PREPRINT N.856
17 Gennaio 1992

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

ROME-856-1992

Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors

M. Griniasty and M.V. Tsodyks[1]
Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem
and
Daniel J. Amit
INFN, Sezione di Roma, Istituto di Fisica
Universita di Roma, La Sapienza, Ple Aldo Moro, Roma[2]

Fermilab Library
0 1160 0007388 4

Abstract

It is shown that a simple modification of synaptic structures (of the Hopfield type) constructed to produce auto-associative attractors, produces neural networks whose attractors are correlated with several (learned) patterns used in the construction of the matrix. The modification stores in the matrix a fixed sequence of uncorrelated patterns. The network then has correlated attractors, provoked by the uncorrelated stimuli. Thus, the network converts the temporal order (or temporal correlation) expressed by the sequence of patterns, into spatial correlations expressed in the distributions of neural activities in attractors.

The model captures phenomena observed in single electrode recordings in performing monkeys by Miyashita et al. The correspondence is as close as to reproduce the fact that, given uncorrelated patterns as sequentially learned stimuli, the attractors produced are significantly correlated up to a separation of 5 (five) in the sequence. This number 5 is universal in a range of parameters, and requires essentially no tuning. We then discuss learning scenarios which could lead to this synaptic structure as well as experimental predictions following from it. Finally, we speculate on the cognitive utility of such an arrangement.

ROUTE TO
NAME
LOCATION
FERMI LIBRARY
MAR 30 1992

[1]On leave of absence from The Institute of Higher Nervous Activity, Moscow
[2]On leave of absence from Racah Institute of Physics

1 Introduction

Temporal to spatial correlations in monkey cortex

The remarkable sequence of neuro-cognitive experiments by Miyashita et al[1, 2, 3] is the most direct evidence of the relevance of attractor dynamics in cortical cognitive processing. It is at the same time detailed and structured enough to guide and confront attractor neural network (ANN) modeling. In the first experiment[1], the monkey is trained to recognize and match a set of visual patterns. As a result, one observes selective enhancement of neural spike activity, which persists for 16 seconds after the removal of the stimulus. The fact that selective, stimulus related, enhancement of neural activity persists for 16 seconds in the absence of the provoking stimulus, is incontrovertible evidence of non-ergodic attractor dynamics. The same encouraging evidence has forced a confrontation on the question of activity rates upon retrieval by attractors. The rates in the Miyashita attractors was many times lower than what models of the Hopfield type[4, 5] predicted. This fruitful confrontation led to a study[6] which showed that when neural description is taken in greater detail, as well as the conditions prevailing in cortex, attractors can appear having stochastic behavior and low rates.

The second study[2] went further to provide information about coding in the particular module of the anterior ventral temporal cortex of the monkey: It was discovered that, despite extreme precaution in producing visual stimuli uncorrelated in their spatial form, spatial correlations appeared in the patterns of sustained activities, evoked by the stimuli, during the delay period. These persistent activities we interpret as the structure of the attractors. There was one kind of correlation that was preserved in the stimuli, the temporal order of their presentation was maintained fixed during training. What the monkey's brain appears to be doing, is to convert the temporal correlation into a spatial one. Namely, spatial correlations were observed among the attractors, corresponding to the stimuli that were close temporally in the training session. These attractors are the result of retrieval dynamics. The spatial correlations between the activities of the neurons investigated persisted to a fifth neighbor in the temporal sequence. The correlation figure of ref. [2] is reproduced in Fig. 1.

Modeling correlation conversion

The main result of the Hopfield program has been to connect the intuitive call for selective (stimulus dependent) attractor dynamics (associative memory) with specific constructions of synaptic matrices, and therefore a bridge to unsupervised learning. The program was limited by the requirement that the attractors be as close as possible to the patterns of which the matrices were constructed, i.e. the presumed items in the learning process. This went under the name of auto-association.

