ROME-933-1993

PREPRINT N. 933
March 30, 1993

Dipartimento di Fisica
Universita di Roma "La Sapienza"
INFN – Sezione di Roma

Fermilab Library
0 1160 0029144 7

Dynamic Learning in Attractor Neural Networks

Daniel J. Amit

INFN, Sezione di Roma, Istituto di Fisica
Università di Roma, La Sapienza, Ple Aldo Moro, Roma[1]

Lectures at the ACAI 93 School, Capri June 1993

March 31, 1993

Contents

1 Introduction **2**
1.1 Experimental evidence for the existence of attractors 2

2 Conversion of temporal to spatial correlations **4**
2.1 Lessons from Miyashita's monkey 4
2.2 Cognitive psychology implications: priming and false alarms 5

3 Toy models **6**
3.1 The Hopfield model 6
3.2 Capacity and performance in associative retrieval 8

4 Correlated attractors and uncorrelated patterns **9**

5 Learning attractor neural networks (LANN) **11**

6 The problem of analog depth problem **14**
6.1 Sompolinsky's solution 14
6.2 The practical problem of continuous refresh 15

7 Learning in material devices **17**
7.1 Slow pattern presentation with deterministic refresh 17
7.1.1 The basic constraint : 18
7.2 General features of signal-to-noise analysis 19
7.3 Deterministic slow presentation of ± 1 patterns 21
7.4 Logarithmic limitation for multi-state synapses 22

8 Going beyond the logarithmic constraint **24**
8.1 Elements of leverage 24
8.2 Stochastic learning of sparsely coded patterns 24
8.3 Performance: absolute maximal storage in a palimpsest 26

[1]On leave from Racah Institute of Physics, Hebrew University, Jerusalem

1 Introduction

This set of lectures extends a certain unity between cognition, neuro-biology and theoretical modeling, see e.g. [1], to some consequences it implies for the concomitant learning processes. I will proceed from some experiments in neuro-physiology to the type of models which may account for them. I will argue that experiment indicates that conventional ideas about the boundary between computation and learning are not implemented in the operation of brains. Hence learning on a realistic substrate and the type of internal representations it can produce is an essential step in capturing cortical computational principles.

From the point of view of artificial intelligence the brain may not be a bad source of ideas for future development. The constraints discovered in following closely the functioning of live brains, or of devices trying to imitate them may narrow significantly the immense space of possible intelligent machines that should be studied in order to reproduce brainy function.

The first lecture will be devoted to discussing the experimental evidence for attractor behavior (reverberations) in the cognitive processes of performing monkeys, as observed in neurophysiology and to the type of models which can capture attractor behavior in a neural network. In the second lecture I will discuss the concepts of learning and memory in attractor neural networks (ANN). The third lecture will be devoted to the description of a material substrate, in electronics, which can learn attractors from an incoming data flow. This will lead to a discussion of the special constraints imposed on learning and memory in a material structure. The last lecture will formalize and generalize the constraints and draw the theoretical consequences which in turn reflect on the engineering of the device.

1.1 Experimental evidence for the existence of attractors

In the basic experiment performed by Miyashita et al. [2] a set of computer generated images is prepared and a monkey is trained to perform delayed visual image matching. In order to avoid correlations in the geometric structure, the pictures are generated using fractal algorithms which produce presumably uncorrelated images. The task of the monkey

is as follows: following a short warning signal, one of the generated pictures is shown on the screen for a short time ($200ms$). Next, the screen is blanked and, after as long as **16** seconds, another picture is presented and the monkey is trained to respond differentially if the second image is the same as the first or is different. When the monkey begins to perform well and gives a high percentage of correct answers the recording starts.

Miyashita's group was able to find a very small area ($1mm^2$ some 100,000 cells) of the anterior ventral temporal cortex (AVT) where selective activity takes place even in the absence of the stimulus, i.e. during the 16 second delay. The first remarkable fact is that the selective persistent activity goes on in a very localized area. This fact constitutes experimental evidence for the existence of a localized mechanism which is able to maintain a reverberation for as long as 16 seconds. Furthermore, when following training the pictures are presented in a random order, one observes that the spike rate distributions are reproducible: the same image produces the same persistent elevated activity distribution, even if between the two presentations many other images were shown at random.

In Fig. 1 there are four displays from ref. [3]. The rasters above each histogram constitute a representation of spike activity of a certain neuron recorded during twelve different repetitions of the same picture. The reproducibility is quite clear in all the four cases. Furthermore this figure is an experimental manifestation of the effects typical of collective behavior of systems with a large number of degrees of freedom. For example, in case (a) a neuron active during the presence of the stimulus maintains high activity in the delay period; in case (b): during the 200 ms when the external stimulus is on, the activity level is very low. However, when the stimulus is removed, this neuron's activity goes up and reaches the same level as in case (a). The interpretation of this instance in terms of attractors and reverberations is quite simple: during the period in which the stimulus is on it maintains some activity distribution. But when the stimulus is removed, the high activity level is maintained in the reverberations which are governed by the structure of the synaptic connections. Actually the system is correcting the 'error' on the basis of the acquired past experience which has structured the attractors by forming the synaptic structure. A similar, though reversed reading is given to the neuron in case (c) that is activated by the stimulus but is quiescent in the reverberation. It is the opposite type of error. The interpretation of (d) is left as an exercise.

This connects to the properties of associative memory. Partial information in a stimulus is sufficient for the retrieval from memory of a full description of the memory item connected to the stimulus. It is in this context that we interpret the modification of the activity of the neuron in case (b).

There is another interesting behavior which characterizes this particular part of anterior ventral temporal cortex [2]: varying the size, the angle of rotation or the color of an image, one finds an internal activity distribution which is essentially the same. This means that in this part of the cortex the attractors are already an internal representation of some prototypes and stimuli which differ for size, angular position or color have essentially the same internal representation and all of them flow into the same attractor.

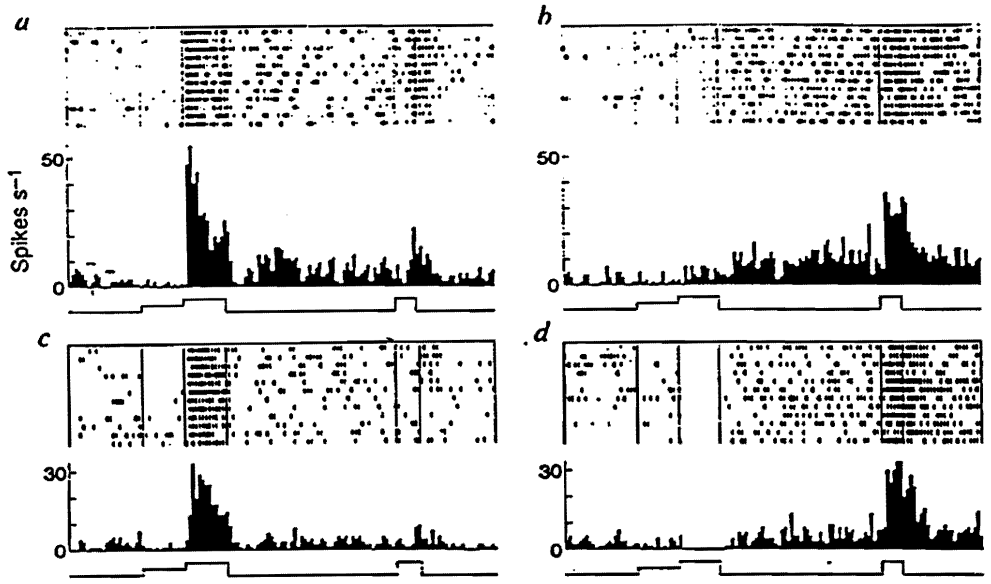

Figure 1: Four types of neuronal behavior (in each part of the figure there are 12 rasters and a spike rate histogram below): a) Active in presence of stimulus and persists in absence; b) Neuron active in the delay period of a different stimulus; c) Active in presence of stimulus, inactive in absence; d) Inactive in presence of stimulus and active in delay period. (From ref. [3]).

2 Conversion of temporal to spatial correlations

In the second experiment [4] the pictures are shown in a fixed order during training: each picture is always followed by the same picture. A sampling of the internal representations is reconstructed by recording the activity of single neurons from the AVT cortex. Subsequently, the correlation between internal representations corresponding to different patterns is calculated off-line. It is found that the correlation between representations depends only on the distance separating the two patterns in the training sequence.

