On learning rules and memory storage abilities of asymmetrical neural networks

P. Peretto

To cite this version:

On learning rules and memory storage abilities of asymmetrical neural networks

P. Peretto

Commissariat à l'Energie Atomique, C.E.N. Grenoble, D.R.F./SPh/P.S.C., 85X, 38041 Grenoble Cedex, France

(Reçu le 24 mars 1987, révisé le 12 août 1987, accepté sous forme définitive le 14 janvier 1988)

Résumé. — La plupart des modèles de mémoire utilise des synapses symétriques. Nous montrons que l'existence d'une capacité de mémorisation ne dépend pas de cette hypothèse. Nous présentons une théorie de la capacité de stockage en mémoire qui ne repose pas sur la méthode des répliques. Elle est fondée sur une approximation de champ moyen et de propriétés d'auto-moyennage. L'efficacité de mémorisation dépend de quatre paramètres d'apprentissage qui peuvent éventuellement être déterminés expérimentalement. Nous montrons que les observations expérimentales actuellement disponibles sur les règles de plasticité synaptiques sont compatibles avec un stockage efficace en mémoire.

Abstract. — Most models of memory proposed so far use symmetric synapses. We show that this assumption is not necessary for a neural network to display memory abilities. We present an analytical derivation of memory capacities which does not appeal to the replica technique. It only uses a more transparent and straightforward mean-field approximation. The memorization efficiency depends on four learning parameters which, if the case arises, can be related to datas provided by experiments carried out on real synapses. We show that the learning rules observed so far are fully compatible with memorization capacities.

1. Introduction.

Plasticity of neural systems is an old idea already proposed by Ramon Y Cajal at the turn of the century. In particular it has been recognized for a long time that the overall neural connectivity tends to be controlled by the neural activity (for an historical survey see [1]). But this observation does not give any clue regarding the fundamental principles governing neural dynamics and in particular one of its most intriguing properties, memory. A decisive step towards a genuine theory of memory was the introduction by Hebb of the idea of associative memory [2]: the efficacy of an excitatory synapse increases when the two neurons it links are simultaneously active. This process has been supported by experimental evidences on a number of preparations ranging from ganglia of invertebrates to central nervous systems of mammals. Of particular interest are the experiments carried out by Levy et al. on the hippocampus of rats [3]. This is a well defined situation where the interactions between the excited groups of cells are homosynaptic, monosynaptic and excitatory. Correlated activities between the upstream and downstream neurons indeed strongly potentiate the synapses. This effect can be modelled by an increase during unit time $\Delta C_{ij}$ of the (excitatory) efficacy between the presynaptic neuron $j$ and the postsynaptic neuron $i$.

$$\Delta C_{ij} \propto S_i \cdot S_j$$  \hspace{1cm} (1)

$S_i = 1$ if the neuron $i$ is active and $S_i = 0$ if it is silent.

An expression such as equation (1) which determines the dynamics of synaptic plasticity will be called a learning rule.

Equation (1) however implies more than the Hebbian rule: it states that the synaptic efficacy is not modified when at least one of the two involved neurons is silent. In particular the efficacy must not change when the presynaptic neuron $j$ is silent and the postsynaptic neuron $i$ is firing. Experiments, those of Levy among others, show that this is not
true: in this particular scheme of activities the efficacy of excitatory synapses undergoes a strong depotentiation. This effect is sometimes called the anti-Hebbian rule. The excitatory synapses are therefore not only able to store correlated activities but also anti-correlated activities. Instead of equation (1) the learning rule of excitatory synapses could take the following form:

$$\Delta C_{ij} \propto \sigma_i \cdot \sigma_j$$  \hspace{1cm} (2)

where $\sigma_i$ is defined by

$$\sigma_i = 2 \cdot S_i - 1$$

that is $\sigma_i = -1$ when the neuron $i$ is silent and $\sigma_i = 1$ when it is firing.

It is clear that rule (2) which takes more information on activity correlations into account than the rule (1) must endow the neural network with more efficient memory storage capacities. Actually rule (2) is the best local learning rule whereas rule (1) does not allow the network to store memory properly as we shall see below. By local rule we mean a process which involves only the activities of the pre- and post-synaptic neurons. It is possible to imagine more efficient non-local rules relying upon the activity of distant neurons.

From the point of view of experimental observation however, rule (2) is still not fully satisfactory. Indeed it implies that an excitatory synapse potentiates when the neurons it links are both silent ($\sigma_i = \sigma_j = -1$) and depotentiates when the downstream neuron $i$ is silent and the upstream neuron $j$ is active ($\sigma_j = 1, \sigma_i = -1$). Conditioning experiments carried out by Rauschecker et al. [4] on the visual cortex of cats show that the strength of the involved excitatory synapses are not modified when the post-synaptic neurons $i$ are silent whatever the pre-synaptic activity $j$ (see Tab. 1a). The learning rule must be modified accordingly.

$$\Delta C_{ij} \propto \sigma_j \cdot (\sigma_i + 1)$$  \hspace{1cm} (3)

Up to now we have only considered excitatory synapses. Actually very little is known on the plasticity of inhibitory synapses. Cooper, according to Singer, suggests that inhibitory synapses could undergo an anti-Hebbian learning rule symmetrical to that obeyed by excitatory synapses [5]. In this case the correlation between the hyperpolarization of the post-synaptic membrane due to an activity of an upstream neuron ($\sigma_j = 1$) impinging onto an inhibitory synapse and the hyperpolarization ($\sigma_j = -1$) of the post-synaptic neuron due to the other afferents would enhance the (absolute value of the) synaptic efficacy. For other authors the inhibitory synapses are much less plastic than the excitatory synapses (see Tab. 1b). In table I the synaptic efficacies are computed algebraically, positively for excitatory synapses, negatively for inhibitory synapses.

Actually the cortex involves many types of neurons and it seems that a least as many rules as neuronal types have to be invoked to model the learning processes. Some authors argue however that the relevant information is embedded in dominant neurons, the pyramidal cells [6]. The other cells would only play ancillary roles. The direct connections between pyramidal cells through apical dendrites are supposed to be excitatory. On the other hand the basal dendrites of pyramidal cells connect to other pyramidal cells through neurons, the stellate cells in particular, via inhibitory synapses thus providing an indirect inhibitory mechanism between pyramidal cells [6]. All these complex processes can
be modelled by introducing effective learning rules describing how the efficacies of effective synapses (between pyramidal cells) are modified on learning. These rules are displayed in table Ic. Because of experimental uncertainties it is better for the moment to leave $a^{++}$, $a^{+-}$, $a^{-+}$ and $a^{--}$, the experimental learning parameters as we call them, undetermined [7]. The question naturally arises to find the ranges of these parameters which allow the neural network to display efficient memory properties. This is the problem we consider in this article. The most convenient approach of a general theory of memory storage capacity appeals to the dynamics of the network. This is the way Feigelman [8] and Derrida [9] consider the problem of asymmetric networks. Indeed the usual treatments of these problems rest upon the analogy of neural networks with models which can be solved with the tools of statistical mechanisms. The interactions we consider here are non-symmetrical. Therefore an energy cannot be defined and statistical mechanics cannot be used straightforwardly. These authors solve the problems of very diluted networks with usual Hebbian rules. The systems which are studied here are more biologically minded. They are devised to model strongly connected networks with realistic plasticity rules as cortical microcolumns are supposed to be. The analytical results show that the range of convenient experimental learning parameters is rather large even though there exist domains which must be avoided absolutely. It is comforting to observe that the experimental learning parameters given in tables Ia and Ib ensure good memory properties for large ranges of learning parameters.

2. Simple model of effective synaptic efficacies.

In spite of the efforts of a number of authors [10] a satisfactory theory of learning dynamics remains to be made. It must be based on the synaptic modifications defined in the introduction but it also has to take saturation and delay effects into account. On the other hand it must explain the interplay between short and long term memory. Simple simulations show that the learning dynamics is generally either unstable or inefficient if some sort of control mechanism is missing [11].