Here we shall show that a simple modification of the synaptic matrices used for auto-association in ANN's leads to a relaxation dynamics that associates with stimuli near one of the random, uncorrelated underlying patterns, an attractor which is

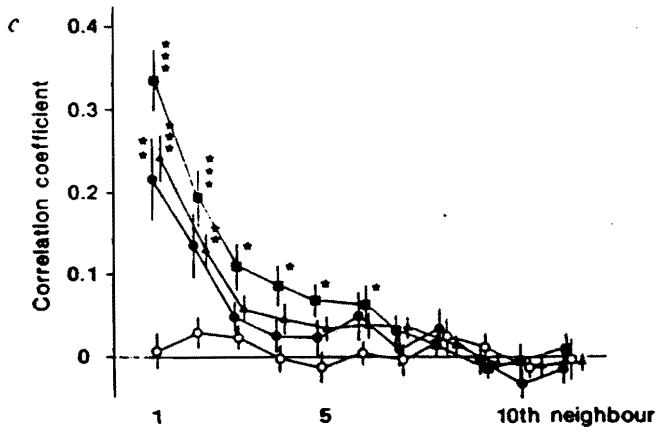

Figure 1: Spatial correlations between attractors, in monkey's anterior ventral temporal cortex, corresponding to structurally uncorrelated patterns, as a function of the difference in the position of the learned stimuli in the fixed training sequence. From [2]

correlated with several patterns. The patterns which have the largest correlations with a given attractor are the neighbors of the stimulus leading to the attractor, in the sequence of stored patterns. It then follows that attractors are correlated among themselves. Again, the attractors that are correlated are the neighbors in the sequence of the underlying patterns. These are just the type of correlations observed by Miyashita. In fact, the number of attractors that are found to be correlated significantly in the model, is the same as in the experiment.

The extended model is discussed in two different variants: one is the original formulation of ± 1 neurons, with the artificial symmetry between active and passive states of neurons; the second is a 0–1 formulation[7, 8], in which this symmetry is removed and which can naturally be interpreted in term of high and low activity rates of neurons in attractors. The results differ in detail, but the main qualitative features, of converting sequential order among uncorrelated patterns to a set of correlated attractors, is present in both. Both models have symmetric synaptic matrices, which are unrealistic but convenient. The study of auto-associative ANNs, over the last several years, has made it clear that most of the attractor properties of these extensively connected networks are rather robust to the introduction of synaptic asymmetry (see e.g. [5]).

We then proceed to interpret the proposed synaptic matrices in terms of learning dynamics. It is argued that rather plausible synaptic dynamics, accompanying the relaxation in the ANN, may produce a synaptic matrix with correlated attractors for uncorrelated external stimuli. Within such learning scenarios, one is led to predict that the presentation of uncorrelated patterns in a random sequence, would produce

attractors which are uncorrelated, and are each close to the representation of the original patterns, as would be the case in the Hopfield model.

Finally, we discuss the potential utility of such conversions of temporal correlations to spatial correlations in modeling several aspects of cognitive behavior.

2 The model with ± 1 neurons

The original way of pursuing the Hopfield ANN program was to choose the variables describing the instantaneous states of each neuron $S_i(t)=\pm 1$, where i labels the neuron, $(i = 1, \ldots, N)$. The patterns, to be stored in an N neuron network, are N-bit words of ± 1's, the value of each bit chosen independently, with probability 0.5. Denoting the components of the activity of neuron number i in pattern number μ, ξ_i^μ, the proposed synaptic matrix is written as:

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

where p is the total number of patterns stored in the connections. The patterns, μ are considered to form an ordered sequence, which corresponds to the order of presentation in the training phase. For simplicity, the sequence is taken to be cyclic. Each pattern in the construction of the matrix is connected to one preceeding and one succeeding pattern. Note, in particular, that this extended matrix still preserves the symmetry of the Hopfield matrix, which implies that all attractors will be fixed points, and makes analysis so much simpler. How this relates to a learning scenario, is discussed in Section 4. The matrix of Eq. 1, for a=0, reduces to the original Hopfield matrix.

This matrix is accompanied, as usual, by a schematized spike emission dynamics which, in the noiseless case, determines the new state, $S_i(t)$, of the neuron according to:

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

where

$$h_i(t + \delta t) = \sum_{j=1}^{N} J_{ij} S_j(t). \tag{3}$$

h_i mimics the value of the post-synaptic potential, relative to the threshold, on neuron i. The linear superposition of bilinear terms in the neural activities of the stored patterns, is sometimes referred to as a Hebbian learning from a 'tabula rasa'. We shall return to the question of learning later.