In order to prevent accidental correlations in the internal representation of the images, two monkeys were trained on the same set of images but the patterns were presented in a different order. For both monkeys the same result was obtained: the correlation between two internal representations becomes insignificant at the separation of about 5 in the training sequence. We therefore conclude that in this part of the cortex the synaptic structure created during learning has correlated attractors even though the items learned were uncorrelated. In other words, presenting uncorrelated pictures in repeated, fixed temporal order produces correlated internal representations, the attractors, that preserve information concerning the temporal order.

2.1 Lessons from Miyashita's monkey

- There are long-lived elevated spike rate distributions selective to the stimulus. Ergodicity is broken and selectivity is possible. The activity distributions are self-sustained as they persist long after the disappearance of the stimulus.
- The reverberations can be observed in localized modules ($1mm^2$, 10^5 neurons)

- The reverberations are reproducible on the single unit level. The activity of a single neuron depends only on the stimulus leading to the reverberation and it is always the same when the same pattern is presented. A particular neuron may be shared by several different representations.
- Persistent high rates are not single neuron properties (collective behavior).
- The internal representations are of prototypes: in this area they do not depend on angular position, size, color etc.
- The internal representations are distributed, i.e. many neurons participate in representing a pattern.
- The reverberations are attractors in the sense that some neurons which are activated during the presentation of the stimulus stop during the delay period. So the network is able to perform error correction. A class of similar stimuli leads to the same activity configuration.
- The internal representations code for temporal correlations: patterns which are temporally correlated in the training phase have correlated internal representations.

2.2 Cognitive psychology implications: priming and false alarms

It is not too surprising that if two uncorrelated objects appear very often in temporal contiguity, then thinking about one, the other can be easily evoked. There is a very familiar phenomenon in cognitive psychology called priming. Subjects are asked to identify distorted visual patterns and the time for the identification of the correct patterns is measured. This time is compared to the time required to identify distorted images preceded by a pattern that is cognitively correlated with the image to be identified. For example, showing first the full word "chair" and then a distorted version of the word "table", the identification time is significantly shortened.

The interpretation in terms of attractors is quite straightforward: when the cortical module is in an attractor which is correlated with the internal representation (the attractor) of the incoming stimulus, then the time the system takes for moving into the new attractor is much shorter than the time required for a transition from an uncorrelated state. This interpretation immediately makes a prediction about a neuro-physiological experiment. One can measure priming on the level of AVT cortex by observing transition times between internal representations.

What is of great relevance is that here we find a potential embryo of context sensitivity. In other words, the internal representations of learned items depend not only on what the corresponding stimuli had been, but also on the context in which they had been learned: who appeared shortly before or after during learning. This is a significant fact since it displaces a part of 'syntax' from the application of rules to the formation of representations during learning. Moreover, the development of these correlated representations is also a potential opening in the direction of 'semantics'. What I mean by that, in a very tentative

way, is that semantics is in part the spectrum of other experiences in which a perceived object is interpreted by the perceiving subject.

But before embarking on these slippery extrapolations, I would prefer to lower the sights and follow more closely the phenomenology of the findings and of the candidates for a theoretical description. This implies focusing on two aspects:

- The presence of persistent states of stimulus selective activity in an assembly of neurons, in the absence of the provoking stimulus;
- The formation of the synaptic structure which can sustain these internal representations, i.e. the corresponding dynamics of unsupervised learning or self-organization.

3 Toy models

Miyashita's experiments have raised two essential problems: the first one concerns the type of dynamics which can maintain a selective distribution of neuronal activity in absence of the stimulus. The second problem is the way the system creates correlated attractors from patterns that are uncorrelated. In order to solve these two problems I will describe some toy models which reproduce the main features of the Miyashita experiments. Toy models are a powerful tool for investigating very complicated systems and turn an untractable problem into a tractable one. They represent the essential features of the systems, indicate certain phenomena and make predictions. If the results are satisfactory, their robustness to the removal of the simplifying assumptions can be tested.

3.1 The Hopfield model

The Hopfield model[5, 6] provides a solution to the first problem, that is how to maintain selective activity distributions. The distribution must be sensitive to the stimulus but not too sensitive, in the sense that internal representations of very similar stimuli must lie in the same basin of attraction so that the system is able to perform error correction. On the contrary, if two stimuli are different enough, then the system has to evolve to different attractors enabling the possibility of discriminating between two different representations.

In the network, the Hebbian assembly, corresponding to the module observed in the Miyashita experiments, N neurons are connected by synapses. A state of the network at a given moment is a distribution of $+1$,-1's which specifies which neurons are emitting spikes and which are not. The state of the N neurons belonging to the network can be completely specified by assigning an N-bit word to each time slice. The dynamics of the network is as follows: first a neuron computes its depolarization relative to its threshold, due to the incoming signals from the other neurons. The synaptic variable J_{ij} is the depolarization of neuron number i due to a single spike in neuron number j. This is the synaptic efficacy and it can be positive or negative – excitatory or inhibitory. We will imagine them to be randomly mixed.

The difference between the depolarization and the threshold is:

$$h_i(t+\delta t) = \sum_{i=1}^{N} J_{ij} s_j(t) \tag{1}$$

which determines the subsequent state of the neuron, according to:

$$s_i(t+\delta t) = \mathrm{sign}[h_i(t+\delta t)] \tag{2}$$

The new state of the neuron is +1 if the depolarization is above the threshold ($h_i > 0$), i.e. the neuron emits a spike, and is −1 otherwise, no spike. We consider the process to be deterministic. In reality there are many noisy processes which affect the time evolution of the activity (See e.g. ref [6]). We shall ignore them for the purposes of the present discussion.

For this simple model[5, 6] it is possible to construct a specific synaptic matrix J_{ij} which has an arbitrary set of uncorrelated patterns as attractors. In other words, it is possible to choose the states one wants as attractors among the very large set of random uncorrelated N-bit words in which each bit is chosen at random, with equal probability, to be either +1 or −1 and to construct a J_{ij} which has them as attractors. This is done as follows: defining ξ_i^μ as the state neuron i should be in when the network settles in attractor number μ, the connection between neuron i and neuron j will be the sum of the products of activities of neuron i and neuron j in each one of the p patterns to be memorized in the network:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{3}$$

To see that this synaptic matrix maintains the activity distribution in every one of the patterns that compose the synaptic matrix, one considers the signal and noise parts of the quantity $h_i s_i$ produced by this synaptic structure when the network is in one of these states, number ν for example, i.e. when $s_i = \xi_i^\nu$. The condition for stability is $h_i s_i > 0$, which implies that with the dynamics Eq. 2, the new s_i will be equal to the old one. The relevant quantity is therefore,

$$\xi_i^\nu h_i = \sum_j J_{ij} s_j s_i = \frac{1}{N} \sum_j \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_i^\nu = 1 + R_i. \tag{4}$$

The first, the *signal* term is the contribution of the term $\mu = \nu$ in the synaptic matrix. It is also the average of $h_i \xi_i^\nu$ over all the sites. R_i is the part contributed by the $p-1$ patterns ξ^μ with $\mu \neq \nu$. It is the noise describing the fluctuations from site to site:

$$R_i = \frac{1}{N} \sum_j \sum_{\mu \neq \nu} \xi_i^\nu \xi_i^\mu \xi_j^\mu \xi_j^\nu \tag{5}$$

Since all the bits of the patterns are uncorrelated, δ_i is a one-dimensional random walk of $(N-1)(p-1)$ steps of size N^{-1}. Its RMS is:

$$\overline{R^2} \simeq p/N.$$

If p is much less than N, then R cannot change the sign of h_i, which is just the sign of ξ_i^ν. This means that starting in a state ξ_i^μ, stored in the synaptic structure, the network remains there for an indefinite time, until a new stimulus arrives. Furthermore, these fixed points are true attractors with a certain basin of attraction. Starting from a state which resembles one of the stored pattern, the system is able to perform error correction by relaxing into the nearest attractor and remaining there. The system works as an (auto) associative memory even in the presence of noise in the dynamical process of spike emission, see e.g. ref. [7, 6].

This is expressed in Fig. 2 where the time development of the network is followed by monitoring the *overlaps* with the stored patterns, defined as:

$$m^\mu = \frac{1}{N}\sum_{i=1}^{N} s_i \xi_i^\mu. \tag{6}$$

These are natural variables which measure the macroscopic distance between the present state of the network and each of the stored patterns (see e.g. ref. [6]). Note that they can vary from +1, when the network state is identical (bit-by-bit) to one of the patterns, to 0 when the network state is uncorrelated with the pattern to −1 when it is anti-correlated with it. It is a simple transformation of the Hamming distance per bit.

