Journal of Statistical Physics, Vol. 43, Nos. 3/4, 1986

A Biologically Constrained Learning Mechanism in Networks of Formal Neurons

L. Personnaz,[1] I. Guyon,[1] G. Dreyfus,[1] and G. Toulouse[1]

Received November 27, 1985; final December 12, 1985

A new learning mechanism is proposed for networks of formal neurons analogous to Ising spin systems; it brings such models substantially closer to biological data in three respects: first, the learning procedure is applied initially to a network with random connections (which may be similar to a spin-glass system), instead of starting from a system void of any knowledge (as in the Hopfield model); second, the resultant couplings are not symmetrical; third, patterns can be stored without changing the sign of the coupling coefficients. It is shown that the storage capacity of such networks is similar to that of the Hopfield network, and that it is not significantly affected by the restriction of keeping the couplings' signs constant throughout the learning phase. Although this approach does not claim to model the central nervous system, it provides new insight on a frontier area between statistical physics, artificial intelligence, and neurobiology.

KEY WORDS: Neural networks; associative memory; biological memory; learning rules; spin glasses; storage capacity.

INTRODUCTION

During the past few years, a large number of investigations have endeavored to explain the behavior of large collections of neurons with the tools of statistical mechanics. The models proposed initially by Little[1] and by Hopfield[2] have been explored and their scope extended both from the point of view of their implications at nonzero temperatures[3,4] and of their learning ability at zero temperature.[5–7] These investigations have shown that networks of simple formal neurons might exhibit very interesting

[1] Laboratoire d'Électronique, École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, 10 rue Vauquelin, 75005 Paris, France.

0022-4715/86/0500-0411$05.00/0

properties in terms of information processing (storage and retrieval). However, from a biological standpoint, the learning rules (derived from Hebb's rule) which have been used in all the investigations of neural networks inspired by the Hopfield model are questionable in at least three respects[8]:

(i) The learning process starts from a zero synaptic matrix, a fact which is not substantiated by experimental evidence: the Hopfield model is an essentially "instructive" model in which the learning phase proceeds from an initial tabula rasa, whereas alternate theories advocate a "selectionist" point of view, in which learning starts from a network which already contains some "knowledge" (*prerepresentations*).[9,10]

(ii) The learning rule leads to a symmetrical synaptic matrix, which is a severe restriction and at best an approximation for real neural networks.

(iii) The synaptic changes undergone during the learning phase include the possibility of sign reversals for the synaptic strengths, which means that an excitatory synapse (with positive synaptic strength) might become an inhibitory synapse (with negative synaptic strength); such phenomena have not been observed in biological systems.

In the present paper, we show that it is possible to define a new, local, selectionist learning rule, which avoids the above-mentioned pitfalls, while guaranteeing the perfect memorization and retrieval of orthogonal prototype patterns of information (up to a maximal storage capacity).

Starting from an initial synaptic matrix with random elements (synaptic strengths), which produces a very large number of stable memorized states (called prerepresentations), we show that the effect of the learning procedure is a sequence of modifications of the initial synaptic matrix, which stores the prototype states as fixed points (attractors) of the dynamics, and retains those prerepresentations that are uncorrelated to the prototype patterns while gradually forgetting the others. We also show that, for weakly correlated prototype patterns, the storage capacity of networks obtained with this learning rule is similar to the storage capacity obtained with Hebb's rule; moreover, preventing sign reversals of the synaptic strengths does not significantly degrade this capacity.

Obviously, the present approach is not an attempt at modeling the whole complexity of the central nervous systems; it shows that the models of learning in networks of formal neurons which have been used in the recent past can be brought substantially closer to biological models without involving more complicated mathematical formalism.