In this article we shall ignore these problems altogether: a set of $M$ patterns $I^\mu$, $\mu = 1, \ldots, M$, is experienced by a system of $N$ neurons. The exposure time is the same for all patterns. When the network experiences a pattern $I^\mu$, the neuronal states are forced to the corresponding values. These are the assumptions of the Hopfield model [12]:

$$\sigma_i = \xi_i^\mu; \quad \xi_i^\mu = \pm 1.$$  

Finally we assume that the patterns are stochastically independent.

The elementary modification of the effective synaptic interaction induced by a learned pattern $I^\mu$ depends on the two binary activities $\xi_i^\mu$ and $\xi_j^\mu$ and therefore it is a pseudo-Boolean function of these two variables:

$$\Delta C_{ij}(\xi_i^\mu, \xi_j^\mu) = \frac{1}{N} \cdot (A \xi_i^\mu \xi_j^\mu + B \xi_i^\mu + C \xi_j^\mu + D)$$

(4)

The factor $1/N$ is for convenience's sake.

The parameters $A$, $B$, $C$, $D$ of equation (4), which we simply call the learning parameters, are related to the experimental learning parameters by:

$$\Delta C_{ij}(-1, -1) = a^{-+} = \frac{1}{N} \cdot (A - B + C + D)$$

$$\Delta C_{ij}(-1, +1) = a^{+-} = \frac{1}{N} \cdot (-A - B + C + D)$$

$$\Delta C_{ij}(+1, -1) = a^{+-} = \frac{1}{N} \cdot (-A + B - C + D)$$

$$\Delta C_{ij}(+1, +1) = a^{++} = \frac{1}{N} \cdot (A + B + C + D)$$

and therefore

$$A = \frac{N}{4} \cdot (a^{--} - a^{+-} - a^{-+} + a^{++})$$

$$B = \frac{N}{4} \cdot (-a^{--} - a^{--} + a^{+-} + a^{++})$$

$$C = \frac{N}{4} \cdot (-a^{--} + a^{+-} - a^{-+} + a^{++})$$

$$D = \frac{N}{4} \cdot (a^{--} + a^{+-} + a^{-+} + a^{++}).$$

(5)

For example the pattern independent learning parameter $D$, taking the experimental datas of tables Ia and Ib into account, is given by:

$$D = n^+ \cdot (a_{ex}^{++} + a_{ex}^{+-}) - n^- \cdot a_{in}^{-+}$$

(6)

where $n^+$ and $n^-$ are the average numbers of excitatory and inhibitory interactions between the relevant (e.g. pyramidal) neurons.

Because the saturation processes are ignored the contributions of the various learned patterns add linearly and the resultant synaptic efficacies can be written:

$$C_{ij} = \frac{1}{N} \cdot \left( A \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu + B \sum_{\mu=1}^{M} \xi_i^\mu + C \sum_{\mu=1}^{M} \xi_j^\mu + MD + D^0 \right)$$

(7)

In equation (7) an extra pattern-independent term $D^0$ has been introduced to account for non-plastic interactions.

Remark: hetero-synaptic junctions as well as the elimination of interneurons introduce interactions of higher orders. They can be modelled along the same
lines as those followed by homo-synaptic (2nd order)
junctions at the expense of a larger number of
learning parameters (see Appendix A).

3. Learning a finite number of configurations.

A finite number of configurations means that
\( M = O \) (1).

The so-called instantaneous average activity of a
neuron \( i \) obeys the following equation of motion [13]

\[
\frac{d\langle \sigma_i \rangle}{dt} = -\frac{1}{\tau} \left( \langle \sigma_i \rangle - \langle \xi(\beta h_i) \rangle \right). \tag{8}
\]

The statistical average of an observable (e.g. the
activity) is defined by :

\[
\langle O \rangle = \sum_i \rho(I, t) \cdot O(I) \tag{9}
\]

where

- \( \rho(I, t) \) is the probability for the network as a whole
to be in state \( I \) at time \( t \),
- \( \tau \) is a relaxation constant which is of the order
of the refractory period of neurons,
- \( h_i \) is the membrane potential of neuron \( i \). It is
given by :

\[
h_i(I) = \sum_j C_{ij} \cdot \sigma_j(I) \tag{10}
\]

\( \beta \) is a noise parameter : \( \beta^{-1} = 0 \) for noiseless
networks.

\( \xi \), a sigmoid function, is the probability for the
membrane potential to overcome a threshold \( \theta_i \).
Neurons indeed are noisy devices and this is why a
probabilistic description of neural networks is neces-
sary. Because the synaptic noise is Gaussian, \( \xi \) is an
error function. However it is often easier to derive
analytical results using the tanh function which is
close to the erf function rather than the erf function
itself. Actually it is not necessary to specify the
sigmoid function for most of the following deri-
vations.

The equations (8) are solved using the mean field
approximation, which is valid for strongly connected
networks that is networks with a number of links of
the order of \( O(N^2) \). An exact expression can also be
derived for very poorly connected networks [9].
According to this approximation the statistical aver-
age carried out on the sigmoidal function can be
transferred to its argument :

\[
\langle \xi(\beta h_i) \rangle = \xi(\langle h_i \rangle) \tag{11}
\]

with

\[
\langle h_i \rangle = \sum_j C_{ij} \langle \sigma_j \rangle. \tag{12}
\]

The equations (8), (11) and (12) are a closed set of
equations which determine the dynamics of the \( N \)
average activities.

Using the expressions (7) giving the synaptic
efficacies the average membrane potential can be
written as a sum of \( M \) contributions or partial fields :

\[
\langle h_i \rangle = \sum_{\mu=0}^{M} \langle h_i^\mu \rangle \tag{13}
\]

with

\[
\langle h_i^\mu \rangle = \frac{1}{N} \cdot \left( A \xi^\mu \sum_j \xi^\mu \langle \sigma_j \rangle + B \xi^\mu \sum_j \langle \sigma_j \rangle + C \sum_j \xi^\mu \langle \sigma_j \rangle \right); \quad \mu \neq 0
\]

and

\[
\langle h_i^0 \rangle = \frac{(MD + D^0)}{N} \cdot \sum_j \langle \sigma_j \rangle. \tag{14}
\]

We define \( M + 1 \) order parameters conjugate to the
\( M \) partial fields \( \langle h_i^\mu \rangle \) and of the uniform field
\( \langle h_i^0 \rangle \).

\[
m^\mu = \frac{1}{N} \cdot \sum_j \xi^\mu \cdot \langle \sigma_j \rangle; \quad \mu \neq 0
\]

\[
m^0 = \frac{1}{N} \cdot \sum_j \langle \sigma_j \rangle. \tag{15}
\]

Using equations (13) and (14), the fields are given by :

\[
\langle h_i^\mu \rangle = m^\mu \cdot (A \xi^\mu + C) + m^0 \cdot B \xi^\mu; \quad \mu \neq 0
\]

\[
\langle h_i^0 \rangle = (MD + D^0) \cdot m^0. \tag{16}
\]

A pattern is stored if it is a fixed point of the
dynamics (Eq. (8)) :

\[
\langle \sigma_i \rangle = \langle \xi(\beta h_i) \rangle = \sum_j \xi(\beta h_i(I)) \cdot \rho(I, \infty)
\]

where \( \rho(I, \infty) \) is the asymptotic distribution. With
the mean field approximation these expressions
reduce to :

\[
\langle \sigma_i \rangle = \xi(\langle h_i \rangle) \tag{17}
\]

which can, at least in principle, be solved without
any explicit calculation of the asymptotic distribu-
tion.