The network of 400 neurons stores 7 patterns in the form Eq. 3. The overlaps are plotted for two different initial configurations, which have significant resemblance to the first stored pattern ξ_i^1. The initial state has an overlap ($m^1 = 0.35$ – 32% errors, $m^1 = 0.20$ – 40% errors) with the first stored pattern. The spike dynamics increases the overlap with the first pattern rapidly to 1 and all the errors are corrected dynamically. The overlaps with other patterns fluctuate around 0 and their value is negligibly small when compared to overlap with the first stored pattern. This happens also in the second case in which the number of "errors" in the initial state is significantly larger than in the first case.

Note that the network relaxes into a fixed point. This does not imply that the system is quiet. At a fixed point half of the neurons are emitting a spike every 2ms and half the neurons are quiescent. The definition of the model implies that at a fixed point the neurons which are 1 emit a spike in every interval of time discretization, i.e. in every $2ms$ (the absolute refractory period). Neurons which are fixed at -1 emit no spikes. Thus the fixed point is a very active state of the modeled neural system, in which a selective half of the neurons emit 500 spikes/sec. Two factors determine this distribution of activity:

- The synaptic matrix – the learning process;
- The initial state – the stimulus.

3.2 Capacity and performance in associative retrieval

The signal-to-noise analysis described above was a simple way of arriving at an estimate of the number of *random* patterns of ±1 bits in a network of the Hopfield type. That

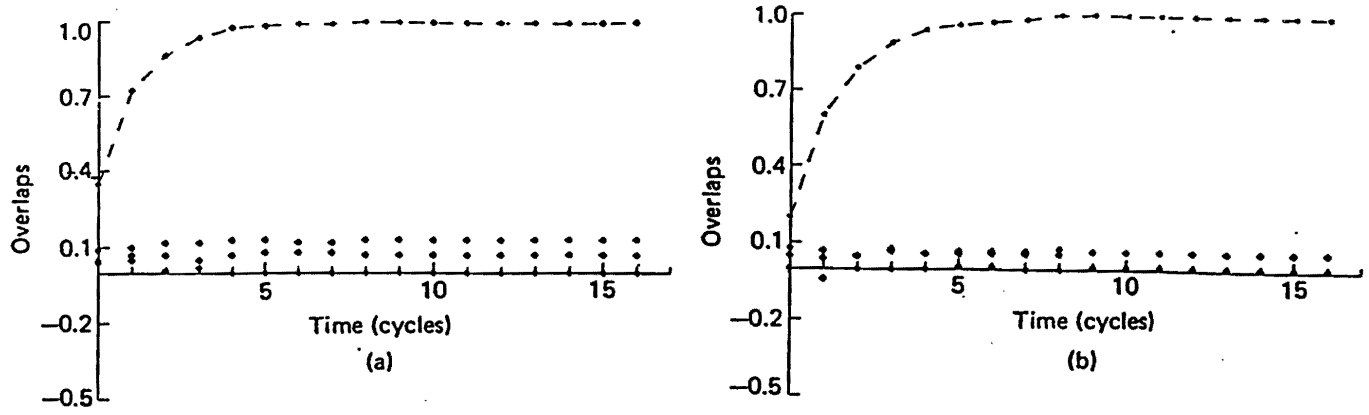

Figure 2: Time development of overlaps of a network of 400 neurons with 7 stored patterns in two simulations of the Hopfield model with noiseless asynchronous dynamics. (From ref. [6])

analysis can be deepened and extended as follows:

- If the actual gaussian probability probability distribution of the neural input, h_i is considered, it can be shown[8, 6] that no errors will occur in the retrieval of any of the patterns provided $p < O(N/\ln N)$;
- If errors are allowed, but one requires that there be stable attractors near to the memorized patterns (in the Hamming sense), then in the absence of noise in the neural dynamics, the system will function well until

$$p \approx 0.14N.$$

  The fraction of errors in retrieval at extreme memory loading is less than 1.5%[9, 6];
- If one does not restrict the synaptic matrix to the Hopfield form, one can show that synaptic matrices exist such that 2N random patterns can be stored as attractors with no errors[10];
- Storing patterns with 0–1 bits, sparsely coded (i.e. with low fraction of 1's) one can reach a storage capacity of $(N/\ln N)^2$ patterns, as the coding rate tends to zero, yet keeping a finite information storage per synapse[11. 10, 12, 13].

It can further be shown that in many cases these properties of the networks are rather robust to various types of noise and synaptic disruption. See e.g. ref. [6] ch. 7.

4 Correlated attractors and uncorrelated patterns

The second question raised at the beginning of this section concerned the correlations between internal representations of patterns which are presented in a fixed order during the

training period. It turns out that it is possible to reproduce correlations of the Miyashita type in a very simple toy model[14]. Take the uncorrelated patterns $\xi_i^\mu = \pm 1$ as the distribution of activities in the stimuli to be learned. The dynamics of neural activity is described again by Eqs. 1, 2. However the synaptic matrix includes a term that reflects the special training process of fixed order presentation:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}[\xi_i^\mu \xi_j^\mu + a(\xi_i^{\mu-1}\xi_j^\mu + \xi_i^\mu \xi_j^{\mu-1})] \tag{7}$$

The idea is that in the synaptic matrix the information concerning temporal correlation is stored by the second term in the square brackets. If $a = 0$, then we have the Hopfield matrix. Otherwise there exists a term which connects the structure of one pattern to the preceding one. The last term preserves the symmetry of the Hopfield matrix so that all attractors are fixed points in the absence of noise and therefore the analysis is simple.

The additional term can, in principle, be set up in a learning scenario. During the training period the pattern number μ–1 is always shown before pattern number μ. First the network learns the uncorrelated stimuli as uncorrelated attractors. When stimulus μ–1 is presented the network relaxes into the attractor number μ–1, and persists there. When stimulus μ is shown the network makes a transition to the next attractor, the one representing the consecutive stimulus. During the transition the information of both patterns is available and can be used for learning the mixing term in Eq. 7. Note that this type of information is available due to the framework of attractor dynamics which I have discussed in the previous section. When stimulus μ–1 is removed the information about it is stored and maintained active in the corresponding attractor.

This picture implies in particular that if the monkey has not been trained long enough, one should expect uncorrelated stimuli to produce uncorrelated internal representations. That is a prediction that could be verified neuro-physiologically.

The neural dynamics driven by the synaptic matrix 7 is as follows: the difference between the depolarization and the threshold, Eq. 4, in terms of the overlaps Eq. 6, reads:

$$h_i = \sum_{\mu=1}^{p} m^\mu(t)[\xi_i^\mu + a(\xi_i^{\mu+1} + \xi_i^{\mu-1})]. \tag{8}$$

Starting from a state identical to pattern ξ_i^2, i.e. which has overlaps $m^2 = 1$, and $m^\nu = 0 \ \forall \nu \neq 2$, the expression for the depolarization, Eq. 8 becomes:

$$h_i = \xi^2 + a(\xi_i^3 + \xi_i^1).$$

If $a < 1/2$, then the original uncorrelated patterns ξ^μ are still attractors because it the first term on the right hand side which dominates the sign of h_i. On the other hand if $0.5 < a < 1$ the uncorrelated patterns ξ_i^μ are no longer stable points. Indeed, starting from $m^2 = 1$, there is a probability 0.25 that the second term dominate. So, after one step of the dynamics one finds[14] that there are three non vanishing overlaps: $m^2 = m^1 = m^3 = 0.5$.

But now, at the next step, two new terms will contribute to the sum in expression (8) and therefore also the overlaps with patterns at distance two from ξ_i^2 will be non zero.