The natural variables for the description of the non-ergodic asymptotic behavior of the network are the 'overlaps' $m^\mu(t)$ of the current state of the network, $S_i(t)$, with the stored pattern μ. They are defined as:

$$m^\mu = \frac{1}{N} \sum_i S_i \xi_i^\mu \tag{4}$$

See e.g. ref. [9, 5]. The value of overlap m^μ measures how close is the state of the network to the stored pattern μ. If m^μ=1, the state is identical, as a binary word to the pattern μ, i.e. $S_i = \xi_i^\mu$ for all i. If m^μ=1 for the asymptotic state, the attractor, then the corresponding pattern is retrieved perfectly.

With the matrix Eq. 1, the 'field' h_i can be expressed in terms of the overlaps[9], which implies that so can the dynamics of the network as well as its attractors. Namely, we can write Eq. 3 as

$$h_i(t) = \sum_\nu m^\nu(t)[\xi_i^\nu + a(\xi_i^{\nu+1} + \xi_i^{\nu-1})], \tag{5}$$

from which one derives the mean-field equations determining the attractors. For a symmetric matrix, those are simple fixed points. They read, in the limit of a network of a large number of neurons with a relatively low number of stored patterns:

$$m^\mu = \langle\langle \xi^\mu \text{sign}\left(\sum_\nu m^\nu[\xi^\nu + a(\xi^{\nu+1} + \xi^{\nu-1})]\right)\rangle\rangle. \tag{6}$$

The double angular brackets imply an averaging over the distribution of the bits in the patterns (see e.g. [9, 5]).

Auto-association was the interpretation of the fact that, in the absence of noise, for low loading, the equations 6 had solutions with one single $m^\mu \neq 0$, which attracted a wide set of initial states in the neighborhood of each pattern. Away from these large basins, 'spurious states' were found to exist[9]. Moreover, the artificial symmetry of the +1 and −1 states, produced attractors of the sign reversed states of each pattern. These, retrieval properties of the Hopfield ANN have been found very robust to extensive noise and synaptic disruption, including asymmetric disruption.

If one tries a pure pattern solution for Eqs. 6, with μ=2 for example, one has[3]:

$$m^2 = \langle\langle \xi^2 \text{sign}\left(m^2[\xi^2 + a(\xi^1 + \xi^3)]\right)\rangle\rangle. \tag{7}$$

For a <0.5, it is the first term in the square brackets which dominates the sign of the argument of the sign-function and m^2=1 is a fixed point solution, as in the case a=0. For a >0.5, this is no longer the case. For 25% of the sites $\xi^3 = \xi^1 = -\xi^2$ and the argument has the sign opposite to that of ξ^2. Starting from a state with m^2=1, and all other overlaps 0, one arrives, after one step to a state with $m^2 = m^1 = m^3$=0.5. This is no fixed point either.

The solutions of Eqs. 6 have all overlaps different from zero. The previous discussion suggests a numerical procedure for arriving at the solution: start from a pure pattern state and iterate until convergence. This is what the network would do, if given one of the pure patterns it learned, $S_i = \xi^\mu$, as an initial state, until it relaxes to a fixed point. The symmetry of the dynamics under pattern permutations implies that this has to be done for one pattern only.

The equations were solved in this way. One finds that starting from a pure pattern, one arrives at a stable solution after several iterations. The solution reached is a state

[3]The 2's are superscripts not squares.

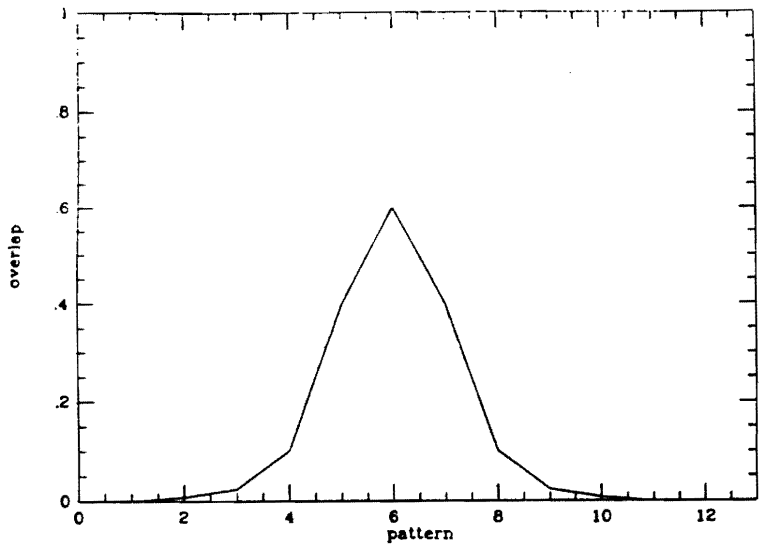

Figure 2: Overlaps between an attractors and stored patterns, as a function of the separation of the pattern in the sequence from the pattern underlying the attractor.

with nonzero overlaps with all stored patterns, symmetrically distributed around the pattern, which served as the stimulus. Only a small number, actually 5, of these overlaps are significantly large, provided $a > 1$. This distribution of overlaps in an attractor, corresponding to one of the underlying patterns, is shown in the Figure 2. In this case p=13 patterns are stored, and a=0.7.