The memory storage properties of networks obey-
ing the learning process (2) \((B = C = D = 0, A > 0)\) have been thoroughly studied [12,
14, 15, 16]. For these systems the memorized
patterns are stable fixed points of the dynamics.
Indeed let us assume that the neural states take the
values of say the pattern \( I \). The membrane potential
on a neuron \( i \) is a sum of a coherent partial field, the
one corresponding to pattern 1, and a random field
resulting from the superposition of all other partial fields.

\[ \langle h_i \rangle = \langle h_i^0 \rangle + \langle \Delta h_i \rangle = \sum_{i > j}^M \langle h_i^0 \rangle. \]

In the zero noise limit the first contribution is of the order of 1 whereas the others are of the order of \( O(\sqrt{M/N}) \). Therefore as long as the number \( M \) of stored patterns is finite these fluctuations can be neglected. According to equation (17) the order parameter \( m^1 \), which measures the degree of retrieval accuracy of pattern \( I^1 \), is the solution of the implicit equation:

\[ m^1 = \zeta (\beta m^1). \quad (18) \]

The use of general learning rules introduces a new partial field \( \langle h_i^0 \rangle \) and its conjugate order parameter, the uniform activity \( m^0 \). This field competes with the other partial fields and tends to destabilize the memorized patterns. We still focus our attention on the pattern \( I^1 \) and study the interplay between the order parameters \( m^0 \) and \( m^1 \). As above the random field brought about by the other patterns is neglected. Geometrically speaking this amounts to restricting the dynamics of the system to a subspace of phase space defined by \( m^2 = m^3 = \cdots = m^M = 0 \). This approximation is valid for large values of \( m^1 \) and is therefore legitimate to check the stability of the memorization process. It is more dubious for small \( m^1 \) values because the system then tends to mix several patterns (including the uniform activity configuration). The local field therefore reduces to:

\[ \langle h_i \rangle = \langle h^0 \rangle + \langle h^1 \rangle \]

and the dynamic equations to

\[ \frac{d\langle \sigma_i \rangle}{dt} = -\frac{1}{\tau} \left( \langle \sigma_i \rangle - \zeta(\beta(\langle h^0 \rangle + \langle h^1 \rangle)) \right). \quad (19) \]

Using the definitions (15) of order parameters \( m^0 \) and \( m^1 \) the set of \( N \) coupled equations shrinks to the two following equations:

\[ \frac{dm^0}{dt} = -\frac{1}{\tau} \left( m^0 - \frac{1}{N} \sum_i \xi_i \left( \beta(m^0(B\xi_i^0 + MD + D^0) + m^1(A\xi_i^1 + C)) \right) \right) \]

\[ \frac{dm^1}{dt} = -\frac{1}{\tau} \left( m^1 - \frac{1}{N} \sum_i \xi_i \left( \beta(m^0(B\xi_i^0 + MD + D^0) + m^1(A\xi_i^1 + C)) \right) \right) \quad (20) \]

Within a statistical error \( \pm \sqrt{N} \) there are as many values \( \xi_i^1 = +1 \) as values \( \xi_i^1 = -1 \). Therefore, using the odd parity of the sigmoidal function \( \zeta \), the equations (20) can be written:

\[ \frac{dm^0}{dt} = -\frac{1}{\tau} \left( m^0 - \frac{1}{2} \left( \xi(\beta(sm^0 + tm^1)) + \xi(\beta(um^0 + vm^1)) \right) \right) \]

\[ \frac{dm^1}{dt} = -\frac{1}{\tau} \left( m^1 - \frac{1}{2} \left( \xi(\beta(sm^0 + tm^1)) - \xi(\beta(um^0 + vm^1)) \right) \right) \quad (21) \]

with

\[ s = B + MD + D^0; \quad t = C + A; \quad u = -B + MD + D^0; \quad v = C - A. \quad (22) \]

The fixed points of the dynamics are given by:

\[ m^0 = \frac{1}{2} \cdot \left( \xi(\beta(sm^0 + tm^1)) + \xi(\beta(um^0 + vm^1)) \right) \]

\[ m^1 = \frac{1}{2} \cdot \left( \xi(\beta(sm^0 + tm^1)) - \xi(\beta(um^0 + vm^1)) \right) \quad (23) \]

In the limit of low noise the fixed point equations are cast in the following form:

\[ m^0 + m^1 = \text{sgn} (sm^0 + tm^1) \]

\[ m^0 - m^1 = \text{sgn} (um^0 + vm^1). \quad (24) \]

The set of solutions is then restricted to

\[ m^0 = 1 \text{ and } |m^1| = 0 \]

\[ |m^0| = 1 \text{ and } m^1 = 0 \]

\[ m^0 = m^1 = 0. \]
The solutions actually achieved at zero noise can be found by a direct inspection of the equations (24). They depend on the signs of $s, r, t$ and $v$. There are 16 possible situations and therefore as many learning rules. For example let us assume that the four parameters are positive. Then the solution $|m^0| = 1$ and $m^1 = 0$ satisfies the equations (24) whereas the solution $m^0 = 0$ and $|m^1| = 1$ does not. We conclude that a set of positive parameters $s, t, u, v$ leads to the unsatisfactory situation of stable uniform activity (for the stability see below). The solutions of equations (24) for the 16 possible learning rules are collected in table II. A convenient set of parameters is such that $m^0 = 0$ and $|m^1| = 1$ is a solution. Then

$$m^1 = \text{sgn} \ (t \cdot m^1)$$

and

$$m^1 = -\text{sgn} \ (v \cdot m^1).$$

Therefore it is necessary that $t > 0$ and $v < 0$ and there are 4 rules (labelled 1 to 4 in Tab. II) which satisfy these criteria.

The stability of the solutions (for general noise levels) is studied by linearizing the equations of motion around the fixed points $m^0*$ and $m^1*$. Letting

$$x = m^0 - m^0* \quad \text{and} \quad y = m^1 - m^1*,$$

We obtain

$$\frac{dx}{dt} = -\frac{1}{\tau} \cdot (x(1 - k)(s + u) - yk(t + v))$$

$$\frac{dy}{dt} = -\frac{1}{\tau} \cdot (-xk(s - u) + y(1 - k)(t - v))$$

with $k = \frac{B}{2} \xi^*(m^0*, m^1*, \beta)$. $k$ is a positive number. The eigenvalues of the relaxation matrix are given by:

$$\det \begin{vmatrix} -\frac{k}{\tau} (1 - s - u) - \Lambda & \frac{k}{\tau} (t + v) \\ \frac{k}{\tau} (s - u) & -\frac{k}{\tau} (1 - t + v) - \Lambda \end{vmatrix} = 0.$$ 

We find:

$$A \pm = \frac{k}{2 \tau} \left( (s + t + u - v) - \frac{2}{k} \pm \sqrt{(s + t + u - v)^2 - 8(ut - sv)} \right). \quad (25)$$

In the low noise limit the factor $k$ goes to zero exponentially be $\xi$ an error or a tanh function and therefore

$$A = -\frac{1}{\tau}.$$

The fixed points of table II are stable fixed points.

Apart from the conditions already stressed the learning rules are satisfactory if the eigenvalues of the relaxation matrix are real and negative. The eigenvalues are certainly real for rule 3 since $u \cdot t - s, v < 0$. The domain of parameter which corresponds to real eigenvalues is still large for the rules 1 and 2 but it is narrow for rule 4 which therefore must be rejected. As the domains of parameters which correspond to negative eigenvalues are also larger for rule 3 than for rules 1 and 2, rule 3 is by far the best rule the network can achieve.