The fixed point attractors of this network, corresponding to the stimulation by one of the original uncorrelated stimuli can be found exactly[14, 15]. Each attractor has exactly 5 non-zero overlaps. Starting from a state identical to one of the stored, uncorrelated patterns, one obtains the following iterations for the overlaps with the stored patterns:

$$
\begin{array}{ccccccccccc}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{8} & \frac{3}{8} & \frac{5}{8} & \frac{3}{8} & \frac{1}{8} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{32} & \frac{3}{32} & \frac{13}{32} & \frac{19}{32} & \frac{13}{32} & \frac{3}{32} & \frac{1}{32} & 0 & 0 \\
0 & \frac{1}{128} & \frac{3}{128} & \frac{13}{128} & \frac{51}{128} & \frac{77}{128} & \frac{51}{128} & \frac{13}{128} & \frac{3}{128} & \frac{1}{128} & 0 \\
0 & \frac{1}{128} & \frac{3}{128} & \frac{13}{128} & \frac{51}{128} & \frac{77}{128} & \frac{51}{128} & \frac{13}{128} & \frac{3}{128} & \frac{1}{128} & 0
\end{array}
$$

From those it is posssible to calculate the exact correlations between different attractors. The correlations between one generic attractor and the attractors corresponding to the patterns neiboghring in the training sequence are:

$$
\cdots \frac{4}{256} \quad \frac{11}{256} \quad \frac{32}{256} \quad \frac{85}{256} \quad \frac{170}{256} \quad 1 \quad \frac{170}{256} \quad \frac{85}{256} \quad \frac{32}{256} \quad \frac{11}{256} \quad \frac{4}{256} \cdots
$$

The resulting attractors are correlated beyond the first neighbor memorized in the synapses and become small at a separation of 4–5. It is remarkable that this result does not depend on the value of the constant a: the only requirement is that a lie in the interval $[0.5, 1]$. Moreover, these results capture the qualitative features observed in the experiment[4].

In Fig. 3 we display side by side the correaltions of attractors found in the experiment of Miyashita[4] and in the toy model calculation. There is a surprising qualitative agreement. The actual values of the correlation coefficients are different, but this seems to be attributable to the way in which correlations are computed in ref. [4], i.e. as rank coefficients.

5 Learning attractor neural networks (LANN)

The preceding discussion underlines the following essential points:

- Attractors are employed by cortex and can be considered as active internal representations of learned external stimuli;
- In the structure of the attractors is embedded information about the context in which learning took place, which in turn expresses already a partial computation of certain tasks;

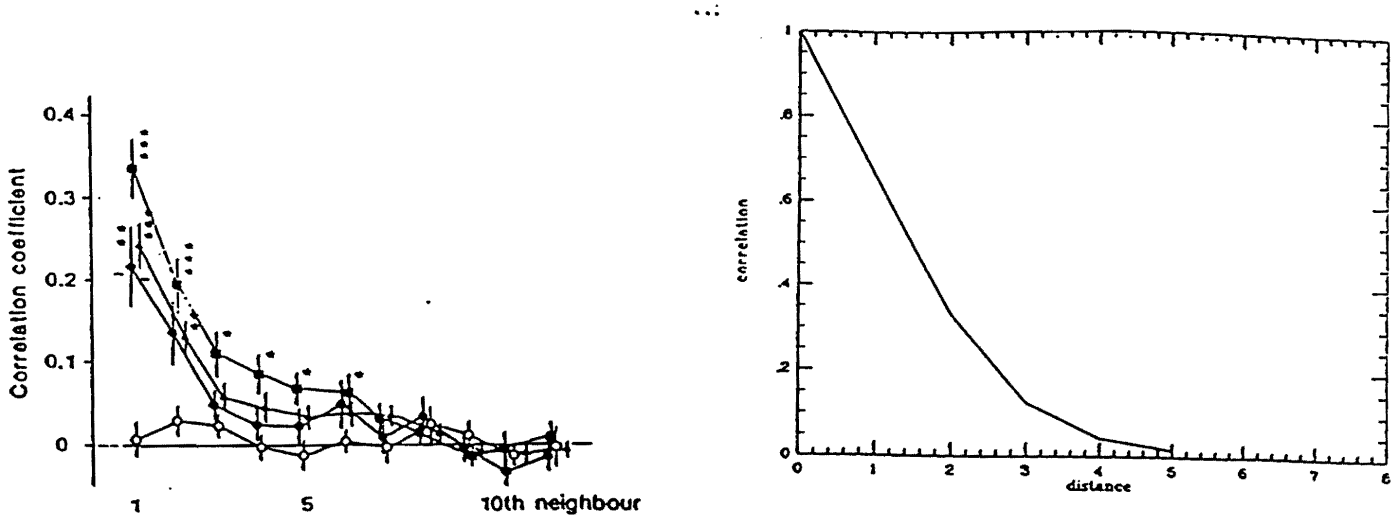

Figure 3: Correlations between attractors: left, delay activity population correlations[4] ; right, attractor correlations in the ± 1 model [14].

- The structure of the attractors is the result of learning, during attractor formation;
- The context dependence of the attractors, or the internal representations, may be modified if the structure of the input flow changes;
- The training process is unseparable from the performance, i.e. the recognition efforts of the monkeys.

These considerations have led us to consider a material system in which learning of attractors can take place in a mode which is indistinguishable from retrieval. The natural substrate is silicon and the formulation of the requirements has been the following:

- Learning, as synaptic dynamics, should be continuous. No learning vs function phases;
- Learning unsupervised;
- Network able to handle a natural flow of incoming stimuli with no constraints on ordering or temporal distribution of arriving stimuli;
- Memory maintenance for long times, even in absence of stimuli and of network activity;
- End product of learning classification capability of the network;
- The network must be implementable in hardware;
- Overloading must not lead to memory blackout catastrophe (palimpsest property).

Figure 4: The pilot LANN with 3 neurons and 3 synapses in discrete elements with asynchronous stochastic refresh.. For details see ref. [16].

This project, carried out by a Rome I Rome II collaboration, see e.g. refs. [16, 17], has completed several stages of design and development and has brought forth several basic constraints of learning in realistic devices. Neurons implemented by high gain operational amplifiers and synapses by condensers. The benchmark device which has been completed, at the writing of these lines is shown in the photograph in Figure 4. It is a network of 3 neurons and 3 synapses which performs in perfect accordance with with theoretical expectations of stochastic learning and retrieval, when exposed to a natural stream of stimuli. It essentially satisfies all the requirements listed above.

A network of 3 neurons and 3 synapses is a far cry from the 10^{11} neurons in the cortex or the 10^5 neurons in the module registered by Miyashita. For us it has been a drawing board for the exploration of dynamical principles connecting learning with retrieval, principles that had to be assimilated before a large scale project is embarked upon. It has been so designed that it naturally expands into a 27 neuron network with the accompanying 27×13 synapses (see e.g. [16]), which is the last step before VLSI integration. Yet, however small this network it has already produced many lessons, which we proceed to discuss.

Learning in this network is affected in a Hebbian framework[18], which provides a natural mechanism for learning attractors. I believe, as a matter of fact, that Hebb introduced his learning paradigm precisely in order to describe the learning of attractors, which he termed *reverberations*. 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 activity of the two neurons connected by it, will lead to the creation of a corresponding attractor. One of the attractions of the Hopfield model[5] 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. One would then also have the high storage capacity, both in terms of number of retrievable patterns and in terms of information per synapse. But, 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. But even in this ideal case, such a network would suffer from the 'blackout catastrophe', namely learning too much produces total memory loss.

6 The problem of analog depth problem

6.1 Sompolinsky's solution

The problem of the analog depth of the synaptic values was treated by Sompolinsky[19, 6]. 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. This is described in Figure 5a. 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.

Starting from a synaptic matrix like Eq. 3, i.e.

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu,$$

one defines the non-linear transform of each element as:

$$T_{ij} = \frac{\sqrt{p}}{N} F(K_{ij}) \quad K_{ij} = \frac{1}{\sqrt{p}} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu.$$

The function $F(x)$, for the 3-way clipping has the form depicted in Figure 5. Following this operation each synapse has one of three values, one of which is 0. The synapses which have zero efficacy are effectively disconnected. Following the clipping the connectivity of the resulting network is diluted.