It is remarkable that the structure of the attractor does not depend on the number of patterns p, nor on the value of a, in the entire range $0.5 < a < 1$. For $a > 1$, the network develops attractors which have overlaps with all stored patterns. The values of the overlaps decrease as the number of patterns increases. This means that after learning sufficiently many patterns, the network loses it's ability to associate attractors with the stimuli.

One can read from Figure 2 that the retrieval attractor has substantial overlaps with several patterns, symmetrically disposed, before and after, in the sequence relative to the pattern corresponding to the stimulus. Clearly, if each attractor is correlated with several patterns, then the attractors corresponding to different patterns must themselves be correlated. These correlations would correspond to the correlations measured by Miyashita[1], Figure 1. The correlation of activities in two attractors, σ^μ and σ^ν, is defined as:

$$C(\mu,\nu) = \frac{1}{|C|}\sum_{i=1}^{N}(\sigma_i^\mu - \overline{\sigma})(\sigma_i^\nu - \overline{\sigma}), \tag{8}$$

where $\overline{\sigma}$ is the average activity in a given attractor and the normalization constant $|C|$ is chosen so that $C(\mu,\mu)$=1. In the present case, $\overline{\sigma}$=0, and $|C| = N$. Hence, the

correlation of attractors μ and ν can be written as:

$$\begin{aligned} C(\mu,\nu) &= \frac{1}{N}\sum_i \sigma_i^\mu \sigma_i^\nu \\ &= \frac{1}{N}\sum_i \mathrm{sign}(h_i^\mu)\mathrm{sign}(h_i^\nu) = \langle\langle \mathrm{sign}(h^\mu)\mathrm{sign}(h^\nu)\rangle\rangle, \end{aligned} \tag{9}$$

where h_i^μ is the local field on neuron i when the network is in attractor number μ. The last equality is an expression of self-averaging, giving an average over the distribution of patterns. Finally, substituting the fields from Eq. 5 in Eq. 9, we arrive at the correlation coefficient:

$$C(\mu,\nu) = \langle\langle \mathrm{sign}(\sum_\alpha \xi^\alpha(m_\mu^\alpha + a m_\mu^{\alpha+1} + a m_\mu^{\alpha-1})\mathrm{sign}(\sum_\beta \xi^\beta(m_\nu^\beta + a m_\nu^{\beta+1} + a m_\nu^{\beta-1})\rangle\rangle \tag{10}$$

where m_μ^ρ is the overlap of the attractor corresponding to stimulus number μ with pattern number ρ.

These attractor overlaps are illustrated in Figure 3, where we plot the correlations among different pairs of attractors vs the distance between their corresponding patterns. Figure 3 clearly demonstrates that while stored patterns are completely random, and hence uncorrelated, the states reached by the network upon presentation of these same patterns, have a substantial degree of correlations, which decreases with the separation of the patterns in the sequence. Note that while an attractor 'sees' two–three patterns on each side, it sees five attractors on each side. The qualitative form of the correlations captures the experimental trend, Figure 1. The absolute values of the correlations differ. This can be due to a different normalization used in ref. [1], where the normalization is not given explicitly. It may also be that the absolute values will be different in a more realistic networks.

Finally, the analysis given above is based on the exact symmetry of the matrix 1, in which case the system has only fixed point attractors. Asymmetry can enter in two ways: either as a local, random disruption of the synaptic elements, or as a coherent asymmetry of the two transition terms in the symmetric matrix. The first type of asymmetry is in the realm of the robustness of the attractor dynamics of the network. Concerning the second type: The mean-field equations 6 hold even if the coefficients of the two transition terms in J_{ij} are not equal. We have found that the behaviour of the network is robust against some amount of asymmetry between the two coefficients. However, if the asymmetry becomes two large, Eqs. 6 first acquire another solution, with the maximum overlap shifted to another pattern. At still higher asymmetry, the fixed point solution is lost, in favor of a time-dependent attractor. In this attractor the network moves from one pattern to another, in a direction determined by the major non-diagonal term in the matrix 1.