LEARNING FROM PREREPRESENTATIONS

1. Presentation of the Network

The networks investigated in the present paper are identical to those studied in a previous paper;[5] we summarize briefly their characteristics and the notations used. The state of a neuron i is represented by a variable σ_i (spin), the numerical value of which can be either 1 if the neuron is active or -1 if the neuron is inactive. We consider a fully connected network of n such neurons operating synchronously in parallel with period τ, without sensory inputs. The strength of the synaptic junction of neuron i receiving information from neuron j is represented by a coupling coefficient C_{ij}. The state of a neuron i at time $t+\tau$ depends on the state of the network at time t in the following way: the neuron i computes its membrane potential

$$v_i(t) = \sum_{j=1}^{n} C_{ij}\sigma_j(t)$$

then it compares $v_i(t)$ to its threshold value θ_i and determines its next state $\sigma_i(t+\tau)$ according to the following decision rule

$$\begin{aligned} &\sigma_i(t+\tau) = \operatorname{sgn}[v_i(t) - \theta_i] && \text{if} \quad v_i(t) \neq \theta_i \\ &\sigma_i(t+\tau) = \sigma_i(t) && \text{if} \quad v_i(t) = \theta_i \end{aligned} \tag{1}$$

In this paper we take $\theta_i = 0$. The network operates at zero temperature.

2. A New Optimal Learning Rule

In Ref. 5, the general condition under which a given set of p prototype patterns $\{\boldsymbol{\sigma}^k\}$ are stable was established; it was shown that, if the activity thresholds are taken equal to zero, the simplest form of this condition can be written as

$$C\Sigma = \Sigma \tag{2}$$

where C is the (n, n) synaptic matrix and Σ is the (n, p) matrix whose columns are the prototype patterns to be memorized

$$\Sigma = [\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, \ldots, \boldsymbol{\sigma}^k, \ldots, \boldsymbol{\sigma}^p]$$

Therefore, the computation of the synaptic matrix reduces to the computation of a nontrivial solution of (2).

This equation always has a solution, the general form of which is

$$C = \Sigma\Sigma^I + B(I - \Sigma\Sigma^I) \tag{3}$$

where I is the identity matrix, B is an arbitrary (n, n) matrix and Σ^I is the Moore–Penrose pseudoinverse[12] of Σ. Matrix $\Sigma\Sigma^I$ is the orthogonal projection matrix into the subspace spanned by the prototype vectors. In general, C will not be a sparse matrix, so that the network will be fully connected. This rule is optimal in that it guarantees the perfect storage and retrieval of any given set of prototype patterns.

As long as the learning phase has not started, one has

$$C = B$$

In Ref. 5, it had been chosen for simplicity

$$B = 0$$

which corresponds to the empiricist point of view, in which the learning phase starts from a tabula rasa. In contrast to this initial approach, we consider, in the present paper, that the network is initially defined by a nonzero synaptic matrix B with random element values, which creates a rich set of attractors in phase space and determines fixed points and limit cycles of the dynamics, which are called *prerepresentations*. Thus, the learning process will consist in altering the preexisting phase space picture in order to accomodate the new items of information, instead of creating it ab initio.

If the initial matrix B is taken to be symmetrical, then the resulting phase space structure can be viewed as a spin glass energy landscape, and the attractors are the bottoms of the valleys (static prerepresentations); for reviews see Ref. 11. However, *even if matrix* B *is symmetrical, the synaptic matrix* C *after learning will not, in general, be symmetrical*, which is one of the conditions for a learning model to be plausible from a biological standpoint.

The scaling of the elements of matrix B with n can be determined by the following argument: the initial membrane potential v_i^* of neuron i when the network is in a state σ is given by

$$v_i^* = \sum_{r=1}^{n} B_{ir}\sigma_r$$

This potential, being a physically measurable quantity, should remain finite if n becomes very large; therefore, the elements of B should be $O(1/\sqrt{n})$.

3. A Local, Selectionist Learning Rule

In the thermodynamic limit (strictly speaking, if $n \to \infty$ and $p/n \to 0$), if the components of the prototype patterns are taken randomly, the latter

are uncorrelated, hence orthogonal. For such patterns, the pseudoinverse of matrix Σ reduces to

$$\Sigma^I = (1/n)\,\Sigma^T$$

where Σ^T is the transpose of Σ.

Therefore, relation (3) may be rewritten as

$$C = B + (1/n)(I - B)\,\Sigma\Sigma^T \tag{4}$$

It will be shown in the next section that this form of the optimal learning rule is local and selectionist in nature. It guarantees the perfect memorization and retrieval of orthogonal patterns, but is suboptimal for weakly correlated patterns. This rule will be used throughout the present paper.