Let us consider the implications of rule 3 on the experimental learning parameters. $s$ and $u$ are negative. This according to equation (23), compels $M \cdot D + D^0$ to be negative. Such a condition can be achieved by strong nonmodifiable inhibitory interactions. However, due to the factor $M$, the effect of even small positive $D$'s would be catastrophic, hindering any memory storage. $D$ is given by equation (6). $D$ is negative if the process of depotentiation of the excitatory synapses is at least as efficient as the potentiation process. It must be noted that the Cooper rule helps in this respect. On the other hand rule 3 implies:

$$t = C + A > 0 \quad \text{and} \quad v = C - A < 0$$

therefore

$$A > 0 \quad \text{and} \quad |C| < A.$$

According to table I the parameters $A$ and $C$ are given by:

$$A = n^- \cdot (a_{ex}^+ - a_{ex}^-) + n^- \cdot a_{in}^+$$
Table II. — Stable solutions of equation (24) according to the signs of the experimental learning parameters: the best rule is rule 3.

<table>
<thead>
<tr>
<th>Rule nb</th>
<th>s</th>
<th>t</th>
<th>u</th>
<th>v</th>
<th>Stable solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>( m^0 = 0,</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>( m^0 = 0,</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>( m^0 = 0,</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>( m^0 = 0,</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>(</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>(</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>(</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>9</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>( m^0 = m^1 = 0 )</td>
</tr>
</tbody>
</table>

\[ C = n^+ \cdot (a^+_{ex} - a^-_{ex}) - n^- \cdot a^+_{in} \quad (6') \]

All the factors entering \( A \) are positive and the first condition \( A > 0 \) is fulfilled. The second condition \( |C| < A \) is also met if the Cooper rule is an active process. It is worth noting that no conditions are derived from the parameter \( B \).

4. Storing an infinite number of patterns.

(By infinite we mean that \( M = O(N) \).)

When the number of stored configurations increases the fluctuations of the random field, the amplitude of which behaves as \( \sqrt{M/N} \), overcomes the coherent term disabling the memory storage properties of the network. This happens for \( M = N \). The theory of the storage capacity of neural networks with symmetrical synapses has been carried out by Amit et al. [16]. Their derivation uses the tools of statistical mechanics for spin glasses, the replica trick in particular. Here, due to the asymmetry of interactions, this technique is inoperative. We prefer a mean field approach which is both efficient and transparent.

First of all, however, we discuss the scaling properties of the memory storage capacities of the network with respect to the various learning parameters \( A, B, C, D \) and \( D^0 \). This can be achieved by checking, as usual, the stability of states \( \xi_i^1 \) of a given pattern \( I^1 \). The stability conditions are:

\[ \xi_i^1 \cdot h_i(I^1) > 0; \quad \forall i. \]

According to equations (7) and (10), the inequalities can be written:

\[ \xi_i^1 \cdot h_i(I^1) = A \left( 1 + O \left( \frac{1}{\sqrt{N}} \right) \right) + B \left( 1 + O \left( \frac{1}{\sqrt{N}} \right) \right) + C \left( \xi_i^1 + O \left( \frac{1}{\sqrt{N}} \right) \right) + (MD + D^0) \cdot O \left( \frac{1}{\sqrt{N}} \right) > 0 \]

where \( O(x) \) is a random number the probability distribution of which is Gaussian with mean square deviation \( x \). We deduce that the states are destabilized when:

\[ 1 \leq \frac{M}{N} \left( 1 + \left( \frac{B}{A} \right)^2 + \left( \frac{C}{A} \right)^2 + \left( \frac{MD + D^0}{A} \right)^2 \cdot \frac{1}{M} \right) + \left( \frac{C}{A} \right)^2. \]

Let us consider the effects of the learning parameters separately.

We first observe that \( C \) and \( A \) must obey:

\[ |C| < A; \quad A > 0 \]

if the patterns are to be stabilized at all even for \( M \), the number of patterns, of the order of 1. \( C \) is the most destabilizing term.

Then comes \( D \). A non-zero value of \( D \) limits the memory capacity to

\[ M = \sqrt{N}. \]

The parameter \( B \) is less harmful since it does not change the scaling behaviour of memory capacity of the Hopfield model.

\[ M = N. \]

Finally \( D^0 \), assuming that all the other learning parameters vanish, has no influence on the limit storage capacity whatsoever. This parameter could however play an important role if it can be adjusted
in such a way as to cancel out the \( M \cdot D \) term thus avoiding the \( M = \sqrt{N} \) scaling behaviour.

We now turn to a more precise analysis.

Let us assume, as we did in the last section, that the system condenses in (retrieves a) pattern \( I^1 \). The local field (the stochastic average of the membrane potential) on a neuron \( i \) is given by:

\[
\langle h_i \rangle = \langle h_i^0 \rangle + \langle h_i^1 \rangle + \langle \Delta h_i \rangle
\]

\[
\langle \Delta h_i \rangle = \sum_{\mu > 1} \langle h_i^\mu \rangle.
\]  

(26)

A particular set of patterns \( I^\mu (\mu > 1) \), gives rise to a defined fluctuation \( \langle \Delta h_i \rangle \). But we are interested in the general properties of neural networks that is in properties which do not depend on particular realizations of stored patterns. In the following derivation we shall use a central hypothesis, called the self-averaging assumption. It states that all statistical observables are sample independent (in the limit of large \( N \)‘s). This is a strong hypothesis which means that the properties of the system, for example its memory storage capacity, can be obtained by averaging over a number of pattern realizations or, provided they are properly scales, are independent on the size of the system for a given realization.

Actually the self-averaging hypothesis breaks down for spin glasses and for neural networks alike. However for symmetrical neural networks, the improvements brought about when one gets rid of self-averaging, using a replica symmetry breaking scheme, do not change the quantitative results significantly [17]. This gives some confidence that the self-averaging hypothesis can be used safely in asymmetrical neural networks.

Let \( g \) be the label of a particular realization of patterns \( I^\mu (\mu > 1) \) and \( N_g \) the number of realizations. The sample average of an observable \( O \), \( \bar{O} \) which must not be confused with the statistical average \( \langle O \rangle \) given by equation (9), is defined by:

\[
\bar{O} = \frac{1}{N_g} \cdot \sum_{g=1}^{N_g} O(g).
\]  

(27)

For example the sample averages of the order parameters \( m^0 \) and \( m^1 \) are written:

\[
\bar{m}^0 = \frac{1}{N_g} \cdot \sum_g \sum_i \langle \sigma_i \rangle g = \frac{1}{N_g} \cdot \sum_g \sum_i \langle \sigma_i \rangle g
\]

\[
= \frac{1}{N} \sum_i \langle \sigma_i \rangle
\]

\[
\bar{m}^1 = \frac{1}{N_g} \cdot \sum_i \xi_i \langle \sigma_i \rangle = \frac{1}{N} \cdot \sum_i \xi_i \cdot \langle \sigma_i \rangle.
\]  

(28)

Similarly

\[
\bar{m}^\mu = \frac{1}{N_g} \cdot \sum_i \sum_g \xi_i^\mu (g) \langle \sigma_i \rangle g = \frac{1}{N} \cdot \sum_i \xi_i^\mu \cdot \langle \sigma_i \rangle.
\]  

(29)

In this expression the index \( \mu \) looses its signification since \( I^{\mu (g)} \) can be any arbitrary pattern. Therefore \( \bar{m}^\mu \) is a constant. On the other hand this sample average must be identical, according to the self-averaging hypothesis, to the average \( m^\mu \) computed on one sample, no matter how large, for a definite realization of the \( I^\mu \)’s. The conclusion is that

\[
m^\mu = \frac{1}{N} \cdot \sum_i \xi_i^\mu \langle \sigma_i \rangle = \bar{m}^\mu = 0.
\]  

(30)

This result is not as trivial as it seems: it implies a decoupling between the average activity \( \langle \sigma_i \rangle \) and the local memory of pattern \( I^\mu \) on a neuron \( i \) in the making of the partial field \( \langle h_i^\mu \rangle \).

