PREPRINT N.890
2 Giugno 1992

Dipartimento di Fisica
Università di Roma "La Sapienza"
I.N.F.N.- Sezione di Roma

24

Fermilab Library
0 1160 0013652 7

Constraints on Learning in Dynamic Synapses

Daniel J. Amit[1] and Stefano Fusi
INFN, Sezione di Roma, Istituto di Fisica
Università di Roma, La Sapienza, P.le Aldo Moro, Roma

Abstract

Hebbian type learning is discussed in a network whose synapses are analog, dynamic variables, whose values have to be periodically refreshed due to possible exponential decay, or other instability of continuous synaptic efficacies. It is shown that the end product of learning in such networks is very sensitive to the relation between the rate of presentation of patterns and the size of the refresh time interval. It is shown that in the limit of slow presentation, the network can learn at most $O(\ln N)$ patterns in N neurons, and must learn each one in one shot, thus learning all errors present in a corrupt stimulus presented for retrieval.

It is then shown that as the rate of presentation is increased the performance is increased rapidly. Another option we investigate is that in which the refresh mechanism is acting stochastically. In this case the rate of learning can be slowed down very significantly, but the number of stored patterns cannot surpass $\sqrt{N}$.

1992

1 Introduction

Modeling brain function is largely concerned with exposing constraints. This holds for the study of brain, as well as for technological implementations. Constraints may include storage capacities, for a memory; retrieval quality at various storage levels; operation times, in memory retrieval or learning; tolerance to correlations, etc etc. The origin of constraints may be found in information theory; in statistical inference; in statistical mechanics, or in dynamics. Often the context in which the constraints are initially found seems over-simplified. Experience has been that as often as not the constraints captured general features of a much wider class of computational systems. Moreover, the corresponding approaches and techniques have turned out to be effective in much less simplified situations.

Here we would like to open a discussion of constraints intrinsic to unsupervised learning in attractor neural networks (ANN). We have come across these constraints in connection with the design of an organically learning ANN, implemented in silicon[1]. Attractors are learned in associative cortex [2]. Moreover, since attractors are learned in cortical regions that are far removed from the input and output regions, it is likely that they are learned in an unsupervised fashion.

[1]On leave of absence from Racah Institute of Physics

The general Hebbian framework provides a simple mechanism for affecting such learning of attractors[3]. In other words, if an afferent activates selectively certain sets of neurons in an 'assembly', then modifying each synapse by some average correlation of the two neurons connected by it, is a candidate for the creation of a corresponding attractor. One of the attractions of the Hopfield model[4] was the fact that the synaptic matrix could be viewed as a linear accumulation of the patterns to be memorized, producing those same patterns as the network's attractors.

If synapses had the analog depth required by the Hopfield matrix, one could envisage a network learning that matrix by acquiring the patterns one at a time. Moreover, the synapses would be required to have a temporal stability for the multiplicity of values arrived at in the course of this one shot learning, both for the intermediate values as for the final ones. The problem of the analog depth of the synaptic values was treated by Sompolinsky[5]. He has evaluated the performance of a Hopfield network, following a clipping of all the synaptic values to one of three values. Given a synaptic threshold, each synapse whose absolute value is above the threshold is given a prescribed uniform value, preserving only its sign. A synapse whose value is below the threshold, is set to zero. In this case one ends up with a diluted network: a fraction of the synapses, that depends on the threshold are eliminated. The performance of the network is then calculated and can be compared with performance of the intact network.

This would have been a solution to the analog depth problem, provided one could affect learning by first accumulating linearly, and stably, the Hopfield synaptic values. Then, once a desired set of patterns has been learned, their number may be arbitrarily large, clipping intervenes to produce a network that can store that set of patterns and yet have only three synaptic efficacies. The practical problem, electronic or biological, is more difficult. It may be the case that the synapses do not have the analog depth to maintain the intermediate values at any time. Even a fully analog synapse, as would be a condenser, may not be able to stabilize its acquired values (See also e.g. [10, 11]. What is common to both cases is that the synapses must be clipped, or refreshed, all along the learning process.

The Sompolinsky clipping can be assumed to take place in some soft form, with its own time constant, which may be the RC of a condenser. In other words, one could assume that a synapse has an electronic, or biochemical mechanism which pushes the synaptic value towards some saturation, whenever that value is above some threshold, and allows the value of the synapse to decay to zero if it is below the threshold. In this case, new questions present themselves. The properties of the resulting network strongly depend on the dynamics of the learning process and not merely on the informational content of the data and on its statistical distribution. In fact, the rate of arrival of the data to be learned, as well as the time of presentation of each pattern, relative to the time constant of the synaptic clipping (refresh) are found to be crucial. We shall see that consideration of some simple relations between these time constants leads to performances much below

those of the static clipped network. In this case the results of information theory,as well as those of statistical mechanics, give a grossly innacurate upper bound. On the other hand, the simple arguments introduced here, show several ways in which things can be improved. Those can be introduced either in the dynamics of data presentation, or in the dynamics of the clipping.

One immediate consequence of such learning is that more recently presented patterns are privileged relative to older ones, much as in 'palimpsest' scenarios[6, 7]. But palimpsest effects are due to the presence of an upper cut-off on the synaptic values, of alternatively, or the exponential decay of the contributions of older patterns[12]. As we shall see below, the refresh mechanism is significantly more restrictive. It does roll memories from the present into a disappearance in the past. But the refresh mechanisms to be discussed here do the rolling faster and consequently the storage capacities are significantly reduced.

The basic learning scenario is one in which random patterns are presented at low rate, with at most one pattern per refresh period. The resulting network is found to be very limited: patterns have to be learned in a single shot, and no more than $\ln N$ attractors can be stored in a network of N neurons. Things improve significantly when the rate of presentation is increased. This is expressed by a scenario in which the patterns are grouped into subgroups, each presented within one refresh period. In this case the capacity is increased very fast as the number of patterns per refresh period increases. Finally we consider a stochastic refresh mechanism, in which synapses which should be modified due to the stimulus being presented, are modified with a probability lower than unity. In this case, learning can be made much slower, and a single presentation of a corrupt pattern, for retrieval, has essentially no effect on the synaptic structure. Provided the probability is low enough. Moreover, in this case the capacity of the network can become as high as $\sqrt{N}$, but not higher.