ANALYSIS OF THE NEW LEARNING RULE

1. Iterative Form

Relation (4) expresses the fact that the synaptic matrix depends both on the initial configuration of the synapses (matrix B) and on the knowledge that has been acquired during the learning procedure (matrix Σ). If B is taken equal to zero (tabula rasa), it should be noticed that relation (4) reduces to Hebb's rule.

In general, learning is a sequential process: each time a new pattern is learned, the synaptic matrix undergoes a change, so that the initial configuration of the synapses fades out gradually. Therefore, in order to understand the learning process correctly, one has to investigate the iterative nature of the learning rule: we show in the following that rule (4) can be put in an iterative form which does not involve explicitly the initial synaptic matrix B.

We assume that a set of $k-1$ patterns has been learned and that one extra pattern $\boldsymbol{\sigma}^k$, orthogonal to the previous ones, is to be learned; learning this extra pattern will result in altering the synaptic matrix $C(k-1)$ (corresponding to the network having learned $k-1$ patterns) to give a new matrix $C(k)$

$$C(k) = C(k-1) + (1/n)(I - B)\,\boldsymbol{\sigma}^k \boldsymbol{\sigma}^{kT} \tag{5}$$

We show in Annex 1 that

$$C(k-1)\,\boldsymbol{\sigma}^k = B\boldsymbol{\sigma}^k \tag{6}$$

If the patterns stored previously are orthogonal, one can write the iterative form of the learning rule as

$$C(k) = C(k-1) + (1/n)[I - C(k-1)]\,\boldsymbol{\sigma}^k \boldsymbol{\sigma}^{kT} \tag{7}$$

with

$$C(0) = B$$

These relations show that the learning process superimposes the patterns learned sequentially to the initial "knowledge" due to matrix B.

2. Memory and Selection

The initial matrix B does not appear explicitly in relation (7). However, the network does keep a "memory" of the initial configuration of the synapses, in the following way: it is clear from relation (6) that if the new item of information $\boldsymbol{\sigma}^k$ was a prerepresentation, it is still memorized after learning the $k-1$ previous patterns. Thus, we have the following result: *the learning procedure does not erase any of the prerepresentations which are uncorrelated to the stored patterns.* Moreover, we show in Appendix 2 that the prerepresentations which are correlated to the learned patterns are gradually erased. Thus, *the learning phase consists in altering the initial matrix so as*:

(i) *to memorize the prototype patterns*

(ii) *to select the prerepresentations uncorrelated to the informations learned during the learning phase*

(iii) *to erase the prerepresentations correlated to the prototype patterns*

Thus, the new learning rule is in fairly good agreement with the selectionist point of view.

Let us consider the particular case in which the new pattern to be learned is a prerepresentation

$$B\boldsymbol{\sigma}^k = C(k-1)\,\boldsymbol{\sigma}^k = D\boldsymbol{\sigma}^k$$

where D is a positive diagonal matrix

$$D_{ii} = |v_i^*|$$

The variation of the synaptic strength can thus be written under the simple, Hebb-like form

$$\Delta C_{ij} = (1/n)\,\sigma_i^k \sigma_j^k (1 - |v_i^*|)$$

Specifically, if $|v_i^*| = 1$ for all i, the synaptic matrix remains unchanged: the pattern is learned without any effort. If the initial matrix is of the spin-glass type, this case will occur with negligible probability, however. On the average, since v_i^* is a random variable with standard deviation 1, the expectation value of $|v_i^*|$ will be $\sqrt{2/\pi} \approx 0.8$, so that the increment of C_{ij}, for learning a prerepresentation, will be approximately $0.2/n$ (instead of $1/n$ with Hebb's rule).