Sompolinsky has shown that for p large, the network with 3-way clipping is equivalent, as an associative memory, to the original network with quenched (static) Gaussian noise on each synapse. In other words, the attractors of the network with the synaptic matrix T_{ij}, above, are the same as those of a network with the synaptic matrix:

$$J_{ij} + \eta_{ij}$$

where J_{ij} is given by Eq. 3 and η_{ij} is with the following characteristics:

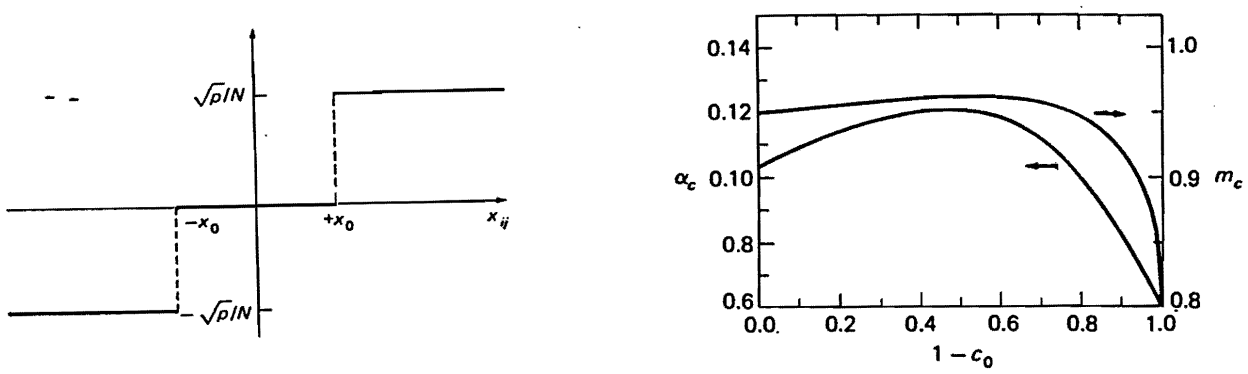

Figure 5: (a) Clipping function (b) Performance of clipped network vs the corresponding dilution. From ref. [19].

$$\langle \eta_{ij} \rangle = 0; \quad \langle (\eta_{ij})^2 \rangle = \frac{\eta^2}{N}$$

where

$$\eta^2 = \frac{p}{N}\left[\frac{1}{2}\pi c_0 e^{x_0^2} - 1\right]$$

and

$$c_0 = 1 - \mathrm{erf}\left(\frac{x_0}{\sqrt{2}}\right).$$

The quantity c_0 is the remaining connectivity in the network. That is the fraction of synapses which have not been severed by the clipping process.

Figure 5(b) shows the performance of the clipped network as a function of the clipping threshold. The curve pointing left is the storage capacity and the one pointing right is the retrieval quality, reaching unity at maximum. One sees that the properties of the network vary rather gently with the clipping and dilution. Moreover, the storage capacity even improves at the beginning and reaches a maximum at x_0=0.62 where the dilution level is about 62%, i.e. 38% of the synapses of the original network have been cut off. At this point the capacity of the network is only about 15% lower than that of the intact network.

6.2 The practical problem of continuous refresh

The above discussion 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, clipping intervenes to produce a network that can store that set of patterns and yet have only three values for its 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. Or it may be that a fully analog synapse, as would be

a condenser, may not be able to stabilize its acquired values for long periods. What is common to both cases is that the synapses must be clipped, or refreshed, all along.

One could assume that a synapse has an electronic, or biochemical refresh 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. This is one option for stabilizing a structured matrix for indefinitely long times. But 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 inaccurate upper bound. On the other hand, the simple arguments introduced here, show several ways in which things can be improved. Those can be introduced in the dynamics of data presentation, in the dynamics of the clipping, or in the coding of the data. Surprisingly, it seems that nature has implemented in cortex some of these options.

One immediate consequence of such learning is that more recently presented patterns are privileged relative to older ones, much as in 'palimpsest' scenarios[20, 21]. But palimpsest effects are due to the presence of an upper cut-off on the synaptic values, or alternatively, of the exponential decay of the contributions of older patterns[22]. 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, because they represent an very low effective cutoff on the synaptic value, and consequently the storage capacities are significantly reduced.

The basic learning scenario, to be elaborated upon below, is one in which random patterns are presented at low rate, with at most one pattern per refresh period. If learning is deterministic, 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 with the number of patterns presented per refresh period. The extreme case of fast presentation is the one in which all patterns are presented in a single refresh period, in which case one would run into a 3-way clipped net described in the previous section.

Next we discuss 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. The lower the probability the higher the capacity, which can become as high as $\sqrt{N}$, even for slow presentation of patterns. Finally, we show that if low refresh probability is combined with sparse coding in the stimuli performance can

be further improved, reaching $O((N/\ln N)^2)$ patterns.

7 Learning in material devices

I would like to devote the last part of this exposition to the type of constraints that emerge when one tries to implement a LANN, a learning attractor neural network, in a material device. Such constraints show up in a context of the development of a project of the type mentioned in the previous section. The system has two sets of dynamical variables: neurons and synapses. Neurons vary on a much shorter time scale than synapses. This feature provides the possibility of performing retrieval with quasi-static synapses. The neural variables s_i have effectively two states (high gain). They obey the dynamics:

$$s_i = \text{sign}\left(\sum_{j\neq i} J_{ij}s_j + H_i\right)$$

where J_{ij} are the synaptic variables and H_i are external currents representing the stimulus.

The synapse has two behaviours depending on the time scale:

1. Fully analog on short time scales: the growth of a given synaptic efficacy is driven by the correlated activities of the two neurons connected by the synapse. This correlation acts as input to an integrator. The potential across its capacitor stores the synaptic efficacy.

2. On a longer time scale (of the order of the capacitors decay time) intervenes the refresh mechanism which discretizes the synaptic efficacy: whenever the synaptic value crosses a threshold J_0, a current J_c is injected into the synapse, driving the capacitor towards saturation. The synaptic value is allowed to decay to zero if it is between the positive and the negative thresholds. Thus the elementary situation (see ref. [25]) would be one in which patterns are presented slowly, at most one per refresh period, and the learning is deterministic, i.e. a synapse which receives two correlated bits moves up and one which receives anti-correlated bits from its neighboring neurons it moves down. We first study this toy model for the learning dynamics.

7.1 Slow pattern presentation with deterministic refresh

In a slow deterministic learning process, of ± 1-neurons, the scenario must be as follows: when the network is forced into one of its neural states, ξ_i, every synapse will get either a $+1$ or a -1 ($=\xi_i\xi_j$). Since the presentation of the last pattern the synapses have been refreshed and the first question is whether the new stimulus will provoke any change. This depends on two things only: the preceding state of the synapse and the value of the arriving bit source. Defining by $\text{clip}(J_{ij}(\lambda))$ the clipped synaptic value following the λ-th

refresh cycle, it can be shown that the only case in which patterns can be learned is[25]:

$$\begin{cases} \text{clip}(J_{ij}(\lambda)) = +J_c & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = \ \ 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ \text{clip}(J_{ij}(\lambda)) = \ \ 0 & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = +J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \\ \\ \text{clip}(J_{ij}(\lambda)) = -J_c & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = \ \ 0 \text{ if } \xi_i^\lambda \xi_j^\lambda = +1 \\ & \rightarrow \text{clip}(J_{ij}(\lambda+1)) = -J_c \text{ if } \xi_i^\lambda \xi_j^\lambda = -1 \end{cases} \tag{9}$$

Essentially every synapse is subjected to a random walk of equal steps of both signs by the stimulus, due to the fact that we are assuming that the patterns to be learned are random and uncorrelated. This is an extremely simple instance of the more elaborate case of learning uncorrelated patterns with other types of information coding, or on synapses with general number of states, or stochastically rather than deterministically. We will discuss the more general case subsequently.

7.1.1 The basic constraint

If a sequence of p patterns $(\xi_i^\mu, \mu = 1, \ldots, p)$ is presented, then synapse J_{ij} will see the random sequence:

$$\alpha_{ij}^p, \alpha_{ij}^{p-1}, ..., \alpha_{ij}^1$$

$\alpha_{ij}^\mu = \xi_i^\mu \xi_j^\mu$ is the correlation of neurons i and j in pattern number μ. For random patterns the sequences arriving at different synapses are independently and uniformly distributed. But given the learning scenario Eq. 9, only **2** out of the possible 2^p sequences can preserve any trace of memory of the synaptic values preceding the presentation of the last p patterns.

Indeed, any sequence that contains two consecutive equal bits (e.g. $\alpha^i = \alpha^{i+1}$) erases any memory of the preceding synaptic value. Given the scenario 9, two consecutive +1 bits will drive the synapse to $+J_c$, irrespective of its preceding value, according to

$$\begin{aligned} J(i-1) = +J_c &\rightarrow J(i) = +J_c \rightarrow J(i+1) = +J_c \\ J(i-1) = \ \ 0 &\rightarrow J(i) = +J_c \rightarrow J(i+1) = +J_c \\ J(i-1) = -J_c &\rightarrow J(i) = \ \ 0 \rightarrow J(i+1) = +J_c \end{aligned}$$

Similarly, two consecutive -1's will push any synapse to $-J_c$. So the only memory preserving, p-long, sequences are:

$$\alpha_{ij}^p = +1, \alpha_{ij}^{p-1} = -1, \ldots, \alpha_{ij}^2 = -1, \alpha_{ij}^1 = +1$$

$$\alpha_{ij}^p = -1, \alpha_{ij}^{p-1} = +1, \ldots, \alpha_{ij}^2 = +1, \alpha_{ij}^1 = -1.$$

Consequently, following the presentation of p patterns, only a fraction $2/2^p$ of the totality of the N^2 synapses can maintain any memory of what preceded the sequence. When $2^{(p-1)} =$

N^2, no synapse has any memory and the maximal storage, p_c that can be achieved[25] is bounded by:

$$p_c < C \cdot \ln N. \tag{10}$$

This is a very extreme bound. It relates not to the ability of the network to retrieve what has been learned, as we have done in Sec. 3.1. It relates solely to the question of whether any trace of the first pattern has been preserved in the synapses.