In section 2 we define the network and the three learning models. In section 3 we demonstrate that if the patterns are presented at low rate, i.e. with large intervals in between, then the network can learn at most $O(\ln N)$ patterns. Then, in section 4, it is shown, by a signal to noise analysis of retrieval, that the capacity of the network can be expressed in terms of a single parameter: the fraction of the synapses which retain a dependence on every pattern in the sequence of learned patterns. Using this result, in Section 5 the storage capacities for all learning scenarios are analyzed. The expressions for the relevant fractions are derived in an appendix. The results are supported by simulations.

2 The model

The elements

In what follows we shall concentrate exclusively on the dynamics of the synapses, so in principle there is no need to specify the neural variables. But, in order to bring the discussion closer to the familiar retrieval models, we describe both sets of variables. Thus, we choose a set on N discrete neurons, each taking values ± 1. The synapses are also discrete, and can take three values: $J_{ij} = +J_c, 0, -J_c$. The number of neurons will be considered very large. And any two neurons can have a potential synapse.

Comments:

1. The choice of the neuronal variables serves only to indicate that in every learning step, of a Hopfield type learning, all synapses receive a contribution which is either +1 or −1, depending on the state of the neurons in the network. Thus, if the network is in state ξ_i^λ, at that time the synapse ij will receive the contribution

$$\alpha_{ij}^\lambda \equiv \xi_i^\lambda \xi_j^\lambda. \tag{1}$$

2. In comment 1, the states of the neurons can be those imposed by a strong stimulus, or those maintained in an attractor. In neither case can the variables take other values, i.e. the internal representation employs all the neurons of the network. Thus even a stimulus with a partial content of information affects all neurons and consequently all synapses. This is a severe limitation of the ± 1 formulation. One way of avoiding the constraints to be discussed here is to allow for an updating of a partial set of synapses. This may imply a different formulation of the neural variables, or some other stratagem, as will be discussed below. But the simple formulation is very effective in exposing the problem.

3. The actual values that can be attained by the synapses are not intended to be just the three values mentioned above. We have in mind a situation in which the synapses can take a continuous set of values. But if synaptic values decay with time, they have to be *refreshed*, e.g.: increasing the values of the positive ones above threshold, decreasing the negative one which are below a negative threshold, and allowing the intermediate efficacies to decay to zero. Similarly, the internal synaptic dynamics may already include the refresh. In both cases, if there is a long interval between two different configurations of the network, the synapses reach a situation of three-state clipping, as defined above, with J_c the saturation value of the refresh mechanism.

Thus, in what follows we will assume just this simple situation in which patterns are learned at low frequency, in a Hopfield fashion. They are clipped or refreshed following every single pattern. The question will be how many patterns can be learned this way.

Finally, we shall assume that the patterns to be learned by the network are N-bit words of independently chosen ± 1s, with equal probability.

The learning situation – slow presentation

Observing the period between two refresh operations, we denote the beginning by λ and the end by $\lambda+1$. We assume that at λ the distribution of the synapses has the following features:

- The synapses have a 3-way clipped distribution ($J_{ij} = +J_c, 0, -J_c$);
- The positive and negative values appear with equal probability.

This will be the case if the initial matrix were zero. As will be shown in section 8.3, it is also the situation for an arbitrary initial condition, following several refresh steps. The fact that the synaptic distribution converges rapidly onto this general form, allows us to treat a sequence of p consecutive patterns, placed anywhere in the total stream of input, as if they were the first p patterns presented.

Since in the learning step all synapses receive contributions which are equal in absolute values, there will be four possible types of learning steps, depending on the choice of J_c and J_0:

1. Stimuli do not affect the synaptic matrix at all:

$$J_{ij}(\lambda) = J_{ij}(\lambda+1) \quad \forall i,j. \tag{2}$$

   Note that this is the case in which no positive synapse is brought below threshold and no zero synapse is brought above the threshold. The same is true for the negative values, due to the symmetry.

2. Stimuli are learned, erasing all previous memory in the system. The synaptic change is the following:

$$\begin{cases} J_{ij}(\lambda) = +J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ J_{ij}(\lambda) = \quad 0 & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ J_{ij}(\lambda) = -J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \end{cases} \tag{3}$$

   Note that the first two lines imply the last two by the sign symmetry. Moreover, the second line implies the forth, as the fifth implies the third. The resulting set of synapses is independent of the original set and stores the last pattern.

3. If the threshold for refresh is high, the zero synapses, in the previous case may remain unchanged. The outcome then is:

$$\begin{cases} J_{ij}(\lambda) = +J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ J_{ij}(\lambda) = \quad 0 & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ J_{ij}(\lambda) = -J_c & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \end{cases} \tag{4}$$

In this case the last pattern merely cuts the synapses that were learned previously and are inconsistent with it. If the patterns presented are uncorrelated, the number of non-zero synapses is reduced by a factor of two with every new presentation. Following the presentation of several patterns, the synaptic matrix will become all zeros.

4. Finally, there is the relevant case in which the stimulus changes synapses, but retains their previous history. This case is represented by the following learning step:

$$\begin{cases} J_{ij}(\lambda) = +J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ J_{ij}(\lambda) = \quad 0 & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ J_{ij}(\lambda) = -J_c & \rightarrow J_{ij}(\lambda+1) = \quad 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \end{cases} \tag{5}$$

Essentially, in this case, the new pattern is learned by the zero synapses; retains the convenient non-zero ones and zeros the contradicting non-zero synapses.

The learning situation – fast presentation in groups

As we shall see below, the slow learning – at most one pattern per refresh cycle – leads to a very constrained situation. First, because every pattern presented must make a finite change in the synaptic matrix, or be ignored. In other words, slow presentation requires rapid learning. Second, only a very small number of patterns can be learned in attractors. The first escape from the very stringent constraints discussed above is to speed up the presentation of the patterns, so that out of the totality of p patterns, n patterns appear in a single refresh period. We shall still keep the simplification that during the period in which the subgroup of

n patterns are presented, the temporal decay of the synaptic values is neglected. Typically, there will be an exponential decay within the subgroup, producing the result that the contribution of earlier patterns to a synapse have an exponentially weaker amplitude the further they go into the past. This effect, in a complete group of patterns, was studied by Mezard et al [12], and its effect in the present context will be discussed later, in Section [12].