3. Interpretation of the Terms of the Learning Rule

In order to get more insight into the learning process, we now consider the incremental variation of the coefficient of a given synapse C_{ij} when the network attempts to memorize a pattern $\boldsymbol{\sigma}^k$. Relation (5) leads to

$$\Delta C_{ij}(k) = (1/n)\, \sigma_i^k \sigma_j^k - (1/n)\, \sigma_j^k \sum_{r=1}^{n} B_{ir} \sigma_r^k \tag{8}$$

The first term corresponds to the classical Hebb's rule, giving a contribution of $\pm(1/n)$; if the initial synaptic matrix B is random with values of $\pm 1/\sqrt{n}$, the second term is $O(1/n)$. Obviously, this term is not symmetrical with respect to i and j, in general.

An alternate form of the incremental variation of the synaptic coefficient C_{ij} necessary to learn a new pattern $\boldsymbol{\sigma}^k$ may be derived from relation (7)

$$\Delta C_{ij}(k) = (1/n)\, \sigma_i^k \sigma_j^k - (1/n)\, \sigma_j^k \sum_{r=1}^{n} C_{ir}(k-1)\, \sigma_r^k$$

where $C_{ir}(k-1)$ is the value of the synaptic coefficient C_{ir} prior to learning the new pattern $\boldsymbol{\sigma}^k$. As was mentioned previously, the first term, corresponding to the classical Hebb's rule, expresses the local interaction between neurons i and j at the level of their connecting synapse: the variation of the synaptic strength depends only on the states of the neurons connected by that synapse. In the second term, the summation is the membrane potential v_i of neuron i when the new pattern to be learned, $\boldsymbol{\sigma}^k$, is input to the network $C(k-1)$; therefore, the variation of the synaptic coefficient depends on the state of the afferent neuron j and on the membrane potential of neuron i; it takes into account the influence of the other synapses afferent to neuron i. It can be noticed that v_i, being a graded variable, provides a fine tuning of the variations of the synaptic coefficient, whereas the classical Hebb's rule allows these coefficients to vary only by steps of $1/n$. Thus, the selectionist learning rule is local; in contrast, if the general form (3) is used to compute the variation of a given synaptic coefficient, it takes into account all the synaptic coefficients of the network.

4. Learning Without Sign Reversal of the Synaptic Coefficients

As was mentioned above, an important condition for a model to make sense from a biological standpoint is the absence of sign reversal in the synaptic coefficients during learning. We are now in a position to evaluate the number of states that can be stored without sign reversal of the synaptic coefficients in the following way: the value of a synaptic coefficient C_{ij} after learning p patterns is given, after relation (8), by

$$C_{ij}(p) = B_{ij} - (p/n)\, B_{ij} + (1/n) \sum_{k=1}^{p} \sigma_i^k \sigma_j^k - (1/n) \sum_{k=1}^{p} \sum_{r \neq j} B_{ir} \sigma_r^k \sigma_j^k \quad (9)$$

Assume that the elements of matrix B are equal to $\pm 1/\sqrt{n}$, taken randomly with probability $\frac{1}{2}$, and that the patterns are chosen randomly. The second term arises from the contribution of $r = j$ to the second term of relation (8). The third term can be interpreted as the result of a random walk of p steps of length $1/n$; therefore, for sufficiently large p, it is the realization of a centered Gaussian random variable of standard deviation $\sqrt{p}/n$. Similarly, the fourth term can be interpreted as the result of a random walk of $p(n-1)$ steps of length $1/n\sqrt{n}$; therefore, it has a standard deviation of $\sqrt{p(n-1)}/n\sqrt{n} \approx \sqrt{p}/n$. These random variables being independent, their sum ζ is centered Gaussian with standard deviation $s \approx \sqrt{2p}/n$. If $B_{ij} = -1/\sqrt{n}$, the probability that the synaptic coefficient $C_{ij}(p)$ undergoes a sign reversal is the probability of having $C_{ij}(p) > 0$

$$\text{Prob}[\zeta > s_0] = 1/(s\sqrt{2\pi}) \int_{s_0}^{+\infty} \exp(-t^2/2s^2)\, dt$$

where $s_0 = (n-p)/n\sqrt{n}$.

As usual, this probability can be expressed in terms of normalized quantities ζ/s and $S = s_0/s$

$$\text{Prob}(\zeta > s_0) = \text{Prob}(\zeta/s > S)$$

Therefore, the probability of sign reversal is governed by the normalized variable

$$S_1 \approx (n-p)/\sqrt{2np} = (1-\alpha)/\sqrt{2\alpha}$$

where $\alpha = p/n$.