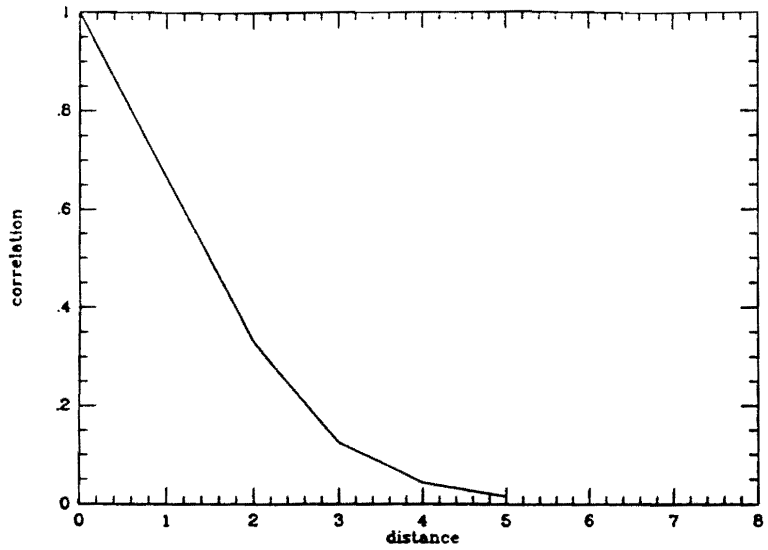

Figure 3: Correlations between attractors as a function of the separation in the sequence of the patterns to which the attractors belong.

3 ANN with discrete 0–1 neurons

The above description was extended to deal with a 0–1 representation of information, allowing for the removal of the symmetry between active and refractory states of the neurons[7, 8]. This description has several further advantages: its terms are very close to a representation in terms of spike rates, which are positive analog variables; moreover it allows for very efficient storage of patterns as the coding rate, i.e. the fraction of 1's in the patterns, becomes very low – sparse coding. Since it is important to show that the correlation effects to be discussed take place also for this case, we shall recall that formulation as well.

The dynamics is described in terms of instantaneous neural variables, $V_i(t)$, which take the values (0,1) as

$$V_i(t+\delta t) = \Theta\left(\sum_j J_{ij} V_j(t) - \theta\right), \tag{11}$$

where $\Theta(x)$=1 for x >0, and 0 otherwise, and θ is a neural threshold. These variables can be directly interpreted as high and low activity (spike rates) of each neuron. Such would be a description in terms of analog rates, in a high gain limit.

In this case the patterns to be stored by a learning dynamics, η^μ, are chosen as N-bit words of independently chosen 0,1's, i.e.

$$\eta_i^\mu = 0;1 \quad \mu = 1, ..., p \tag{12}$$

where the probability for a 1,0-bit is $f, (1-f)$, respectively.

An extension of the symmetric synaptic matrix, appropriate for auto-association [7, 8], to our requirements would be:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\left((\eta_i^\mu - f)(\eta_j^\mu - f) + a[(\eta_i^{\mu+1} - f)(\eta_j^\mu - f) + (\eta_i^\mu - f)(\eta_j^{\mu+1} - f)]\right), \quad (13)$$

and the corresponding overlaps are generalized to:

$$m^\mu(t) = \frac{1}{Nf(1-f)}\sum_i(\eta_i^\mu - f)V_i(t). \quad (14)$$

With the couplings 13, the dynamics can be expressed in terms of the above overlaps and so can the fixed points of the retrieval attractors. The latter have the form:

$$m^\mu = \langle\langle(\eta^\mu - f)\Theta\left(\sum_\nu m^\nu\{(\eta^\nu - f) + a[(\eta^{\nu-1} - f) + (\eta^{\nu+1} - f)]\}\right)\rangle\rangle. \quad (15)$$

When a=0, the system of equations reduces to that of refs. [7, 8]. In this case, at low loading, the exact stored patterns are the retrieval attractors of the network, i.e. the equations admit solutions with a single non-vanishing overlap, which in turn is equal 1. These attractors persist until a reaches the critical value of,

$$a_{cr} = \frac{f+\theta}{2(1-f)}.$$

Above this value of a, the pure patterns are unstable and the network, having a symmetric synaptic matrix, finds new fixed points. The Eqs. 15 have to be solved numerically for the values of the overlaps in the retrieval attractors. This we do, again following the network, as was explained in the previous section.