In the present case, in every period, the contribution to the synaptic value is:

$$A = \sum_{\mu=1}^{n} \alpha_{ij}^{\mu}. \tag{6}$$

Given values for J_0 and J_c, there will be a threshold value, A_0, for A such that upon clipping, the synapses can learn, e.g.,

$$\begin{cases} J_{ij}(\lambda) = +J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } A > -A_0 \\ & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } A < -A_0 \\ \\ J_{ij}(\lambda) = \quad 0 & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } A > A_0 \\ & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } |A| < A_0 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } A < -A_0 \\ \\ J_{ij}(\lambda) = -J_c & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } A > A_0 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } A < A_0 \end{cases} \tag{7}$$

See the discussion in Section 2. There, the value of n was 1 and the value of A_0 was lower than 1. In practice A_0 should satisfy:

$$A_0 = \max(J_0, J_c - J_0).$$

The consequences of this learning scenario are discussed in Section 5 as well as in the Appendix 8.1.

The learning situation – stochastic learning

Another way of avoiding the strict constraints controlling slow presentation of patterns, is to affect the synaptic changes, implied by every presentation, on a fraction of the neurons only. This can be viewed as the following type of mechanism: The synaptic dynamics, during the presentation of a pattern, drives the synaptic value towards a saturation value determined by the strength of the stimulus. If this value is lower than the threshold for the activation of the refresh, no pattern will be learned. But, if the value at saturation is not far from the threshold for refresh, noise in the refresh mechanism then drives the synaptic value across the threshold, with a given probability q, determined by the noise. In this manner learning can be significantly slowed down and the capacity of the network significantly increased, but only from $\ln N$ to $\sqrt{N}$. This fact has also been observed by Horner in a different

context[9]. The consequences of this learning scenario will be discussed in Secion 5 and in Appendix 8.2.

3 The basic constraint

Clearly, of the four learning situations listed in the previous section only the forth is of interest. This is a statement about the choice of the parameters J_c and J_0, given that the contribution of the pattern being learned has been normalized to unity. The dynamics ensures that the presentation of a single pattern modifies the synaptic structure, keeping a good memory of the previous structure. But in trying to memorize many patterns, one must confront the question: How long is the memory of the learning process. In other words, given a sequence of p patterns, each modifying the synaptic matrix as described, can the network remember its structure preceding the entire sequence.

In the course of learning a flow of bits arrives at each synapse: the synapse ij gets a flow of bits α_{ij}^{μ}, defined by 1, where μ enumerates the pattern number in the sequence of presentation. If a sequence of p patterns is presented, any synapse will see one of 2^p different possible sequences of bits, which it will use one-by-one in the dynamics described in case 3. The origin of the constraint we shall demonstrate is in the fact that the special synaptic dynamics, implied by clipping following every presentation, allows only a fixed **number** of the 2^p p-bit sequences to preserve a memory of the synaptic value preceding the last p presentations. This implies that the probability of obtaining a useful sequence of bits arriving at a given synapse is proportional to 2^{-p}. Since the sequences are independently and uniformly distributed, 2^{-p} will also be the fraction of the synapses which will learn from the p patterns and not loose the information accumulated prior to this last sequence. Thus, one arrives at the conclusion that one can learn patterns at most as long as $2^p \approx N^2$, or $p \approx \ln N$.

This result follows from the fact that:

- given any synapse, the only strings of p bits which will produce a final value for the synaptic efficacy that depends on the value of the first bit in the string, are those which have no consecutive pair of equal bits following the first one.

For any p, there are only 4 such strings, depending on the signs of the first two bits in the sequence.

If there are two consecutive equal bits in a sequence, for example, if in the sequence $\alpha_p = \{\alpha^p, \alpha^{p-1}, ...\alpha^1\}$ one has $\alpha^i = \alpha^{i+1}$ $(i = 2, ..., p-1)$, then $J(p)$ is independent of $J(0)$.[2] The subscripts ij are implicit. For $\alpha^i = \alpha^{i+1} = +1$, all three synaptic values we become:

[2]Note that here pattern number 0 can be considered to be the begining of a sequence which can start anywhere in the total stream

$$J(i-1) = +J_c \rightarrow J(i) = +J_c \rightarrow J(i+1) = +J_c$$
$$J(i-1) = \ \ 0 \ \rightarrow J(i) = +J_c \rightarrow J(i+1) = +J_c$$
$$J(i-1) = -J_c \rightarrow J(i) = \ \ 0 \ \rightarrow J(i+1) = +J_c.$$

While for $\alpha^i = \alpha^{i+1} = -1$, the resulting value of $J(i+1)$ is $-J_c$, for all three cases. As a consequence, the probability that a synapse receives a sequence which does not erase previous memory is 2^{-p+2}.

This proves our assertion that presenting one pattern in any refresh period one can learn at most p_M patterns, with $2^{p_M-2} \approx N^2$.

The above argument does not prove that the stored patterns can be properly retrieved. This question we study next.

4 General features of signal-to-noise analysis

Given a synaptic matrix, following the learning of a set of p patterns, the question of the possible retrieval of these patterns is composed of two parts: First, do the synaptic values depend on the values of the different patterns in the sequence. And if they do, is a pattern recovered when presented to the network, i.e. is the signal appearing on neurons which should have a given value in that pattern large enough to reproduce the pattern, even in the presence of interference due to the memorization of the other patterns.

In the case of slow learning, in presence of refresh, we have found in the previous section that even the first requirement, the weaker one, allowed only $O(\ln N)$ patterns. Next one should calculate the signal-to-noise ratio, to check that retrieval is indeed possible for all these patterns. Since we shall next proceed to learning modes which induce much weaker restrictions, we first develop a general scheme for computing the number of patterns that can be memorized in one of the following three senses:

- That the synaptic matrix depends on all the patterns;
- That the signal-to-noise ratio in a pattern state is much higher than 1;
- That the probability of an error on any neuron tends to zero with increasing N, in the sense of Weisbuch and Fogelman-Soulie[8].

To derive the general results, we use the fact that:

- In the subspace of synapses which depend on the value of the oldest learned pattern, the conditional probabilities for a given synaptic value given a value of the input α from the oldest pattern $\Pr(J(p)|\alpha^1)$, is independent of p and of N.

In the appendix we show that this holds for all cases under consideration. Denote by d the fraction of synapses, out of the totality of N^2 synapses, which depend on α^1. The first condition for memorization is simply:

$$d \cdot N^2 > 1.$$

Next we argue that the signal, defined as:

$$S = \langle\langle \xi_i^1 h_i \rangle\rangle = \langle\langle \frac{1}{N} \sum_{j \neq i} J_{ij} \alpha_{ij}^1 \rangle\rangle \tag{8}$$

is proportional to d, while the square of the noise,

$$R^2 \equiv \langle\langle (\xi_i^1 h_i - S)^2 \rangle\rangle \tag{9}$$

is proportional to $1/N$.