The same arguments leads to

\[
\bar{h}_{i}^\mu = 0; \quad \mu > 1, \mu' > 1, \mu \neq \mu'.
\]  

(31)

The sample average of the local field can therefore be easily computed:

\[
\langle h_i \rangle = \langle h_i^0 \rangle + \langle h_i^1 \rangle + \sum_{\mu > 1} \langle h_i^\mu \rangle
\]

\[
= m^1 (A \xi_i^1 + C) + m^0 (B \xi_i^1 + MD) + D^0
\]  

(32)

The local field varies from place to place. According to the self-averaging hypothesis the local field distribution \( P(\langle h_i \rangle) \) can be reconstructed using a sample averaging procedure. Due to the stochastic character of the memorized patterns the field distribution is Gaussian and is fully determined by the mean value equation (32) and the mean square deviation

\[
P(\langle h_i \rangle) = \frac{1}{\sqrt{2 \pi} \cdot \langle \Delta h_i \rangle^2} \times \exp \left( -\frac{(\langle h_i \rangle - \bar{h}_i)^2}{2 \cdot \langle \Delta h_i \rangle^2} \right)
\]  

(33)

The sample average of the order parameter \( \mu \) is calculated as follows:

\[
\langle \Delta h_i \rangle^2 = \langle \langle h_i \rangle - \bar{h}_i \rangle^2 \rangle = \langle h_i^2 \rangle - \langle h_i \rangle^2
\]

\[
= \sum_{\mu > 1} \langle h_i^\mu \rangle^2
\]

\[
= \sum_{\mu} \left( (A \xi_i^1 + C) \cdot m^\mu + B \xi_i^1 \cdot m^0 \right)^2
\]

\[
= \sum_{\mu} \left( (A^2 + C^2) m^\mu + B^2 \xi_i^1 \right)^2
\]

\[
= \lambda r + \nu \bar{m}^0
\]  

(34)

with

\[
\lambda = A^2 + C^2, \quad \nu = B^2 M
\]

and

\[
r = \sum_{\mu > 1} \bar{m}^\mu
\]

\( r \) is the order parameter introduced by Amit et al. (AGS parameter).
In equation (34) we have not to take the fluctuations of \( m^0 \) into account since the essence of mean field theories is to ignore the fluctuations of order parameters, here \( m^0 \) and \( m^1 \).

For a given realization \( g \) of memorized patterns \( I^\mu (\mu > 1) \), the local polarization is given by

\[
\langle \sigma_i \rangle_g = \xi (\beta \langle h_i \rangle_g)
\]

and therefore the sample average of the local statistical average of the activity is:

\[
\langle \sigma_i \rangle = \int d(\langle h_i \rangle \cdot P(\langle h_i \rangle) \cdot \xi (\beta \langle h_i \rangle)).
\]  

(35)

Introducing equations (32), (33) and (34) into equation (35) and letting

\[
z = (\langle h_i \rangle - m^0(B\xi_1 + MD + D^0) - m^1(A\xi_1 + C))/(\sqrt{\lambda r + \nu m_0^2})
\]

leads to:

\[
\langle \sigma_i \rangle = \frac{1}{\sqrt{2}\pi} \int dz \cdot \exp \left(-\frac{z^2}{2} \cdot \xi (\beta (m^0(B\xi_1 + MD + D^0))) + \ldots \right.

\[
+ m^1(A\xi_1 + C) + z \sqrt{\lambda r + \nu m_0^2}) \right).
\]

(36)

From the averaged local activity it is easy to derive the sampled averaged order parameters \( m^0 \) and \( m^1 \). The uniform activity using equations (36) and (28) reads:

\[
\overline{m^0} = \frac{1}{2\sqrt{2}\pi} \int dz \cdot \exp \left(-\frac{z^2}{2} \cdot \left( \xi (\beta (sm^0 + tm^1 + z \sqrt{\lambda r + \nu m_0^2})) + \ldots \right.

\[
+ \xi (\beta (um^0 + vm^1 + z \sqrt{\lambda r + \nu m_0^2})) \right).
\]

(37)

whereas the retrieval efficacy for pattern 1 reads:

\[
\overline{m^1} = \frac{1}{2\sqrt{2}\pi} \int dz \cdot \exp \left(-\frac{z^2}{2} \cdot \left( \xi (\beta (sm^0 + tm^1 + z \sqrt{\lambda r + \nu m_0^2})) - \ldots \right.

\[
- \xi (\beta (um^0 + vm^1 + z \sqrt{\lambda r + \nu m_0^2})) \right).
\]

(38)

To find a closed set of equations between \( \overline{m^0} \), \( \overline{m^1} \) and \( r \) it is convenient to introduce the Edwards-Anderson (EA) parameter \( q \) defined by:

\[
q = \frac{1}{N} \cdot \sum_i \langle \sigma_i \rangle^2.
\]

(39)

According to equations (36) and (39), \( q \) is given by:

\[
q = \frac{1}{2\sqrt{2}\pi} \int dz \cdot \exp \left(-\frac{z^2}{2} \cdot \left( \xi (\beta (sm^0 + tm^1 + z \sqrt{\lambda r + \nu m_0^2})) + \ldots \right.

\[
+ \xi (\beta (um^0 + vm^1 + z \sqrt{\lambda r + \nu m_0^2})) \right).
\]

(40)

Finally one derives a relation between \( q \), \( r \) and \( \overline{m^0} \). But this can be carried out analytically only when \( \xi \) is a \( \tanh \) (thermodynamical) function. The following calculations are restricted to this response function.

Assuming that the system condenses in pattern \( I^1 \) we compute the polarization fluctuation \( m^\mu \) created by the contribution \( \langle h^\mu \rangle (\mu > 1) \), to the field \( \langle h^\mu \rangle \). To 1st order in \( \langle h^\mu \rangle \) the local polarization due to the pattern \( I^\mu \) is given by

\[
\langle \sigma_i \rangle_{\mu} = \langle \sigma_i \rangle_g + \beta \langle h^\mu \rangle \cdot \left(1 - \tanh^2 \left(\beta \sum_{\mu \neq \mu} \langle h^\mu \rangle \right)\right)
\]

(41)
where $\langle \sigma_i \rangle_0$ is the activity of neuron $i$ in the absence of the pattern $I^\mu$. With equation (41) the fluctuation $m^\mu$

$$m^\mu = \frac{1}{N} \sum_i \xi_i^\mu \langle \sigma_i \rangle_\mu$$

is transformed into:

$$m^\mu = \frac{1}{N} \sum_i \xi_i^\mu \langle \sigma_i \rangle_0 + \beta \sum_i \left(1 - \tanh^2 \beta \sum_{\mu \neq \mu} \langle h_i^\mu \rangle \right) \cdot \left(m^\mu (A + C \xi_i^\mu) + m_0^0 B \right).$$

This expression is squared and the sample average is carried out:

$$m^\mu \left( \left(1 - \frac{\beta}{N} \cdot A \cdot \sum_i \left(1 - \tanh^2 \beta \langle h_i \rangle \right) \right) \right)^2 + \beta \frac{C^2}{N^2} \sum_i \left(1 - \tanh^2 \beta \langle h_i \rangle \right)^2 + \cdots$$

$$+ m^\mu \beta^2 B^2 \left( \frac{1}{N} \sum_i \left(1 - \tanh^2 \beta \langle h_i \rangle \right) \right)^2 = \frac{1}{N^2} \left( \sum_i \langle \sigma_i \rangle^2 \right).$$

Using the self-averaging hypothesis we can write

$$\frac{1}{N} \sum_i \left(1 - \tanh^2 \beta \langle h_i \rangle \right) = 1 - q \quad (42)$$

and summing up over all patterns we obtain the desired relation:

$$r = \sum_{\mu \neq 1} m^\mu = \frac{M(q/N - \overline{m}^0 \beta^2 B^2(1-q)^2)}{(1 - \beta A(1-q))^2 + \beta^2 C^2(1-q)^2}. \quad (43)$$

The equations (37), (38), (40) and (43), which couple the four order parameters $\overline{m}^0$, $\overline{m}^1$, $q$ and $r$, solve the problem of the memory storage properties of asymmetrical neural networks. In the usual models where $B = C = D = D^0 = 0$ the set reduces to the three equations linking $\overline{m}^1$, $q$ and $r$ derived by Amit et al.