A typical solution is shown in Figure 4, for parameter values: f=0.01, θ=0.2, p=11 and a=0.25. Figure 4(a) represents the overlaps vs the pattern number in the sequence, relative to the pattern of the stimulus. Figure 4(b) is the correlation between the attractors. Note that in distinction to the ± 1 case, the significant overlaps here, of which there are five in total, are all equal. They are all unity, up to terms of $O(f)$. This implies that the attractor is approximately the union of the 1-bits in the five patterns centered around the stimulus. In particular, the mean spatial activity level in the attractors is *higher* than in the pure patterns. A fact that can be tested experimentally.

The correlation figure 4(b) may seem somewhat simple compared with the experimental one of Fig. 1. Clearly, the experimental correlations are not a straight line going to zero at a separation of five patterns. We find the appearance of the correlations as well as their clear trend to decrease with the separation in the training sequence, down to very small values at a separation of five, very significant. All that was put in was the synaptic structure connecting successive patterns in the sequence. The remaining differences may be attributed to several factors, all of which are under study. These factors are:

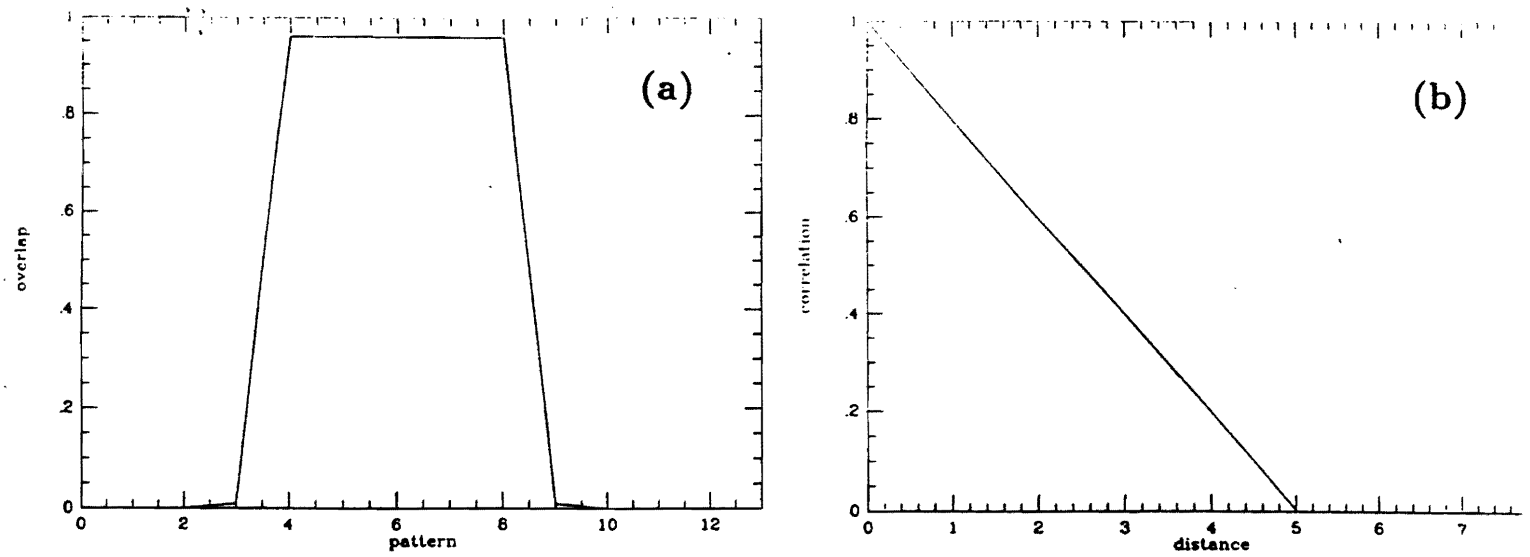

Figure 4: Correlations between attractors (a) and overlaps of attractor with patterns (b) vs the separation in the sequence of the patterns to which the attractors belong.

- The neurons in the experiment are analog neurons, represented by spike rates, and not discrete 0-1 neurons;
- In the experiment the neurons operate in the presence of noise, while here for simplicity we dealt with a noiseless situation;
- The matrix we chose is surely not the matrix in the monkey's cortex. One consequence is that all our attractors are identical;
- In the experiment the sample groups of neurons are small and are chosen in special ways. This leads inter alia to large fluctuations. Our correlations are ideal in that they take into account an infinite number of neurons.