Before proceeding to the demonstration of the above assertions on the signal and the noise, we draw the conclusions:

1. Naive signal-to-noise analysis has it that retrieval is possible while

$$\frac{S^2}{R^2} = CN \cdot d^2 \gg 1, \tag{10}$$

   where C is a constant independent of N and p.

2. The condition for vanishing probability of errors in retrieval, *a la* ref. [8], is

$$\frac{S^2}{R^2} = CN \cdot d^2 \gg \ln N. \tag{11}$$

We now demonstrate the assertions for the dependence of the signal and the noise on N and d, and then, using the expression for d in three different learning scenarios, we shall exhibit the behavior of the capacity of the resulting networks.

Note first that the signal can be written as:

$$S = \sum_{J=\pm J_c} \sum_{\alpha=\pm 1} \Pr(\alpha) \Pr(J|\alpha) J\alpha, \tag{12}$$

where $\Pr(J|\alpha)$ is the conditional probability of finding the value $J^p = J$, given that $\alpha^1 = \alpha$. Note that the sum is only over non-zero values of J. For random patterns, $\Pr(\alpha) = \frac{1}{2}$. Moreover,

$$\Pr(J|\alpha) = \Pr(-J|-\alpha)$$

and

$$\Pr(J|-\alpha) = \Pr(-J|\alpha).$$

The signal is therefore,

$$S = J_c\left[\Pr(J_c|1) - \Pr(J_c|-1)\right]. \tag{13}$$

If we denote by $\overline{\Pr}(J|\alpha)$ the conditional probability in the subspace of α^1-dependent J's, then

$$P(J|\alpha) = d \cdot \overline{\Pr}(J|\alpha), \tag{14}$$

and hence Eq. 13 can be written as:

$$S = J_c d\left[\overline{\Pr}(J_c|1) - \overline{\Pr}(J_c|-1)\right].$$

Since the $\overline{Pr}$ have been assumed independent of p and N, and the signal must be positive, on the right hand side of the last expression we have d multiplied by a constant independent of p and N, as asserted.

The mean square noise can be written as

$$\begin{aligned} R^2 &= \langle\langle \frac{1}{N^2}\sum_{jk} J_{ij}J_{ik}\alpha_{ij}\alpha_{ik}\rangle\rangle - S^2 = \\ &= \langle\langle \frac{1}{N^2}\sum_{j} J_{ij}^2\rangle\rangle \leq \frac{J_c^2}{N}. \end{aligned} \tag{15}$$

This is just the form we have asserted for the noise.

5 Results for three learning scenarios

The previous section derived a limit on the number of patterns that can be learned in a learning process with refresh, that arrives to tri-modal distribution of synapses, $\pm J_c, 0$. In section 3 we have seen that for slow learning, of one pattern per refresh period, even the most liberal constraint gave very low storage, e.g. p of order $\ln N$. Using the results of Section 4 we now proceed to derive the storage capacity in three different learning scenarios: the slow learning of Section 3; fast learning, in which the p patterns are presented in groups of n patterns in each refresh period; and finally, probabilistic slow learning, in which the threshold operation of the refresh is noisy and induces learning changes in every refresh cycle on a fraction of the synapses only. For each of the three learning scenarios we shall exhibit the results for the three criteria discussed in the previous section. Since the performance of the network depends only on the value of d, the fraction of synapses whose value depends on the oldest pattern, d will be given in this section for each of the three cases, then the consequences will be drawn. The derivation of d will be left to the appendix.

Capacity for slow learning

In this case, at most one pattern is presented per refresh cycle. The value of d, given in terms of p, is (see e.g. 8)

$$d = \frac{C}{2^p}$$

Using the general results of Section 4, we find:

- For at least one synapse to depend on the value of α^1_{ij}:

$$N^2 d \simeq C \quad \Rightarrow \quad p < 2\log_2 N$$

- For naive signal-to-noise stability,

$$N d^2 \simeq C \quad \Rightarrow \quad p < \frac{1}{2}\log_2 N$$

- For vanishing error probability:

$$N d^2 \simeq C \cdot \ln N \quad \Rightarrow \quad p < \frac{1}{2}\log_2\left(\frac{N}{\ln N}\right). \tag{16}$$

Note that the first constraint is already so tight, that the stricter requirement of signal-to-noise reduces the limit only by a constant factor of 4.

Fast learning in groups

Here the p patterns are presented in groups of n per refresh cycle. It is assumed, for simplicity, that each pattern is presented only once. In every period the analog contribution to the synaptic value is:

$$A = \sum_{\mu=1}^{n} \alpha^{\mu}_{ij}. \tag{17}$$

For simplicity we take $2J_0 = J_c$, in order to symmetrize the situation for synapses whose value is moving towards the threshold for refresh from above and from below. Given J_c, there will be a threshold value, A_0, for A such that upon thresholding, the synapses can learn. E.g.,

$$\begin{cases} J_{ij}(\lambda) = +J_c & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } A > -A_0 \\ & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } A < -A_0 \\ \\ J_{ij}(\lambda) = \ \ 0 & \rightarrow J_{ij}(\lambda+1) = +J_c \text{ if } A > A_0 \\ & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } |A| < A_0 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } A < -A_0 \\ \\ J_{ij}(\lambda) = -J_c & \rightarrow J_{ij}(\lambda+1) = 0 \quad \text{if } A > A_0 \\ & \rightarrow J_{ij}(\lambda+1) = -J_c \text{ if } A < A_0 \end{cases} \tag{18}$$

See the discussion in Section 2. There, the value of n was 1 and the value of A_0 was lower than 1. In practice A_0 should satisfy:

$$A_0 = \max(J_0, J_c - J_0).$$

Defining $x_0 = A_0/n$, the fraction of synapses depending on α^1, is (see e.g. 8)

$$d \simeq C \cdot 2^{-\frac{p}{n}} \frac{1}{\sqrt{n-1}} \left[\left(1 - x_0 - \frac{1}{n}\right)^{1-x_0} \left(1 + x_0 - \frac{1}{n}\right)^{1+x_0} \right]^{-n/2} \left(1 - \mathrm{erf}\left(\frac{x_0}{\sqrt{2}}\right)\right)^{\frac{p}{n}-1}$$