A full discussion of these equations is very demanding. We shall focus on the sole influence of parameter $C$ (assuming that $B = MD + D^0 = 0$). The idea is to study the influence of the most disturbing term and also to make contact with the analysis of Amit et al.

The equations then reduce to:

$$\overline{m}^1 = \frac{1}{2 \sqrt{2 \pi}} \int dz \cdot \exp - \frac{z^2}{2} \cdot \left( \zeta (\beta (im^1 + z \sqrt{\lambda r})) - \zeta (\beta (vm^1 + z \sqrt{\lambda r})) \right)$$

$$q = \frac{1}{2 \sqrt{2 \pi}} \int dz \cdot \exp - \frac{z^2}{2} \cdot \left( \zeta^2 (\beta (im^1 + z \sqrt{\lambda r})) - \zeta^2 (\beta (vm^1 + z \sqrt{\lambda r})) \right)$$

$$r = \frac{\alpha q}{(1 - \beta A(1-q))^2 + (\beta C(1-q))^2}$$

with

$$\alpha = M/N \quad (\text{and } t = A + C, v = -A + C, \lambda = A^2 + C^2). \quad (43')$$

The memory storage capacity is computed in the limit of low noise level. The equations (43) become (see Appendix B):

$$r \left( \left(1 - \frac{2}{\pi r(1 + \varphi^2)} \cdot \exp - \frac{m^1}{2r} \cdot \cosh \left( \frac{\varphi}{1 + \varphi^2} \right) \right)^2 + \cdots \right)$$

$$+ \frac{2 \varphi^2}{\pi \varphi (1 + \varphi^2)} \cdot \exp - \frac{m^1}{2r} \cdot \cosh^2 \left( \frac{\varphi}{1 + \varphi^2} \right) = \alpha \quad (44)$$

where the parameter $\varphi$ is defined by $\varphi = |C|/A$. 


For low values of $\alpha$ these equations have a solution $m_1 \neq 0$ meaning that the retrieval process is efficient. This solution is lost in a steep (1st order) transition. $(1 - m_1)/2$ is the percent errors in the retrieval process. Figure 1 shows the variations of the percent error for some values of $\varphi$. The behaviour of this quantity at critical values of $\alpha$ is displayed in figure 2. The behaviour of the critical value $\alpha_c$ as a function of $\varphi$ is given in figure 3. We observe that the neural network looses its memory storage abilities as soon as $\varphi = |C|/A > 1$.


The memory storage capacities of asymmetrical networks have been studied numerically on systems comprising 200 and 400 neurons. A parallel deterministic (noiseless) dynamics is used. (The algorithm of synchronous updating of neural states is efficient at low noise level because it limits the computation of local fields variations to those units which have changed their states during the last time step of the dynamics.)

After $M$ random patterns $I^\mu$, $\mu = 1, \ldots, M$, have been memorized according to equation (7), their stabilities are checked one by one by setting the system in state $I^\mu$ at time origin and letting the network relax. 10 time steps are generally enough for the system $t_0$ reach a fixed point. The Hamming distances between the initial states $I^\mu$ and the final states are computed. Their average determines the order parameter of the condensed (Mattis) phases. The whole process is repeated to improve the accuracy of the averages and the computation is carried out for increasing values of $M$.

We first of all observe that the order parameter variation agrees the theoretical prediction of Amit et al. when $A = 1$, $B = C = M$, $D + D^0 = 0$ but that it does not show any 1st order transition. This is not surprising since the 1st order transition is associated with the onset of a disordered stable phase which the algorithm fails to explore (for a discussion on remanence effects, see Kinzel [18]). This lack of discontinuity makes the determination of the memory storage capacity difficult.

A way out is to use the theoretical critical magnetization $m_c = 0.97$: one finds $\alpha_c = 0.145 \pm 0.005$. Crisanti et al. [17] have shown that breaking the replica symmetry actually yields slightly increased values of $\alpha_c$.

The effects of learning parameters $B$, $C$ and $D$ on the memory storage capacity have been studied separately.

Figure 4 shows how the critical capacity varies for increasing values of $|C|/A$ ($B = D = 0$). The simulations are carried out on a 400 neurons network. The critical capacities $\alpha_c(C)$ are determined using the theoretical critical magnetizations. The agreement with the analytical approach is fair. Figure (4) also shows a more direct comparison between theory and simulation by displaying the values of $\alpha$ corresponding to a magnetization $m = 0.97$.

Next we observed the influence of parameter $B/A$ ($C = D = 0$). As expected this influence is sensitively weaker than that of parameter $C$. 

![Figure 1](image1.png)

Fig. 1. — Variation of the percent error in the retrieval process of one stored pattern as a function of the parameter $\alpha = M/N$, the ratio of the total number of stored patterns to the number of neurons. The various curves correspond to different values of learning parameter $|C|/A$; from left to right = 0.90; 0.80; 0.65; 0.50; 0.35; 0.20 and 0.00.

![Figure 2](image2.png)

Fig. 2. — Variation of $m_c$, the order parameter associated with the retrieved pattern, computed at the critical values of $\alpha$ as a function of learning parameter $|C|/A$.

![Figure 3](image3.png)

Fig. 3. — Phase diagram of retrieval efficiency. The upper curve corresponds to the destabilization of memorized states. The lower curve to a retrieval with 97% efficiency.
Fig. 4. — Phase diagram of retrieval efficiency obtained by computer simulation on a 400 neurons network \((B = D = 0)\). The circles correspond to capacities observed at critical magnetizations (given by Fig. 2). The crosses correspond to capacities associated with 97% retrieval efficiencies. The agreement with figure 3 is fair although systematic deviations towards lower values are observed (except for \(C = 0\)). The statistical errors amount to ± 0.01 for the upper curve and to ± 0.005 for the lower curve.

At least we have studied the effect of parameter \(D\) on the retrieval properties of a network comprising 200 neurons. For networks of that size the influence of \(D\) is not as catastrophic as expected: whereas the retrieval quality of all patterns is spoiled evenly when \(B\) and \(C\) are increased, \(D\) (be it positive or negative) destabilizes the patterns selectively. The retrieval of some patterns remains excellent while the other patterns are completely lost. The averaging process has therefore little meaning and the relative number of faithfully retrieved patterns would be a more appropriate parameter. It must be noted that negative \(D\)'s are somewhat less disturbing from this point of view than positive one's anyway.

6. Discussion.

It is surprising that the two approximations used throughout this article, namely the mean-field approximation and the self averaging hypothesis, which are known to be both incorrect separately, yield the correct results, withing the replica symmetry analysis, when they are used together. This observation applies to the Hopfield model but also to the S.K. (Sherrington, Kirkpatrick) spin glass model. We now present some arguments which could be shed some light on this problem.