7.2 General features of signal-to-noise analysis

A more meaningful tool for testing the end result of the learning process is to consider the signal-to-noise ratio in the retrieval attempt of a pattern which has been presented for learning. If that ratio is much higher than 1, retrieval will be successful. But the considerations of Sec. 3.1 were restricted to a Hopfield matrix. Here we require a more general approach. See e.g. [26].

Consider the presentation for retrieval of pattern number 1, the oldest in the sequence of p patterns that have been learned. If it has a good signal-to-noise ratio, then patterns shown subsequently will be retrievable a fortiori. When the pattern is presented the synaptic feed-back in the network creates an input 'post synaptic potential' at each neuron which is

$$h_i^p(\xi_i^1) = \sum_j J_{ij}^p \xi_j^1, \tag{11}$$

where the superscript p indicates the value of either h or J following the presentation of p patterns. The argument of h indicates the sign of the i-th bit in the pattern ξ^1 being retrieved. The signal can be defined as:

$$S^p \equiv \langle h_i^p(+1)\rangle - \langle h_i^p(-1)\rangle. \tag{12}$$

If this difference is significantly greater than the sum of the noises around each of the two values of ξ, then a threshold can be found which will separate correctly the two outcomes to reproduce the retrieved pattern.

The noise can be estimated as follows[26]: For the situation under discussion it will clearly be the same around the two values of h_i composing the signal. Hence, apart from a factor of 2, it will be

$$\begin{aligned} R^2 \equiv \langle (h_i^p - \langle h_i^p\rangle)^2\rangle &= \langle \frac{1}{N^2} \sum_{jk\neq i} J_{ij}^p J_{ik}^p \xi_j^1 \xi_k^1\rangle - \langle h_i^p\rangle^2 \\ &= \langle \frac{1}{N^2} \sum_j (J_{ij}^p)^2\rangle \leq \frac{J_c^2}{N}. \end{aligned} \tag{13}$$

Next we turn to the signal. For this purpose we consider the conditional probability that the coupling J_{ij}^p after the presentation of p patterns be J if in the first pattern we

had ξ_i^1 and ξ_j^1, i.e.

$$\Pr(J_{ij}(p) = J|\xi_i^1, \xi_j^1) \equiv \rho^p(J|\xi_i^1, \xi_j^1).$$

In terms of this probability we can write

$$\langle h_i^p(\xi_i^1)\rangle = \sum_j J_{ij}^p \xi_j^1 = \sum_{J=0,\pm J_c} \sum_{\xi=\pm 1} \Pr(\xi)\rho^p(J|\xi_i^1, \xi) J\xi$$

$$= 0.5 J_c \left(\rho^p(J_c|\xi_i^1, +1) - \rho^p(J_c|\xi_i^1, -1)\right) - 0.5 J_c \left(\rho^p(-J_c|\xi_i^1, +1) - \rho^p(-J_c|\xi_i^1, -1)\right)$$

where we have made use of the fact that the bits arrive with equal probability 0.5. Taking $\xi_i^1 = 1$ and making use of the fact that our problem has a symmetry that ensures equal probabilities for the simultaneous reversal of J and ξ^1, we find:

$$\langle h_i^p\rangle = J_c \left[\rho^p(J_c| + 1, +1) - \rho^p(J_c| + 1, -1)\right]. \tag{14}$$

Our signal S^p is twice this quantity.

Note that the first ρ on the right hand side is the fraction of all synapses which received α=+1 when the first pattern was presented and ended with a value of $+J_c$ for the synapse after all p patterns have been presented, while the second one is the fraction of those synapses that received α=−1 on the first presentation and ended having the value $+J_c$.

The probability distribution of synaptic values after p learning steps, $\rho^p(J|\tilde{\xi}, \xi)$, is subject to a master equation describing the stochastic process of learning the random patterns, i.e.

$$\rho^{p+1}(J|\tilde{\xi}, \xi) = \sum_K T_{JK} \rho^p(K|\tilde{\xi}, \xi), \tag{15}$$

where T_{JK} is the transition matrix from state K to J. As p becomes very large this probability distribution reaches an asymptotic form dominated by the largest eigen-value of the corresponding transition matrix, T, which equals 1. That limit is independent of the initial conditions and hence is devoid of all memory. But this asymptotic part cancels on the right hand side of eq. 14. The behavior of the signal for large values of p is dominated by the next leading eigen-value, λ_M. After p presentations this part of the probability distribution will be proportional to $(\lambda_M)^p$ and factors connected with the corresponding eigen-vector.

As was argued above, the dependence of the conditional probability distribution $\rho^p(J|\tilde{\xi}, \xi)$ on the first pattern decreases like $(\lambda_M)^p$. Combining this observation with the fact that the number of synapses which preserve the same conditional dependence of the synaptic value on the first pattern as ρ is $N^2 \cdot \rho$, we conclude that the number of memory carrying synapses is proportional to $N^2(\lambda_M)^p$, and from there to

$$p_c < C \cdot \frac{\ln(N^2)}{-\ln(\lambda_M)}. \tag{16}$$

This result corresponds to Eq. 10. But the context is much wider and allows for treatment of a variety of situations, as we show below.

Inserting the behavior of the leading memory preserving part of the probability distribution we have, using Eqs. 14 and 13, that the ratio of the signal-to-noise upon presentation of the oldest pattern is

$$\frac{(S^p)^2}{R^2} \approx C \cdot N(\lambda_M)^{2p}. \tag{17}$$

To allow for retrieval with a vanishing number of errors, even when the number of patterns and the number of bits in each pattern increase with N to infinity, we should have

$$\frac{(S^p)^2}{R^2} \approx C \cdot N(\lambda_M)^{2p} > \ln N.$$

See e.g. [8]. This gives for the corresponding storage capacity,

$$p_c < \frac{\ln(C \cdot N/\ln N)}{-2\ln(\lambda_M)}. \tag{18}$$

Note that in general the constant C, independent of p and N, may depend on various parameters of the learning process, such as the coding level of the presented patterns f and/or the synaptic transition probability q in a stochastic learning scenario, and/or on the rate of pattern presentation, etc..

7.3 Deterministic slow presentation of ± 1 patterns

In this case the transition matrix is:

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The elements $\frac{1}{2}$ represent the probability that at any step the synapse get a +1 to go up, or a −1 to go down, in the 3-state ladder of possible synaptic stable states. This matrix has the eigen-values 1, $\frac{1}{2}, -\frac{1}{2}$. Its eigen-vectors are independent of the parameters of the problem, i.e. p, N etc.

In the present case $\lambda_M = \frac{1}{2}$. Hence the non-asymptotic part of the probability distribution vanishes as 2^{-p}. Combining this observation with the fact that the number of synapses which express the same conditional dependence (of the synaptic value on the first pattern) as ρ does, we arrive immediately at the result of the previous section, i.e. that the number of memory carrying synapses is proportional to $N^2 2^{-p}$, and from there to 10, i.e. $p_c < \ln(CN^2)$.

Inserting the behavior of the leading memory preserving part of the probability distribution we have, using Eqs. 14 and 13, that the ratio of the signal-to-noise upon presentation of the oldest pattern is

$$\frac{(S^p)^2}{R^2} > C \cdot N 2^{-2p}. \tag{19}$$

For this ratio to be larger than $\ln N$, see discussion above, we must have

$$p_c < \frac{1}{2}\ln(CN/\ln N).$$

7.4 Logarithmic limitation for multi-state synapses

The severe logarithmic restriction on the number of patterns that can be learned in a slow presentation scenario, exposed in the previous section, may appear to be a consequence of the small number of stable states available to each synapse. We now proceed to show that this constraint persists for networks with synapses with arbitrarily many stable states, provided the number of stable synaptic states is kept fixed as the number of neurons N or the number of patterns p grows.