The constant C depends on the initial connectivity, but not on N, n, p, x_0. Consider the case $x_0 = 0$, i.e. when the number of patterns per group is large compared with the refresh threshold and n odd. For even n there is a slight complication due to the appearance of a fraction of $A = 0$. This fraction decreases rapidly as n increases. If we keep fixed the number of patterns per group, n, as we increase p, we find:

- For at least one synapse to depend on the value of α^1_{ij}:

$$N^2 d \simeq C \quad \Rightarrow \quad p < 2n \log_2 \left(\frac{N}{\sqrt[4]{n-1}} \right)$$

- For naive signal-to-noise we find

$$N d^2 \simeq C \quad \Rightarrow \quad p < \frac{n}{2} \log_2 \left(\frac{N}{n-1} \right)$$

- For vanishing error probability

$$N d^2 \simeq C \cdot \ln N \quad \Rightarrow \quad p < \frac{n}{2} \log_2 \left(\frac{N}{(n-1)\ln N} \right) \tag{19}$$

Note that when n becomes of order p, the first condition limits p only by N^4, while the second limits it by N. This is a reduction to the Hopfield matrix, which preserves some synapses depending on the values of any one of the p patterns until p becomes of order N^4. On the other hand, for small values of n, one obtains a linear increase in capacity. The ratio between the signal to noise capacities for $n = 1$ and 2 is about 2.

If instead it is the number of groups, $g = p/n$, that is kept fixed, one finds (listing the cases in the same order as above):

1.

$$N^2 d \simeq C \quad \Rightarrow \quad p < C \cdot g 4^{-g} N^4$$

2.

$$Nd^2 \simeq C \quad \Rightarrow \quad p < C \cdot g4^{-g}N$$

3.

$$Nd^2 \simeq C \cdot \ln N \quad \Rightarrow \quad p < C \cdot g4^{-g}\frac{N}{\ln N}$$

Note here that as g increases, one is approaching the case of slow learning, one per refresh. It is the factor 4^{-g} which restores the very strong restriction on the capacity found for this case.

Learning with stochastic refresh

Finally, we consider the case in which the network is presented one pattern at a time, but not all synapses, which would be modified in the scheme 5, of are actually modified. Instead, those synapses are modified with probability q. The picture is, basically, that the learning mechanism, when saturating, brings a synapse to just short of the threshold which determines which way it will be refreshed. A source of noise in the refresh mechanism then drives the synaptic value, with a given probability across the threshold. For simplicity we take again $2J_0 = J_c$, as was done in writing Eqs. 18, to equalize the probabilities of learning by the charging and the discharging synapse.

The fraction of synapses dependent on the oldest pattern is (Appendix 8.2):

$$d = C \cdot q\left(1 - \frac{q}{2}\right)^{p-1}.$$

One therefore finds for the three cases:

1.

$$N^2 d \simeq C \quad \Rightarrow \quad p < -\frac{2\ln N + \ln q}{\ln(1 - \frac{q}{2})}$$

2.

$$Nd^2 \simeq C \quad \Rightarrow \quad p < -\frac{\ln N + 2\ln q}{2\ln(1 - \frac{q}{2})}$$

3.

$$Nd^2 \simeq C \cdot \ln N \quad \Rightarrow \quad p < -\frac{\ln(N/\ln N) + 2\ln q}{2\ln(1 - \frac{q}{2})}. \tag{20}$$

The second case, signal-to-noise, gives a lower limit on the probability for modification. Since the ratio must be positive, $q > N^{-1/2}$. When q reaches this limit, the capacity becomes proportional to $N^{1/2}$. A similar result was found by Horner in a different context of stochastic learning [9].

6 Simulations

We have performed simulations to corroborate the analytic results derived above. The simulations were restricted to the case of retrieval with no errors in all three modes of learning – slow, multi-pattern and stochastic. In all three cases the theory predicts

$$p(N) \simeq a \ln\left(\frac{N}{\ln N}\right) + b. \tag{21}$$

The constants a and b are independent of the number of neurons, but vary from case to case and may depend on the number of patterns presented in a refresh interval or on the refresh probability in the stochastic case. The constant a is evaluated in the theory, while b is not. Thus, in what follows, we will be fitting the data (see below) by adjusting the parameter b and using the theoretical value for a. In fact, if $p(N)$ is the measured value of the number of retrievable patterns for a given learned matrix we shall be considering the quantity $[p(N) - b]/a$ as a function of $\ln(N/\ln N)$.

We can read from Eq. 16, for slow learning,

$$a = \frac{1}{2\ln 2}.$$

From Eq. 19 we read, for learning in groups,

$$a = -\frac{n}{2\ln\left(1 - \frac{c}{2}\right)}, \tag{22}$$

where $c = c(n, x_0)$ is the fraction of synapses changed by the presentation of a single group, Eq. 26. Note that for n=1, c=1 and the expression for a reduces to that for slow presentation. Finally, from Eq. 20 we have, for stochastic refresh,

$$a = -\frac{1}{2\ln\left(1 - \frac{q}{2}\right)}.$$

In the simulations we generate sets of random N-bit patterns of ± 1's chosen with equal probability. Those patterns are presented in groups which contain a single pattern for slow and stochastic learning and which contain a variable number, n, of patterns for the case of fast learning in groups. Refresh, as described above is affected after each group. Each pattern is presented once in a learning sequence. The synaptic matrix is updated and clipped according to one of the three learning scenarios. See e.g. Section 2 slow learning, fast, group learning and stochastic learning.

Following every group, the synaptic matrix is frozen and the patterns are presented for retrieval. The patterns are presented, as initial states, starting from the last learned pattern going sequentially into older and older patterns. The network

is allowed to relax for each pattern until it arrives at a fixed point. The overlap with the original pattern is computed and the particular pattern is counted into $p(N)$, if the overlap is greater than 0.99. When a larger fraction of errors is found in a group of past patterns of the size of a learning group, the checking is interrupted. A new group is presented for learning and the process of retrieval testing is repeated. For each learning scenario and for each choice of the corresponding parameters, such as N, n, q, we obtain a value of p. This value fluctuates from sample to sample, i.e. as more and more patterns are presented to the network for learning. The fluctuating values are averaged and the RMS is taken as a measure of the error.