Let \(\{C^g_{ij}\}\) be the set of interactions for a given realization \(\Sigma^g\) of the system \(\Sigma\). The corresponding asymptotic probability distribution is \(\rho^g(I)\) \((= \rho^g(I, \infty))\) which can be viewed as the relative number of times a state \(I\) is found in an ensemble of \(N^s\) stationary systems identical to \(\Sigma^g\). We then consider \(N^g\) different realizations of \(\Sigma\), each characterized by a different set of interactions \(\{C^g_{ij}\}\), chosen so as to keep invariant some parameters of the distributions. In memory (Hopfield) models the interactions are built using sets of memorized patterns which are randomly generated except for one of the patterns (say \(I^0\)) which is left unchanged. Similarly the interactions of the S.K. model are chosen at random with the constraint that the two first moments of the distributions do not vary. The collection of the \(N^g\) ensembles of system \(\Sigma^g\) forms a new large ensemble of \(N^s\cdot N^g\) systems, the asymptotic distribution of which is \(\bar{\rho}(I)\). \(\bar{\rho}(I)\) is the relative number of times a state \(I\) is found in this large ensemble. The self-averaging hypothesis states that the statistical properties of \(\Sigma\) can be computed using \(\bar{\rho}(I)\). If \(N^s\) is identical for all systems \(\Sigma^g\), \(\bar{\rho}(I)\) is given by:

\[
\bar{\rho}(I) = \frac{1}{N^g} \sum_g \rho^g(I).
\]  

(45)

Let us focus our attention on unit \(i\). Its activity is

\[
\langle \sigma_i \rangle = \sum_I \bar{\rho}(I) \cdot \sigma(I) = \frac{1}{N^g} \sum_g \langle \sigma_i \rangle_g.
\]

On the other hand we know that the average activity of \(i\) is determined by its local field \(h_i\). It is therefore natural to carry out the summation over the \(N^s\cdot N^g\) systems so as to collect the systems giving rise to the same field \(h_i\). This is summarized in the following compound probability formula:

\[
\sum_I \bar{\rho}(I) \cdot \sum_{\sigma_i} \rho^g(h_i) \cdot \rho(\sigma_I|h_i)
\]

where \(\rho^g(h_i)\) is the probability to find a local field \(h_i\) and \(\rho(\sigma_i|h_i)\) the probability for \(i\) to be in state \(\sigma_i\) given \(h_i\). \(\rho^g(h_i)\) is determined by the following equation:

\[
\rho^g(h_i) = \frac{1}{N^g} \sum_I \rho^g(I) \times \delta\left(h_i - \sum_j C^g_{ij} \sigma_j(I)\right).
\]  

(46)

This equation is a sum of random contributions each arising from a different realization \(g\). The distribution \(\rho^g(h_i)\), for large values of \(N^g\), is therefore Gaussian and the regularities of the realizations are trapped into the two parameters which characterize the distributions namely its first and second moments, \(\langle h_i \rangle\) and \(\langle \Delta h^2_i \rangle\). Using equations (45) and (46), those quantities are given by:

\[
\langle h_i \rangle = \sum_{h_i} \rho^g(h_i) \cdot h_i = \frac{1}{N^g} \sum_I \rho^g(I) \cdot h^g(I) = \frac{1}{N^g} \sum_g \langle h_i \rangle_g.
\]
Finally the average activity of unit $i$ is given by:

$$\langle \sigma_i \rangle = \sum_{h_i} \rho(h_i) \sum_{\sigma_i} \rho(\sigma_i | h_i) \sigma_i$$

$$= \sum_{h_i} \rho(h_i) \cdot \xi(\beta h_i)$$

$$= \frac{1}{\sqrt{2\pi} \langle \Delta h_i \rangle} \times$$

$$\int dh_i \cdot \exp \left( - \frac{(h_i - \langle h_i \rangle)^2}{2 \langle \Delta h_i \rangle} \right) \cdot \xi(\beta h_i) . \quad (48)$$

It must be noted that this formula is valid for any form of (symmetrical) sigmoid function. It only appeals to the self-averaging hypothesis. Actually the mean-field approximation only comes in the computation of the moments of the Gaussian distribution. Let us apply this approach to the S.K. model. The interactions are given by:

$$C_{ij} = C_{ij} + \Delta C_{ij} = \bar{C} + \Delta C_{ij}$$

with

$$\Delta C_{ij}^2 = \frac{\Delta C^2}{N}$$

$\bar{C}$ and $\Delta C^2$ are fixed parameters.

The first moment is computed

$$\langle h_i \rangle = \frac{1}{N^\theta} \sum_g \left\{ \sum_j C_{ij} \sigma_j \right\}_g$$

$$= \bar{C} \cdot \frac{1}{N^\theta} \cdot \sum_g \langle \sigma_j \rangle_g + \frac{1}{N^\theta} \cdot \sum_g \langle \Delta h_i \rangle_g$$

$$= \bar{C} \cdot m + \langle \Delta h_i \rangle = \bar{C} \cdot m ; \quad (49)$$

since the local field deviation $\langle \Delta h_i \rangle_g$ varies at random from one realization to another realization.

The second moment is given by:

$$\langle \Delta h_i \rangle = \langle h_i^2 \rangle - \langle h_i \rangle^2$$

$$= \frac{1}{N^\theta} \cdot \sum_g \sum_j \rho^g(I) \left( \sum_j \Delta C_{ij} \sigma_j(I) \right)^2$$

$$= \frac{1}{N^\theta} \cdot \sum_g \sum_{jk} \Delta C_{ij} \Delta C_{jk} \langle \sigma_j \sigma_k \rangle_g . \quad (50)$$

There is little chance for a random realization to display critical properties. This argument is used to justify a mean-field (or clustering) approximation which allows the activities to be decoupled

$$\langle \sigma_j \sigma_k \rangle_g = \langle \sigma_j \rangle_g \langle \sigma_k \rangle_g .$$

Then

$$\langle \Delta h_i \rangle = \frac{1}{N^\theta} \sum_g \sum_{jk} \Delta C_{ij} \Delta C_{jk} \langle \sigma_j \sigma_k \rangle_g$$

$$= \frac{1}{N^\theta} \sum_g \sum_{jk} \Delta C^2 \langle \sigma_j \rangle_g^2 = \bar{C}^2 \cdot q$$

with

$$q = \frac{1}{N} \sum_i \langle \sigma_i \rangle^2$$

since self-averaging averages out the off-diagonal terms to zero. The S.K. equations are derived straightforwardly from equations (48), (49) and (50). The conclusion is that self-averaging and mean-field approximation used together actually yield the results derived from the symmetrical replica techniques. But this approach is more general since it can be applied to situations with no energy available and to dynamics driven by transition probabilities of any type. It is worth noting that an expression similar to equation (48) has been derived by Mezard et al. [19] in a theory (the cavity method) of spin glasses which does not appeal to the self-averaging hypothesis. Assuming that the clustering approximation is valid Mezard proved that the local field distribution (for a thermal ensemble) is Gaussian. Moreover if the conditionnal probability $\rho(\sigma_i | h_i)$ is of the Boltzman type: $\rho(\sigma_i | h_i) = \exp \beta \sigma_i h_i$ the TAP (Thouless, Anderson, Palmer) equations, giving the local magnetization, are recovered. However when average collective properties are needed, sample averaging is a necessity and the TAP equations loose their usefulness.

7. Summary and conclusions.

The models of long term memory storage studied so far have been criticized on the grounds of the symmetry of the interactions they use. The hypothesis of symmetric synapses are indeed of little biological relevance. This is nevertheless a necessity if a statistical mechanics approach is to be used because the concept of energy is otherwise lost. Its instigators argue that the symmetry hypothesis is a convenient way of studying memory and that its absence does not destroy the very mechanism of memory. Simulations carried out on networks comprising deleted synapses support this conjecture [20, 21].

In this article we also get rid of the symmetry hypothesis but asymmetry is assumed to originate in learning mechanisms rather than in attrition of synapses. We strive to determine whether the exper-
imental data on synaptic efficacies modifications are compatible with the ability to store information.

The homosynaptic learning rules, which are the one we considered, depends on four learning parameters:

- $A$, which memorizes the correlated activities of pre and post-synaptic neurons;
- $B$, which memorizes the post-synaptic activity;
- $C$, which memorizes the pre-synaptic activity;
- and $D$, which is activity independent.

The best local learning rule is $A > 0$, $B = C = D = 0$. This is the rule proposed by Hopfield. $A < 0$ is an anti-learning rule. These conditions are rather restrictive and not likely to be obeyed by biological systems.

We find that the most sensitive parameter is $\varphi = |C|/A$. It is necessary that $\varphi < 1$ if the system is to display any memorization properties.

The effect of $|D|/A$ is to limit the memory storage capacity to $M \approx \sqrt{N}$.