Consider a network with synapses that have n possible stable states. Let the neurons be ± 1 again and hence the synaptic source is always either $+1$ or -1. The refresh will be taken deterministic and the presentation of patterns to be learned slow, one pattern per refresh cycle. Upon presentation of a pattern a synapse which is not in one of its extreme states will go one step up (down) with probability $\frac{1}{2}$, depending on the sign of the arriving source bit α. Synapses in the two extreme states will remain unchanged with probability $\frac{1}{2}$ or move one step away. Again, the process of presentation of random patterns is a random walk on the synaptic values, described by a probability distribution $\rho^p(m|\tilde{\xi},\xi), m = 0,1,\ldots,n-1$, where m stands for the synaptic state J_m and $\tilde{\xi},\xi$ for the source bits at the first presentation. $\rho^p(m|\tilde{\xi},\xi)$ is an n dimensional vector for fixed $\alpha = \tilde{\xi}\cdot\xi$. The transition matrix, describing the dynamics of ρ is a simple extension of T of the previous section. It is tri-diagonal with $\frac{1}{2}$'s along the two side diagonals; zeros along the main diagonal except in the first and last positions where it is $\frac{1}{2}$.

The signal is:

$$\begin{aligned} S^p &= \langle h^p(+1)\rangle - \langle h^p(-1)\rangle \\ &= \frac{1}{2}\sum_{m=0}^{n-1} J_m\left[\rho^p(m|1,1) - \rho^p(m|1,-1)\right] - \frac{1}{2}\sum_{m=0}^{n-1} J_m\left[\rho^p(m|-1,1) - \rho^p(m|-1,-1)\right] \\ &= \sum_{m=0}^{n-1} J_m\left[\rho^p(m|1,1) - \rho^p(m|1,-1)\right], \end{aligned} \tag{20}$$

where the 2 coming from the symmetry of ρ under the change of sign of both ξ's cancels against the $\frac{1}{2}$ which is the probability of getting a ± 1. In the difference of distribution functions on the r.h.s the asymptotic part cancels out and the leading term, expressed, in terms of the distribution following the presentation of the oldest pattern, ρ^1, is:

$$S^p \approx (\lambda_M)^{p-1}\sum_{K=0}^{n-1}\sum_{m=0}^{n-1} J_m v_K v_m[\rho^1(K|1,1) - \rho^1(K|1,-1)], \tag{21}$$

where v_K are the components of the eigen-vector of the matrix T corresponding to the eigen-value λ_M.

This expression becomes much simplified when the initial distribution is the asymptotic one, since the latter is uniform, i.e. $\rho^0 = \frac{1}{n}$. Upon presentation of the first pattern, the difference between the two conditional probabilities can only appear at the extreme values of J_m. $\alpha^1 = +1$ pushes all synaptic values one step up, leaving the probability of all states invariant, except the lowest state which is emptied and the top one which is doubled, it becomes $\frac{2}{n}$. $\alpha^1 = -1$ does the opposite, and thus

$$\rho^1(K|1,1) - \rho^1(K|1,-1) = \left(\frac{2}{n}, 0, \ldots, 0, -\frac{2}{n}\right).$$

Inserting in Eq. 21 we have,

$$S^p \approx (\lambda_M)^{p-1} \frac{2}{n} \sum_{m=0}^{n-1} J_m v_m [v_0 - v_{n-1}]. \tag{22}$$

Next we use the fact that[26]

$$\lambda_M = 1 - 2\sin^2 \frac{\pi}{2n}$$

and the corresponding eigen-vector has

$$v_0 = -v_{n-1} \approx \sqrt{\frac{2}{n}}, \quad v_m = v_0 \cos\left(\frac{\pi m}{n-1}\right) \tag{23}$$

when n is large. Substituting in 22 one finds:

$$S^p \approx (\lambda_M)^{p-1} \frac{8}{n^2} \sum_{m=0}^{n-1} J_m \cos\left(\frac{\pi m}{n-1}\right). \tag{24}$$

One immediate conclusion from this result is that the signal will decrease with increasing n, at least as n^{-1}. This result is independent of the choice one makes for the synaptic values. It is a consequence of the fact that the learning process has a uniform asymptotic distribution of values. Even after the presentation of a single pattern, on the background of the asymptotic distribution, the signal will decrease as n^{-1}.

As far as the noise is concerned, it is contributed mainly by the asymptotic part of the distribution. Hence:

$$\langle R^2 \rangle = \frac{1}{N} \sum_m \rho(m|1\xi)(J_m)^2 \approx \frac{1}{n} \sum_m (J_m)^2$$

The square noise is therefore independent of n, since the sum will be proportional to n and to the square of the scale of the J_m. Since also the square of the signal is proportional to the square of the scale of the J's, we find that the signal-to-noise ratio is decreased by a factor of $\frac{1}{n}$ relative to the 3-state case. But there is an important change in the eigen-value. It approaches 1 like n^{-2} when n becomes large. Hence the final signal-to-noise ratio behaves like

$$\frac{(S^p)^2}{R^2} \approx \frac{N}{n^2} \left(1 - \frac{a}{n^2}\right)^{2p},$$

which leads, in turn to

$$p_c \approx \frac{\ln(\frac{N}{n^2})}{2\ln(1-\frac{a}{n^2})} \approx n^2 \ln\left(\frac{N}{n^2}\right).$$

This is an improvement of a factor of n^2. Note that this factor can approach N if n is allowed to approach $N^{1/2}$. In fact, this is the upper allowed limit for n since for larger values of n the logarithm in the numerator becomes negative, which is a reflection of the fact that even a single pattern cannot be recalled in these conditions.

8 Going beyond the logarithmic constraint

8.1 Elements of leverage

The logarithmic constraint on the number of storable patterns in a network with synapses with a fixed number of internal states is severe enough to warrant a special attention as to the determining factors. The above discussion points to the following ones :

- The speed of the pattern presentation;
- The generic distribution of synaptic values upon presentation of a pattern;
- The fact that in a learning step all synapses are affected in an identical way. This is consists of two components:
  - The coding type in which we have ± 1 with equal probability;
  - The deterministic action of the refresh mechanism.

8.2 Stochastic learning of sparsely coded patterns

In ref. [26] we have discussed the effect of fast pattern presentation. Clearly, if as many as $O(N)$ patterns are presented in a single refresh period, and at the beginning the synapses were all zero, then one can reproduce the Sompolinsky case, that is memorize $0.12N$ patterns[9, 19, 6]. With its good news and its bad news: capacity is large but one invites the blackout catastrophe, i.e. that if by chance too many patterns were presented in a refresh period all is lost. Moreover, patterns which are presented slowly act to reduce the capacity of those that have been learned fast. Finally, in a network of a realistic size it is rather unrealistic to have $O(N)$ patterns presented in a single refresh period.

If one is more modest, allowing the speeding up of presentation up to a rhythm of a fixed finite number per refresh period, an improvement of the capacity is achieved: the logarithmic constraint remains but is multiplied by the number of patterns presented in a single refresh period[26].

The type of remedies one should rather look for are those which improve the worst case – slow learning, and consequently will not require special conditions on the input stream. Next we can take up the question of the generic synaptic distribution upon the

presentation of a new pattern. If the number of synaptic states is finite, the stochastic process will typically be ergodic and the asymptotic distribution, following the presentation of a large number of patterns will be independent of the initial patterns and will be uniform over the synaptic states. This is also the way that 'palimpsest' behavior is brought about. As patterns move into the past they are washed out into the uniform distribution and new patterns are inscribed on top of them. There is no blackout catastrophe. This was effectively what was observed in the studies of refs. [22, 21] of memories that forget.

Thus we consider a network with 2-state synapses. Recall that in order to learn random ± 1 patterns with a deterministic refresh mechanism we had to have at least three states. With two states every pattern would have determined all the synapses, thus erasing all previous memory. Its asymptotic distribution will be uniform. The options left open are the last two in the list above, namely the coding type and the transition probability in the refresh mechanism. The first option is to code the pattern by 0-1 states for the neurons, with states of coding rate f, where f is the mean fraction of 1's in a pattern. In this case, for f small, upon presentation of a pattern, a synapse will see a +1,+1 pair source with probability f^2, a 0,+1 pair with probability $2f(1-f)$ and 0,0 with probability $(1-f)^2$. The Hebbian dynamics implies that a small fraction f^2 of the synapses is to move up, a somewhat larger fraction, those that see anti-correlated neural activities, 0,+1, should move down. And the majority, $(1-f)^2$ of them, remains unchanged.