In the simulations the number of groups for any scenario and fixed set of parameters is 40, i.e. we used 40 samples for every network, scenario and set of parameters. To check the N-dependence we have used networks containing from 100 to 1000 neurons. The parameter b in Eq. 21 was determined by averaging the values of $p(N)$ over the 40 samples and fitting Eq. 21 as linear function of $\ln(N/\ln N)$, we determine the parameter b using the analytic result for a. The same procedure was repeated without the $\ln N$ in the denominator, i.e. as if for naive signal-to-noise, with $\ln(N/\ln N)$ replaced by $\ln N$. The differences are not very marked, but to our judgement, the presence of the extra $\ln N$ improves the fit and the scatter of the data. We have not expanded on this issue and present only the data analysis with $\ln(N/\ln N)$.

In Fig. 1 we present the data for stochastic learning (see below). The empty squares represent the case with a refresh probability 1, i.e. slow learning – one pattern per refresh. The points represent the mean of $\ln(N/\ln N)$ over the 40 samples for any given N. The bars around the points represent the error as described above. The line is just $\ln(N/\ln N)$ and the agreement of the line with the data captures both the functional form and the theoretical coefficient a.

Fast presentation in groups

Here for every value of N we have allowed the groups to contain n=2,...9 patterns. The dependence on n is predicted by the theory and is contained in a, Eq. 22. In order to reduce the presentation of the results to a universal curve, we have proceeded as follows: For every value of n and given N we computed the average over the 40 samples. These averages were used in a linear fit to deduce $b(n)$. Since $a(n)$ is given by the theory, we could plot $[p(N) - b]/a$ as a universal curve for all values of n. These results are presented in Fig. 2. For each N the different symbols represent data for different values of n. The plurality of bars reflects the fact that fluctuations increase with n. The line is again just $\ln(N/\ln N)$. The agreement is quite satisfactory.

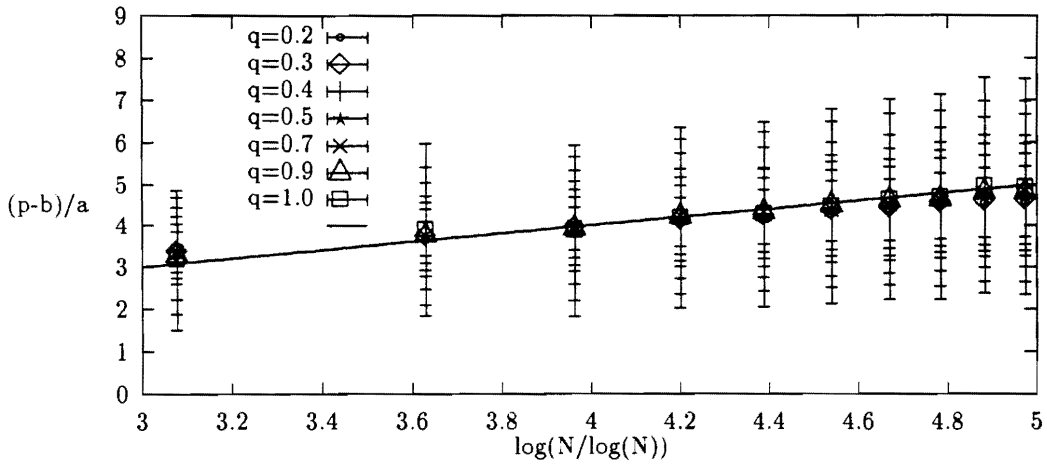

Figure 1: Network capacity vs number of neurons (for $100 < N < 1000$). For stochastic learning, with $q = 0.2, \ldots, 1.0$. q=1, corresponds to deterministic slow learning (empty squares). Successful retrieval is defined as a stimulus leading to an attractor having overlap >0.9 with the memorized pattern. Data points are averages over 40 samples. Different symbols for the same N are averages over values of n. Error bars are RMS fluctuations among the samples.

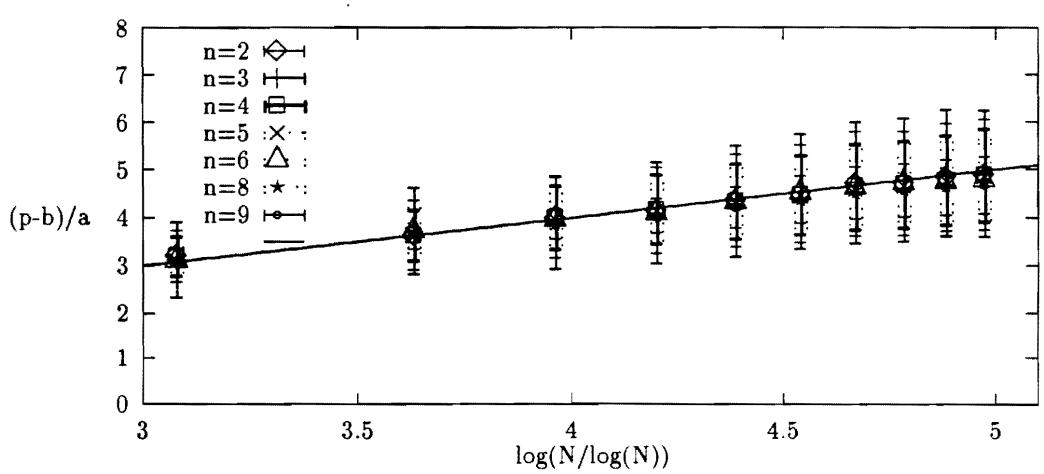

Figure 2: Storage capacity for fast group presentation. The statistics and the networks are as in Fig. 1. The different symbols for the same N represent different values of n.

Stochastic learning

This case is represented in Fig. 1. The patterns are presented for learning one at a time and the synapses are modified, according to 5, but the modification is affected with probability q. The simulations were done for values of $q = 0.2, \ldots, 1.0$. For stochastic learning one does not expect the patterns to be perfectly learned in a single presentation. Especially, for low values of q. Thus our criterion for failure and success in retrieval has been slightly relaxed. We have noticed that the distribution of overlaps, at the fixed points had a large maximum near 1 then a deep trough and then a second maximum towards zero overlaps. These distributions became sharper as the number of presentations increases, but here we concern ourselves with a single presentation od each pattern. We have, therefore, chosen the lowest value (over q) of the overlap at the high end of the trough in the overlap distribution as the lower limit of successful retrieval. It turned out to be about 0.90. Patterns leading to attractors with lower overlaps were considered unretrievable.

Then we proceeded to the presentation of the results as in the case of group learning. The coefficient $a(q)$ is known theoretically, and $b(q)$ is determined by a linear fit of the averages over the samples. All the data was the brought to the universal form by elimination a and b. The different symbols representing the data for the same N are for different values of q. The dashed line in Fig. 1 is again of slope 1 and agrees with the data within the error bars representing the RMS of the statistical fluctuations. Here the fluctuations are larger the smaller the value of q.