This shrinking of the capacity with respect to that of Hopfield model can be avoided by introducing an extra term $D^0$ which cancels out the effect of $D$.

The parameter $|B|/A$ does not change the scaling behaviour $M = N$ of the Hopfield model.

We show analytically that memorization is a robust process and that it can accommodate a number of learning rules.

The derivation rests upon two assumptions which are often used in statistical mechanics: the mean-field approximation and the self-averaging hypothesis. The results we obtain generalize those of Amit et al. without appealing to the replica trick.

The experimental data so far available on the synaptic plasticity fit the criteria of efficient learning processes. According to equations (6) and (6'), the condition the experimental learning parameters have to fulfil are the following:

$$a_{+ + +} > 0$$ (The Hebb rule)
$$a_{- - -} < 0$$ (The Anti-Hebb rule)
$$a_{+ + -} - a_{+ - +} - a_{- + +} + a_{- - -} > 0$$ (The Cooper rule)
$$a_{m +} - a_{m -} > a_{m + +}$$

Obviously these results have to be reexamined if synaptic saturation phenomena are taken into account. It is likely that the conclusions will remain valid at least qualitatively.

Appendix A.

HETEROSONAPTIC JUNCTIONS AND HIGHER ORDER SYNAPSES [21]. — Heterosynaptic junctions are processes where the activity of a synapse modified the efficacy of an other adjacent synapse. The learning rule of an heterosynaptic junction is a pseudo-Boolean function of the three neurons say $i$, $j$, $k$, it involves:

$$\Delta C_{ijk}(\xi^i_+, \xi^j_+, \xi^k_+) \propto A^1 \xi^i_+ \xi^j_+ \xi^k_+ + A^2 \xi^i_+ \xi^j_+ \xi^k_+ + A^3 \xi^i_+ \xi^j_+ \xi^k_+ + \cdots$$

The learning parameters, $A^1$, ..., are related to the experimental learning parameters $a^{+++}$, ..., by:

$$A^1 = \frac{1}{8} (a^{+++} - a^{---} - a^{++-} - a^{+-+} + a^{-+ +} + a^{+ - +} + a^{---})$$

Higher order synapses can also arise from usual homosynaptic contacts if one chooses to focus attention on a subset of relevant neurons. Let us assume that the neuronal population is made of two types of neurons: type 1 neurons are the relevant population and type 2 neurons is an ancillary population. All contacts are homosynaptic. We restrict the discussion to noiseless networks.

If the synapses are symmetrical the dynamics of the network is driven by an energy function. A way to get rid of the activity of type 2 neurons is to let it relax for a given configuration of type 1 neurons. It is always possible to express the energy of the relaxed configurations as a pseudo-Boolean function of the state of type 1 neurons only. This function, however, depends in general on high order terms. For example if each neuron of type 2 interacts with three neurons of type 1, the effective energy of system 1 comprises terms of third order as in (symmetrical) heterojunctions.

A similar derivation can be carried out for asymmetrical synapses where no energy function can be defined provided the dynamics is synchronous. Indeed we can fix the states of system 1, let the system 2 relax and observe the induced changes for the states of system 1. This determines a mapping between successive states of system 1 i.e. its effective dynamics. It is always possible to build interactions between the units of system 1 only which reproduce the mapping. Here too, the price to pay is the introduction of (asymmetrical) interactions of higher orders.

The treatment of noisy networks demands a more thorough analysis. These procedures are reminiscent of the decimation technique introduced in the theory of critical phenomena.
Appendix B.

DERIVATION OF EQUATIONS (44). — In the zero noise limit $B^{-1} = 0$, the sigmoid function $\zeta$ is transformed into a step function:

$$\zeta(x) \rightarrow u(x); \quad u(x > 0) = 1, \quad u(x < 0) = -1, \quad u(0) = 0$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz \cdot \exp -\frac{z^2}{2} \cdot \zeta(\beta (\bar{m}^{-1} + z \sqrt{\lambda r})) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz \cdot \exp -\frac{z^2}{2} \cdot \left( u\left(z + \frac{\bar{m}^{-1}}{\sqrt{\lambda r}}\right) + 1 - 1 \right) = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{d\bar{m}}{\sqrt{\lambda r}} \exp -\frac{\bar{m}^2}{2} - 1 = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} \exp -y^2 dy = \text{erf} \left( \frac{\bar{m}^{-1}}{\sqrt{2\lambda r}} \right)$$

and therefore

$$\bar{m}^{-1} = \frac{1}{2} \left( \text{erf} \left( \frac{\bar{m}^{-1}}{\sqrt{2\lambda r}} \right) - \text{erf} \left( \frac{v\bar{m}^{-1}}{\sqrt{2\lambda r}} \right) \right). \quad (B.1)$$

On the other hand

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz \cdot \exp -\frac{z^2}{2} \cdot \tanh^2 \left( \beta (\bar{m}^{-1} + z \sqrt{\lambda r}) \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz \cdot \exp -\frac{z^2}{2} \cdot (\tanh^2 \left( \beta (\bar{m}^{-1} + z \sqrt{\lambda r}) - 1 + 1 \right)) = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dz \cdot \exp -\frac{z^2}{2} \cdot \frac{1}{\cosh^2 \left( \beta (\bar{m}^{-1} + z \sqrt{\lambda r}) \right)} + 1$$

which with

$$x = \beta (\bar{m}^{-1} + z \sqrt{\lambda r})$$

is transformed into:

$$1 - \int_{-\infty}^{+\infty} dx \cdot \frac{1}{\beta \sqrt{2\pi \lambda r}} \cdot \exp -\frac{(x - \beta \bar{m})^2}{2 \beta^2 \lambda r} \cdot \frac{1}{\cosh^2 x} = 1 - \frac{1}{\beta \sqrt{2\pi \lambda r}} \cdot \exp -\frac{(\bar{m})^2}{2 \lambda r} \int_{-\infty}^{+\infty} dx \cdot \frac{1}{\cosh^2 x} \cdot \frac{1}{\cosh x} \cdot \frac{1}{\cosh x}. \quad (B.2)$$

For $\beta^{-1} = 0$ the expression (B.2) becomes:

$$1 - \frac{1}{\beta \sqrt{2\pi \lambda r}} \cdot \exp -\frac{(\bar{m})^2}{2 \lambda r} \int_{-\infty}^{+\infty} dx \cdot \exp -\frac{(\bar{m}^{-1})^2}{2 \lambda r} = 1 - \frac{1}{\beta \sqrt{2\pi \lambda r}} \cdot \exp -\frac{(\bar{m}^{-1})^2}{2 \lambda r}$$

whence

$$q = 1 - \frac{1}{\beta \sqrt{2\pi \lambda r}} \cdot \left( \exp -\frac{(\bar{m})^2}{2 \lambda r} + \exp -\frac{(\bar{m}^{-1})^2}{2 \lambda r} \right). \quad (B.3)$$

From equation (B.3) we obtain

$$\beta (1 - q) = \sqrt{\frac{1}{2\pi \lambda r}} \cdot \left( \exp -\frac{(\bar{m})^2}{2 \lambda r} + \exp -\frac{(\bar{m}^{-1})^2}{2 \lambda r} \right)$$
which yields

\[
  r \cdot \left( \frac{1 - \frac{A}{\sqrt{2 \pi \lambda r}}}{\exp \left( \frac{- (\ln r)^2}{2 \lambda r} \right) + \exp \left( \frac{- (\ln r_0)^2}{2 \lambda r} \right)} \right)^2 + \ldots + 
  + \left( \frac{C}{\sqrt{2 \pi \lambda r}} \cdot \exp \left( \frac{- (\ln r_0)^2}{2 \lambda r} \right) \right)^2 = \alpha 
\]

(B.4)

The equations (44) are recovered by introducing the parameter \( \varphi = \frac{|C|}{A} \) into the equations (B.1) and (B.4).

References