The same result holds if $B_{ij} = +1/\sqrt{n}$. Therefore, for a given probability of a synapse undergoing a sign reversal, the number of patterns which can be stored is $O(n)$. For instance, before reaching Prob(sign reversal) = 0.05 one can store $p \approx n/7$ uncorrelated patterns.

COMPARISON WITH TABULA RASA LEARNING RULES

In the Hopfield (or Little) models, the learning rule used to compute the coupling coefficients was the usual Hebb's rule, which is a particular case of our learning rule (with $B=0$). In the present section, we compare these two approaches.

The new learning rule has one common feature with Hebb's rule: it is optimal for storing orthogonal patterns. Therefore, if these rules are used with nonorthogonal patterns (such as, for instance, random patterns with a finite number of neurons), the stability of the prototype patterns is no longer guaranteed; it is well-known that this problem is a strong limitation to the storage capacity of Hopfield networks. Similarly, the question which arises in our case is: what is the behavior of the new learning rule when one attempts to memorize nonorthogonal patterns? Moreover, we have shown that, with some restriction on the storage capacity, the new learning rule enables the network to store patterns without causing sign changes in the synaptic strengths. In this context, two questions arise: first, is this restriction more or less stringent than the restriction due to the fact that the weakly correlated prototype patterns must be stable; second, how does our rule compare with the Hopfield model as far as sign changes in the coupling coefficients are concerned?

Thus, a comparison between these rules must consider two problems separately:

(i) the ability of storing random prototype patterns with a finite number of neurons

(ii) the ability of storing such patterns without sign reversal of the synaptic coefficients

Let us consider the first problem, without any restriction concerning the sign reversals of the synapses; the stability condition of a component σ_i^m of a prototype pattern is

$$\sum_j C_{ij}\sigma_j^m\sigma_i^m > 0$$

where C_{ij} is given by relation (9).

Following the lines of derivation of the sign reversal probability for a synapse, it can be shown, after some algebra, that the probability for a bit of a prototype state to be stable is governed by the dimensionless quantity

$$S_2 \approx (1+\alpha)/\sqrt{2\alpha}$$

In the Hopfield model (*tabula rasa* hypothesis), one has $B=0$, so that the above stability condition reduces to

$$n+p-1+\sum_{j\neq i}\sum_{k\neq m}\sigma_i^k\sigma_j^k\sigma_j^m\sigma_i^m>0$$

A similar derivation shows that, in this case, the relevant variable for studying the stability is

$$S_3\approx(1+\alpha)/\sqrt{\alpha}$$

Since S_2 and S_3 have the same order of magnitude, *the storage capacity of weakly correlated patterns with or without tabula rasa are similar.*

We consider now the problem of the sign reversals of the synaptic strengths. Obviously, the *tabula rasa* learning rule implies very frequent sign reversals, since the initial value of the synaptic strengths is zero and since the latter are incremented by $\pm 1/n$ each time a new vector is learned; in sharp contrast, we have shown in a previous section that the probability of a change in the sign of a synaptic strength with the new learning rule is governed by

$$S_1=(1-\alpha)/\sqrt{2\alpha}$$

Since $S_1\approx S_2$, *the constraint of having no sign reversal in the synaptic strengths does not impair the storage capacity of a network without tabula rasa, whereas it has a dramatic effect on the storage capacity of networks with tabula rasa.*

A final point should be mentioned: the fact that the storage capacity is limited by the nonorthogonality of the prototype patterns is due to our use of a particular form of relation (3), in which the pseudoinverse Σ^I was replaced by the simpler form $(1/n)\,\Sigma^T$. If the general form of relation (3) is used, the stability of any set of prototype patterns is guaranteed, and the iterative nature of the learning rule is preserved; however, as was previously mentioned, the rule is no longer local, which is questionable from a biological standpoint.

