» decay is illustrated in fig. 2, by plotting the retrieval quality as a function of storage ancestry.

Marginalist learning is not completely unrealistic for some forms of memory. Suggestive possibilities are also open by unlearning effects [7], allowing for the recovery of sunk memories. Exponential growth, however, becomes asymptotically unrealistic, whether for natural or artificial memories. A simple alternative process (1), which captures learning and forgetting, in a stationary mode, free from any exponential growth, is studied in the next section.

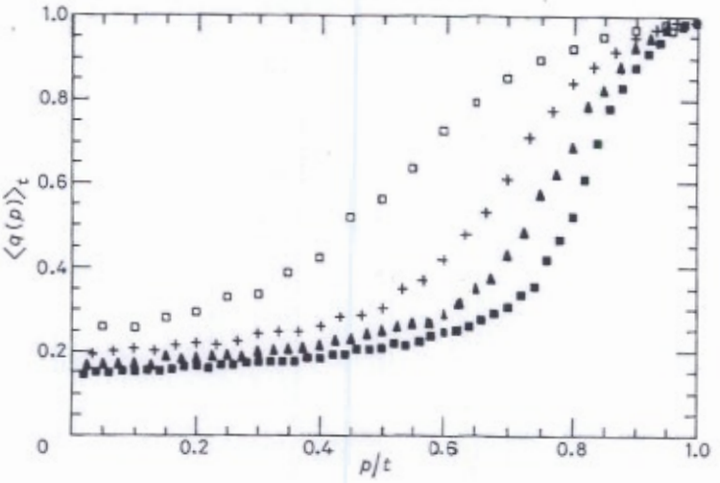

Fig. 2. – Marginalist learning: mean retrieval overlap $\langle q(p) \rangle_t$ of the p-th pattern once t have been learned, as a function of p/t, shown here for $N = 100$, $\varepsilon = 2.7$, and $t = 20$ (□), 30 (+), 40 (▲) and 50 (■). Apart from the most recently learned patterns, $\langle q \rangle$ takes the minimal possible value ~ 0.15.

4. Learning within bounds.

The notion of learning without synaptic sign changes has been evoked already [5]. We impose here a stronger constraint, namely that the synaptic efficacies are bounded above and below:

$$0 \leqslant T_{ij} \leqslant A\,, \quad \text{or} \quad -A \leqslant T_{ij} \leqslant 0\,. \tag{18}$$

The acquisition intensity k is taken uniform for simplicity.

(1) We are thankful to G. PARISI for his contribution to the definition of this simple model.

If k is very small, the bounds have little effect, and the catastrophic deterioration of sect. 2 still occurs. If k is very large, however, the last pattern will be optimally memorized. Consideration of these two limits, therefore, suggests that a transition takes place as a function of k. From the constitutive eqs. (6), (8), one sees that if

$$A = 1/N^{\frac{1}{2}}, \tag{19}$$

then the critical value k_c should scale as $1/N^2$:

$$k_c = (\varepsilon_c/N)^2. \tag{20}$$

We have found numerical evidence for such a transition, by looking at different values of $\varepsilon = k^{\frac{1}{2}}N$, at $N = 100$ and $N = 200$. In fig. 1c) one sees that for $\varepsilon = 1.0$ the characteristic deterioration is observed, whereas for $\varepsilon = 2.7$ a stationary regime is reached with a capacity C_N which is a fraction of the optimal, but transient, capacity observed at $\varepsilon = 1$. The capacity C_N, shown in fig. 3 as a function of ε, behaves as an order parameter. The threshold, or critical value, is found at $\varepsilon_c \sim 1.2$. Remarkably, a maximum occurs at a value ε_m. In our calculations with $N \leqslant 200$, this value seems to be independent of N: $\varepsilon_m \sim 3$, and the optimal capacity C_n^m grows with N: $C_N^m \sim 0.016\,N$. For $\varepsilon \geqslant N^{\frac{1}{2}}$, only the last pattern is perfectly memorized, the asymptotic capacity is one exactly. It is known [8] that human short-term memory can keep seven (plus or minus two) separate items (*e.g.* numbers). It is amusing to note that in our scheme, a capacity of seven patterns is reached with a connectivity of about 500, each neuron being connected to five hundred other neurons.