7 Discussion

This study has attempted to delineate some of the constraints implied by a refreshing mechanism which is protects learned synaptic information from its eventual decay. The refresh mechanism assumed here is rather restrictive, in that synapses are refreshed to one of three values, depending only on whether the learning has made the synapse cross a threshold, either increasing in amplitude or decreasing it. One may assume other mechanisms, which preserve a more detailed distribution of synaptic values. Such has been the case in the learning model of Shinomoto[10], at the price of assuming a special distribution of uninterrupted inputs. Another option has been suggested by Dong and Hopfield[11], where the saving agent is the persistence of the network in a selective activity attractor.

We have considered a situation in which those escapes are not assured. To preserve learned synaptic information in synapses whose efficacy decays with time, even if during long periods the network receives no input and neuronal activity may die down, one is led to a scheme of the type considered here. The reason is that if the neurons are silent for a long enough period, all synapses will reach the saturation values of the refreshing mechanism, except those which may remain unrefreshed because they have been below some threshold and they will decay to

zero.

The good news is that such a minimally selective mechanism, which selects synapses for refreshing only by a threshold (not by their actual efficacy, nor by their learning history) even when acting continuously, can preserve learned information about attractors. This can provide a scenario for the neuro-biological process of longer term fixation of synaptic efficacies by PKC enzymes generated in the cell body (see .e.g. [14]). One of the problems of this attractive mechanism has been the recognition of synapses which have recently learned, and the selectivity of the reinforcement to the synaptic value. These features are not necessary, we have shown. It is enough that the fixation mechanism be sensitive to a threshold, and otherwise treat all synapses in the same manner.

Moreover, we have shown that despite the constant intervention of the refreshing mechanism, learning of a given pattern can be made very slow on the time scale of the refresh period. This feature is quite important if one would like to preserve the separation of dynamical time scales between retrieval and learning. What was needed for slowing down learning was a mechanism which ensures that not all the synapses which should be modified to learn a given pattern be actually modified. This for ± 1-neurons is problematic, because all synapses get an input of the same amplitude. We therefore had to make recourse either to the interference between a multiplicity of patterns presented in a single refresh period, or to a stochastic refresh mechanism.

This problem points to other, perhaps more natural solutions, which seem to have been preferred by the biological system. These are a combination of the facts that there is no ± 1 symmetry in cortical dynamics. Neurons are either active or not active, and a natural internal representation of a pattern is by the selective group of neurons which are active in the corresponding attractor. Such a representation would imply a significant fraction of neurons to have zero variable, and the others positive, discrete 1's, or analog. Moreover, it seems that in cortex the coding rate, the fraction of neurons with enhanced activity, is low. This reduces even further the fraction of synapses which will be affected upon learning a pattern within a refresh period. Learning, and its restrictions for (0-1) models, or for neurons with analog-rate variables, is presently under study.

The bad news is that storage capacities are significantly reduced. The number of remembered attractors is no longer proportional to the number of neurons, as it is for stable analog synapses even if their values are clipped at the end of a long train of learning. These new constraints follow from rather general arguments, which are made available by the fact that the synaptic distribution becomes very much simplified following every long interval in which neural activity stops. In fact we have found that as long as the number of patterns presented in a single refresh period is small compared to the number of neurons, the best one can do is store $O(\sqrt{N})$ patterns.

In dealing with multi-pattern presentation in a single refresh period we have been making the simplifying assumption that within the period there are no effects

of exponential decay. If the decay is made to be as weak as in ref. [12], i.e. the relative coefficient of two consecutive patterns is $O(\exp(-1/N))$, effectively there is no decay as long as the number of patterns per cycle is small compared with N and our considerations hold. When the number of patterns becomes of $O(N)$, as in the model of Mezard et al, one is presenting in a refresh cycle an extensive number of patterns. Then, the clipping due to the refresh will introduce some effects of noise and leave the capacity at $O(N)$. If patterns are presented at a lower rate, i.e. there is more significant decay between consecutive presentations, the constraints become even more severe than in the case of low rate presentation discussed above. These assertions will be demonstrated elsewhere[15].

It may be that this storage capacity is sufficient. If it is not, and if neural networks, biological or electronic, must survive long periods of neural inactivity, then some escape route is called for. One such possible route has been mentioned above, i.e. rate coding at low coding rate. It may be that the analog nature of the neural variables is essential, perhaps even the operation at rates much below saturation. But those are topics for future studies.

Acknowledgements

This work was supported in part by the ANNETTHE initiative of the INFN. We have benefited from very useful discussions with Simona Genovese as well as from much technical help. We have also had stimulating discussions with Prof. H. Horner and with our collaborators on the hardware part of the project Gaetano Salina and Roberto Riccardi.

8 Appendix

We derive the expressions for the fraction of synapses which depend on an early stimulus, for the cases of fast (group) presentation and for the case of stochastic refresh.

8.1 Fast pattern presentation

Section 2 described this learning scenario, neglecting the intra-group exponential decay. The total contribution of a group can be written as:

$$A = \sum_{\mu=1}^{n} \alpha_{ij}^{\mu} = \tilde{\alpha}_{ij} + \alpha_{ij}^{1}.$$

For the noise due to the other patterns, $\tilde{\alpha}$, not to destroy the imprint of pattern number 1, within the group, we must have:

$$\pm A_0 - 1 < \tilde{\alpha} < \pm A_0 + 1. \tag{23}$$

Since the distribution of $\tilde{\alpha}$ is binomial, for the type of patterns employed, the probability that this condition be satisfied is:

$$\underset{in}{\Pr} = \binom{n-1}{\frac{n-1}{2} + \frac{A_0}{2}} \left(\frac{1}{2}\right)^{n-1} \tag{24}$$

Using Stirling expansion and expressing the result in terms of $x_0 = A_0/n$, we find, for the probability that a pattern internal to the first group be recorded in the synaptic value:

$$\underset{in}{\Pr} \simeq C \cdot \frac{1}{\sqrt{n-1}} \left[\left(1 - x_0 - \frac{1}{n}\right)^{1-x_0} \left(1 + x_0 - \frac{1}{n}\right)^{1+x_0} \right]^{-n/2}. \tag{25}$$

In order to calculate the probability that the memory of the patterns of the first group is not destroyed by the subsequent groups, we proceed as follows. We have seen in Section 3 that memory is destroyed by consecutive refresh changes which act in the same direction. Thus, we are restricted to sequences of groups of n patterns which do not impose two consecutive changes of the same sign. But, in contrast to the slow learning case, here there may be intermediate refresh periods in which there is no modification, i.e. in which the condition $|A| > A_0$ is not satisfied. Thus, we must consider all group sequences of alternating signs, that may have any number of intervening no-modifying refresh cycles between any two effective refresh cycles.