CONCLUSION

Starting from the general stability condition of a pattern in a neural network, we have derived a new learning rule, interpreted in terms of local interactions, which embodies three features that are essential for a biologically realistic model of a neural network. Specifically, we have shown that, starting from an arbitrary initial synaptic matrix, it is possible

to store and retrieve faithfully uncorrelated patterns of information, and that, provided their number is not too large, the synaptic coefficients have a low probability of undergoing a sign reversal. It has been further proved that, even if the initial synaptic matrix is symmetrical, the synaptic matrix after learning is not, in general, symmetrical. It has been established that the learning procedure is selective: it does not erase the prerepresentations which are not correlated to the knowledge stored during learning, but it does forget the prerepresentations which are correlated to the stored information. Finally, the storage capacity due to the new learning rule has been shown to be similar to that of *tabula rasa* learning rules.

APPENDIX 1

We assume that the network has learned $k-1$ patterns of information since the beginning of the learning phase, and we denote by $C(k-1)$ the synaptic matrix at this step of the learning phase.

We define a matrix:

$$\Sigma(k) = [\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, \ldots, \boldsymbol{\sigma}^k]$$

After relation (4) we can write

$$C(k-1)\,\boldsymbol{\sigma}^k = B\boldsymbol{\sigma}^k + (1/n)(I-B)\,\Sigma(k-1)\,\Sigma^T(k-1)\,\boldsymbol{\sigma}^k$$

Since the new item of information is orthogonal to the previous ones, we have

$$\Sigma^T(k-1)\,\boldsymbol{\sigma}^k = 0$$

Therefore

$$C(k-1)\,\boldsymbol{\sigma}^k = B\boldsymbol{\sigma}^k$$

which shows that, if $\boldsymbol{\sigma}^k$ was a prerepresentation, it is still memorized after $k-1$ steps of the learning phase.

APPENDIX 2

We show that if a vector $\boldsymbol{\sigma}$ is a prerepresentation, and if it is correlated to patterns stored during the learning phase, it is no longer a stable state with the new synaptic matrix.

Since $\boldsymbol{\sigma}$ is a prerepresentation, one has

$$B\boldsymbol{\sigma} = D\boldsymbol{\sigma}$$

where D is a positive diagonal matrix.

We define a diagonal matrix D' by

$$(1/n)(I-B)\,\Sigma\Sigma^T\boldsymbol{\sigma} = D'\boldsymbol{\sigma}$$

After relation (4), we have

$$C\boldsymbol{\sigma} = (D+D')\boldsymbol{\sigma}$$

Obviously, there is no reason why matrix D' should be positive diagonal. Since the elements of D' and D have the same order of magnitude, matrix $D+D'$ will not be, in general, a positive diagonal matrix, so that vector $\boldsymbol{\sigma}$ will not be stable after learning k patterns.

REFERENCES

1. W. A. Little, *Math. Biosci.* **19**:101 (1974).
2. J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **79**:2554 (1982).
3. D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. A* **32**:1007 (1985); *Phys. Rev. Lett.* **55**:1530 (1985).
4. P. Peretto, *Biol. Cybern.* **50**:51 (1984).
5. L. Personnaz, I. Guyon, and G. Dreyfus, *J. Phys. Lett.* **46**:L359 (1985).
6. W. Kinzel, *Z. Phys. B* **60**:205 (1985).
7. N. Parga and M. A. Virasoro, to be published.
8. G. Toulouse, S. Dehaene, and J. P. Changeux, *Proc. Natl. Acad. Sci. U.S.A.*, to be published.
9. N. Jerne, in *The Neurosciences: A Study Program*, G. Quarton *et al.*, eds. (The Rockefeller University Press, New York, 1967).
10. J. P. Changeux, T. Heidmann, and P. Patte, in *The Biology of Learning*, P. Marler and H. Terrace, eds. (Springer-Verlag, New York, 1984).
11. K. H. Fischer, *Phys. Stat. Sol. (B)* **116**:357 (1983); see also: "Heidelberg Colloquium on Spin Glasses," *Lecture Notes in Physics*, No. 192 (Springer-Verlag, New York, 1983).
12. A. Albert, in *Regression and the Moore–Penrose Pseudoinverse* (Academic Press, New York, 1972).