The second option is to postulate that the synapses which are expected to change, either up or down will not do so deterministically, but rather stochastically, with corresponding probabilities q_+ and q_-. The probability distribution of synaptic values has two components and the transition matrix T will be a 2×2 matrix. It is:

$$T = \begin{pmatrix} 1-2f(1-f)q_- & f^2 q_+ \\ 2f(1-f)q_- & 1-f^2 q_+ \end{pmatrix} \tag{25}$$

This matrix has the eigenvalues 1 and

$$\lambda_M = 1 - f^2 q_+ - 2f(1-f)q_-.$$

One can compute the left and right eigen-vectors and hence all the statistical properties of the network. In particular, the asymptotic distribution is

$$\rho_{as}(J) = \left(\frac{b}{a+b}, \frac{a}{a+b}\right) \equiv (p_+, p_-),$$

where $a = 2f(1-f)q_-$, $b = f^2 q_+$. The two components of ρ_{as} are the fractions of synapses with values J_+ and J_-, respectively.

The signal can be written as:

$$\begin{aligned} S^p &= \langle h^p(1)\rangle - \langle h^p(0)\rangle = f\sum_J J[\rho^p(J|11) - \rho^p(J|01)] \\ &= f(\lambda_M)^{p-1}\sum_{JJ'} J A_J B_{J'}[\rho^1(J'|11) - \rho^1(J'|01)] \end{aligned} \tag{26}$$

in which A_J and B_J are, respectively, components of the right and left eigen-vectors corresponding to the eigenvalue λ_M and ρ^1 is the distribution function resulting from thr asymptotic distribution upon presentation of a single pattern. The solutions for the eigen-vectors are $A = 1, -1$ and $B = (\frac{a}{a+b}, \frac{b}{a+b})$.

$$\rho^1(J_+|11) = p_+ + p_- q_+; \quad \rho^1(J_+|01) = p_+(1 - q_-).$$

Hence,

$$\rho^1(J_+|11) - \rho^1(J_+|01) = \frac{aq_+ + bq_-}{a+b} = p_- q_+ + p_+ q_-,$$

and because $\rho(J_-) = 1 - \rho(J_+)$, replacing in the above difference J_+ by J_- simply reverses the sign. Finally,

$$S^p \approx f(\lambda_M)^{p-1}(J_+ - J_-)\frac{aq_+ + bq_-}{a+b}. \tag{27}$$

For the noise we calculate the asymptotic part. Namely,

$$\begin{aligned} R^2 &= \langle \frac{1}{N^2}\sum_{jk} J_{ij}J_{ik}\eta_j\eta_k\rangle - \langle\left(\frac{1}{N}\sum_j J_{ij}\eta_j\right)^2\rangle = \frac{1}{N^2}\left(\sum_j\langle(J_{ij})^2\eta_j\rangle - \sum_j\langle J_{ij}\eta_j\rangle^2\right) \\ &\to \frac{f}{N}\left[\sum_J J^2\rho_{as}(J) - \left(\sum_J J\rho_{as}(J)\right)^2\right] = \frac{f}{N}\left(p_+J_+^2 + p_-J_-^2 - f(p_+J_+ + p_-J_-)^2\right) \end{aligned}$$

Combining the two and substituting $a = 2f(1-f)q_-$, $b = f^2q_+$, and neglecting the last term (proportional to f^2) on the right hand side one has for the signal-to-noise ratio:

$$\frac{(S^p)^2}{R^2} \approx Nf(\lambda_M)^{2(p-1)}\frac{(q_+q_-)^2(2-f)^2}{[2(1-f)q_- + fq_+]}\frac{(1 - J_-/J_+)^2}{[fq_+ + 2(1-f)q_-J_-^2/J_+^2]} \tag{28}$$

Auto-associative retrieval requires that $(S^p)^2/R^2 > \ln N$.

8.3 Performance: absolute maximal storage in a palimpsest

1. Low coding rate, finite transition probabilities and non-zero synaptic values, i.e.

$$q_+, q_-, J_-, J_+ = O(1); \quad f \approx \ln N/N \to 0.$$

In this case:

$$\frac{(S^p)^2}{R^2} \approx Nf(\lambda_M)^{2(p-1)}q_+^2(J_+/J_- - 1)^2.$$

And the capacity is limited by:

$$p_c \approx \frac{\ln(Nf/\ln N)}{-2\ln(1-2fq_-)} < \frac{N}{4q_-\ln N}. \tag{29}$$

In other words, in this case as in the following one, going to very low coding rate one looses a factor of N in capacity relative to the maximal capacity of the Willshaw model[11], where for coding rate $f = N/\ln N$ the network can store $N^2/(\ln N)^2$ patterns.

2. The most interesting case arises when one imposes that the lowering transition probability be proportional to the (low) coding rate, i.e. $q_+ = O(1)$, $q_- = \overline{q_-}f$. In this case:
$$\frac{(S^p)^2}{R^2} \approx C \cdot Nf(\lambda_M)^{2(p-1)},$$
where C is independent of f and of N and is $O(1)$. The factor fN in this expression implies that $f > N^{-1}$. But the important change is in the eigenvalue λ_M which becomes: $\lambda_M \approx 1 - Qf^2$. The result is that
$$p_c \approx \frac{\ln(Nf/\ln N)}{-2\ln(1-Qf^2)}, \tag{30}$$
in which Q is a constant depending only on q_+ and $\overline{q_-}$. Now when $f \approx \ln N/N$, the capacity reaches its ultimate value $N^2/(\ln N)^2$.

There is though an important advantage, namely that the learning scheme presented here can function indefinitely, while the Willshaw scheme, as well as the Hopfield one break down after a long enough process of learning different patterns. Both undergo a blackout catastrophe: if learning had been excessive nothing can be retrieved.

Acknowledgements

Most of what is new in this text has been developed in collaboration with Stefano Fusi to whom I am also indebted for many useful comments on the manuscript.

References

[1] Amit DJ 1992 Psychology, Neuro-biology and Modeling: The Science of Hebbian Reverberations,

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

[3] Sakai K and Miyashita Y 1991 Neural organization for the long-term memory of paired associates, *Nature*, **354** 152.

[4] Miyashita Y 1988 Neuronal correlate of visual associative long-term memory in the primate temporal cortex, *Nature* **335** 817

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

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

[7] Amit DJ, Gutfreund H and Sompolinsky H 1985 Spin-glass models of neural networks, *Phys. Rev.*, **A32**, 1007

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

[9] Amit DJ, Gutfreund H and Sompolinsky H 1987 Statistical mechanics of neural networks near saturation, *Annals of Physics*, **173**, 30

[10] Gardner E 1987 Maximum storage capacity in neural networks, *Europhys. Lett.*, **4**, 481

[11] Willshaw D 1969 Non-holographic associative memory, Nature, London, **222** 960

[12] Tsodyks MV and Feigel'man MV 1988, The enhanced storage capacity in neural networks with low activity level, *Europhys. Lett.*, **46** 101

[13] Buhmann J, Divko R and Schulten K 1989, Associative memory with high information content, *Phys. Rev.* **A39** 2689

[14] Griniasty M, Tsodyks MV and Amit DJ 1992 Conversion off temporal correlations between stimuli to Spatial correlations between attractors, *Neural Computation* **5** 1

[15] Cugliandolo L 1992 Correlated attractors from uncorrelated stimuli in *Les Houches Meeting on Mathematical Methods Motivated by Neural Networks and Learning Processes*

[16] 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) and Badoni D, Riccardi R and Salina G 1992 Learning Attractor Neural Network: The Electronic Implementation, Preprint ROM2F/92/56, Dipartimento di Fisica, Universita' di Roma *Tor Vergata* (Roma, November 1992); to appear in the Proceedings of the Conference *Neural Network: From Biology to High Energy Physics* Proceedings of the second workshop held in Elba International Physics Center, June 1992

[17] Amit DJ and Fusi S 1992 Learning constraints in storage capacity in networks with dynamic synapses, to appear in *Neural networks: from biology to high energy physics* Proceedings of the second workshop held in Elba International Physics Center, June 1992

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

[19] 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)

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

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

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

[23] Tsodyks M 1990 Associative memory in neural networks with binary synapses, *Modern Physics Letters* **B4** 713

[24] Horner H 1992, private communication.

[25] Amit DJ and Fusi S 1992 Constraints on learning in dynamic synapses, *NETWORK* in press.

[26] Amit DJ Fusi S 1993 Theory of unsupervised attractor learning in synapses with fixed number of stable states, in preparation.