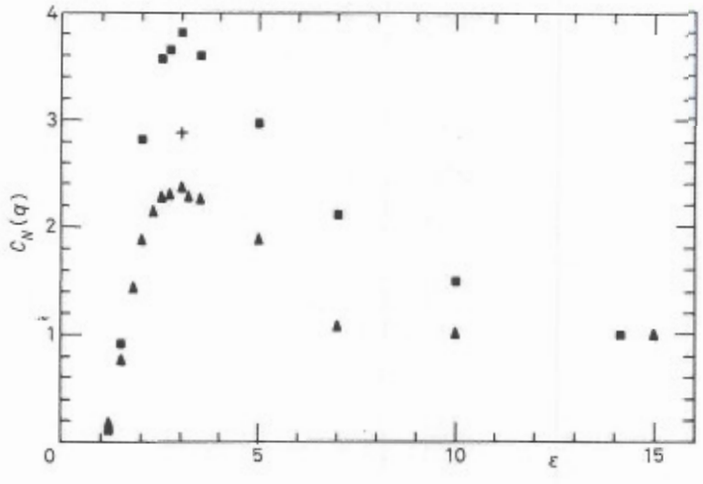

Fig. 3. – Asymptotic capacity: C_N as a function of ε, for $N = 100$ (▲) and $N = 200$ (■, +). For comparison with the other models, C_N is the mean number of patterns with retrieval better than 97%. Shown also is the mean number of patterns with quasi-perfect retrieval (99%) (+) for $N = 200$.

These results remain qualitatively similar with other learning rules, such as the one considered in [5]. Dropping the sign constraint, so that (18) is replaced by the unique constraint $-A \leqslant T_{ij} \leqslant A$, was found not to affect significantly the qualitative results.

As a final remark, it is worth stressing that, in these schemes of palimpsest-type memories (sect. 3 and 4), forgetting appears as an active process, directly related to new learning events, as distinct from a passively relaxational time decay. Forgetting comes as a masking effect, due to interference of superimposed memory traces. There are suggestive analogies between some properties of these memory palimpsests and several features of the short-term memory in man [8, 9]. The possible biological relevance of this study will be critically discussed in a forthcoming publication.

In conclusion, this paper illustrates a perhaps counterintuitive phenomenon: The introduction of constraints in the learning process can lead to improved, instead of diminished, storage performances.

* * *

Useful discussions with M. MÉZARD, G. PARISI and M. A. VIRASORO, during a visit to Rome, are gratefully acknowledged by one of us (GT). One of us (JPN) would like to thank G. WEISBUCH for interesting discussions. The numerical calculations were made possible thanks to the GRECO «Expérimentation Numérique» (supported by the C.N.R.S.).

REFERENCES

[1] J. J. HOPFIELD: *Proc. Natl. Acad. Sci. (USA)*, **79**, 2554 (1982).
[2] D. J. AMIT, H. GUTFREUND and H. SOMPOLINSKY: *Phys. Rev. Lett.*, **55**, 1530 (1985).
[3] W. KINZEL: *Z. Phys. B*, **60**, 205 (1985).
[4] N. JERNE: in *The Neurosciences: A Study Program*, edited by G. QUARTON *et al.* (The Rockfeller University Press, 1967).
[5] G. TOULOUSE, S. DEHAENE and J. P. CHANGEUX: *Proc. Natl. Acad. Sci. (USA)*, in press.
[6] L. PERSONNAZ, I. GUYON, G. DREYFUS and G. TOULOUSE: submitted to *J. Stat. Phys.*
[7] J. J. HOPFIELD, D. I. FEINSTEIN and R. G. PALMER: *Nature*, **304**, 158 (1983).
[8] F. E. BLOOM, A. LAZERSON and L. HOFSTADTER: *Brain, Mind, and Behavior* (W. H. Freeman, 1985), p. 191.
[9] B. KOLB and I. Q. WHISHAW: *Fundamental of Human Neuropsychology* (W. H. Freeman, 1980).