Let r be the number of modifying refresh cycles out of the $p/n - 1$ cycles following the first one. The number of memorizing sequences is the number of sequences which have $(p/n - 1) - r$ zeros, standing for the non-modifying groups, and r alternating modifications. Suppose that the arriving sequences of random patterns have a distribution of bits such that

$$\Pr(A > A_0) = \Pr(A < -A_0) = \frac{c}{2}$$

$$\Pr(-A_0 < A < A_0) = 1 - c,$$

then the fraction of sequences with exactly $p/n - 1 - r$ zeros is:

$$\binom{p/n - 1}{r} c^r (1-c)^{p/n-1-r}.$$

Among those only sequences with alternating sign modifications keep the memory in a synapse. Their probability, for a sequence of r modifications is proportional to $2{-}r$. Thus the probability that the memory created by the first group is not erased by the $p/n - 1$ subsequent groups is:

$$\Pr_{out} = C \sum_{r=0}^{p/n-1} 2^{-r} \binom{p/n-1}{r} c^r (1-c)^{p/n-1-r} = C \left(1 - \frac{c}{2}\right)^{\frac{p}{n}-1}.$$

The total fraction of memory maintaining synapses is given by:

$$d = \Pr_{out} \cdot \Pr_{in} \simeq C \cdot \frac{1}{\sqrt{n-1}} \left[\left(1 - x_0 - \frac{1}{n}\right)^{1-x_0} \left(1 + x_0 - \frac{1}{n}\right)^{1+x_0} \right]^{-n/2} \left(1 - \frac{c}{2}\right)^{\frac{p}{n}-1}.$$

c is the probability that a group will effect a modification. It can be expressed in terms of the n and the modification threshold x_0. A modification will take place if $A > x_0$. If k is the number of positive bits, out of a group of n,

$$A = k - (n - k)$$

and the corresponding probability is

$$\Pr(A = k - (n-k)) = \binom{p-1}{k} \frac{1}{2^n}.$$

Consequently, the probability for a modification is:

$$c(n, x_0) = 2 \sum_{k > n(x_0+1)/2}^{n} \binom{p-1}{k} \frac{1}{2^n}. \tag{26}$$

where the extra factor of 2 represents the fact that there are negative as well as positive modifications. When n becomes very large,

$$c(n, x_0) = 1 - \operatorname{erf}\left(\frac{x_0}{\sqrt{2}}\right).$$

8.2 Stochastic refresh dynamics

In this case there is a probability q that a synapse changes upon presentation of a modifying pattern. Any synapse sees a different history of modifications, even for the same incoming set of input bits. One finds a situation similar to that discussed in Appendix 8.1, in which different synapses may receive varying numbers of non-modifying inputs among the p arriving bits. The fraction d is again the sum over all possible numbers of zeros in a sequence of $p-1$ bits, with the remaining bits of alternating signs. The difference is that in the present case the probability of modification, q, is fixed, while in the group presentation it was to be calculated from the properties of the group. The result is:

$$d = C \cdot q \sum_{k=0}^{p-1} \binom{p-1}{k} (1-q)^{p-1-k} q^k \frac{1}{2^k} = C \cdot q \left(1 - \frac{q}{2}\right)^{p-1}.$$

8.3 Asymptotic synaptic distribution

Whatever the initial distribution of synaptic values, after a long period of inactivity, the refresh mechanism will produce a three-way distribution $\pm J, 0$. Suppose that after k refreshed learning steps the clipped distribution of the synapses has:

$$c_+^k,\ c_-^k \ \text{ and } \ c_0^k = 1 - (c_+^k + c_-^k),$$

for the fractions of synapses with the values indicated by the subscripts. Since the arriving patterns have a symmetric distribution of positive and negative values, in a learning situation, after the presentation of the next pattern the distribution will be:

$$c_+^{k+1} = \frac{1}{2}(1 - c_-^k); \quad c_-^{k+1} = \frac{1}{2}(1 - c_+^k).$$

This distribution converges, exponentially fast, to $c_+ = c_- = c_0 = \frac{1}{3}$.

References

[1] Amit DJ, Fusi S, Genovese S, Badoni D, Riccardi R and Salina G 1992 LANN: Learning attractor neural network, Model and hardware implementation, (INFN internal report, in Italian)

[2] Miyashita Y and Chang HS 1988 Neuronal correlate of pictorial short-term memory in the primate temporal cortex, *Nature*, **331** 68 and Neuronal correlate of visual associative long-term memory in the primate temporal cortex, ibid **335** 817

[3] Hebb D 1949 *The Organization of Behavior* (Wiley, NY)

[4] Hopfield JJ 1982 Neural networks and physical systems with emergent selective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**, 2554

[5] Sompolinsky H, The theory of neural networks: The Hebb rule and beyond, in L. van Hemmen and I. Morgenstern eds. *Heidelberg Colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987)

[6] Nadal J-P, Toulouse G, Changeux JP and Dehaene S 1986 Networks of formal neurons and memory palimpsests, *Europhys. Lett.*, **1** 535

[7] Parisi G 1986 A memory which forgets, *J. Phys.*, **A19** L617

[8] Weisbuch and F. Fogelman-Souliè. Scaling laws for the attractors of Hopfield networks, *J. Physique Lett.*, **2**,337(1985)

[9] Horner H 1992, private communication.

[10] Shinomoto S, Memory-maintenance in neural networks, *J. Phys.*, **A20**, L1305 (1987)

[11] Dong DW and Hopfield JJ 1992 Dynamic properties of neural networks with adapting synapses, CALTECH preprint.

[12] Mezard M Nadal J-P and Toulouse G 1986 Solvable models of working memories, *J. Physique* **47** 1457

[13] Amit DJ 1989 *Modeling Brain Function* (Cambridge University Press, NY)

[14] Alkon DL 1989 Memory storage and neural systems, *Scientific American*, July 26.

[15] DJ Amit, S Fusi and L Cugliandolo 1992, in preparation.