All these effects can be studied either by an extension of the above reasoning to analog neurons with noise, or by simulations. These studies are under way.

The pattern to pattern correlations are computed according to Eq. 8. What remains is to determine $\overline{\sigma}$ and $|C|$ for this case. If the mean proportion of 1-bits in the attractors is g, then $\overline{\sigma} = g$ and $|C| = Ng(1-g)$. It is rewarding to see that the correlations between neighboring attractors are monotonically decreasing with the separation in the sequence, and are disappearing after the fifth neighbor, as in the experimental data.

In the present case, the number of condensed patterns, those having large values of the overlap with the attractor corresponding to a stimulus, depends on the value of a. This variation leaves finite intervals of a in which the attractors are invariant. Increasing a, we observe a sequence of bifurcations, where the number of condensed patterns increases by two. Correspondingly, the number of significantly correlated attractors, increases by four upon crossing a bifurcation value of a. Between any

two bifurcation points the solution does not change, i.e. the number of significantly correlated attractors, as well as the magnitude of the correlations, remain invariant.

4 Learning

In this section we will try to discuss possible learning scenarios, which could lead to a synaptic structure of the type considered in the previous sections. At the present time there is not enough information about the learning mechanism and memory preservation in the cortex, and our discussion can at best be tentative. We feel though that such a discussion may not be completely premature, just because of the level of specific detail provided by the experiments of Miyashita et al, and the ability of theory to approach a similar level of detail. Moreover, it is our feeling that a discussion of the implications to learning, of such findings, may lead to experiments which through neurophysiological correlates of behavior may shed additional light upon constraints on learning.

It is plausible to describe the synaptic dynamics as:

$$\frac{dJ_{ij}}{dt} = -\gamma J_{ij} + K_i L_j, \tag{16}$$

where γ is the rate of decay of the synaptic value and K_i, L_j are, respectively, the post- and pre-synaptic contributions to the synaptic efficacy. Both K and L depend on the activity of the corresponding neuron.

A simple mechanism which would lead to the matrix 1 could be to apply a usual Hebbian modification rule, with both pre- and post-synaptic terms as linear combinations of the current and preceding patterns, e.g.,

$$\begin{aligned} K_i(t) &= K_1(t)\xi_i^\mu + K_2(t)\xi_i^{\mu-1}, \\ L_i(t) &= L_1(t)\xi_i^\mu + L_2(t)\xi_i^{\mu-1}. \end{aligned} \tag{17}$$

This form may result from two different scenarios. In both we assume that strong presentations of the individual, uncorrelated patterns create attractors for those patterns themselves. Then, during training, which consists of many repeated presentations, the network, which remains in an attractor between presentations, is made to move to the next attractor by a new presentation. It should be emphasized, that in this description the role of the attractors is quite crucial. Before the patterns themselves are stored in the synapses as attractors, at the presentation of a consecutive pattern (to be learned) in the sequence during training, there is no memory of the previous pattern. This is especially true if the time between presentations of consecutive patterns is as long as in the experiments of Miyashita et al.

The difference between the two scenarios is in the way we view the origin of the source terms, K and L, for the synaptic change.

In the first we assume that the values of the neuronal spike frequencies represent, in an analog way, the transition between the two attractors. In this picture, $K_1(t) = L_1(t)=0$, before the transition starts, which is about when the next pattern is

presented. When the network is well established in the new attractor, $K_2(t) and L_2(t)$ tend to zero.

In the second scenario, one assumes that it is the local synaptic variable which remembers some short history of the pre- or post-synaptic activity of the corresponding neuron. For example, it may be the case that the synaptic mechanism may modify its effectiveness depending on the mean of the neuron's activity in some prior time window τ. The pre- or post-synaptic change may be enhanced, or suppressed, by a history of high average mean activity. Since the moving mean of the activity history is a linear combination of the activity in two consecutive patterns, while the network is moving from one attractor to the next, the end result is the same. Provided, of course, that the averaging window is short in comparison with the time of presence in each attractor. But this does not seem to be a strong requirement, given that the network stays in these attractors for many seconds.

As a simplest assumption we can take $K = L$, as functions of the pre- and post-synaptic activities, which implies $K_1 = L_1$, $K_2 = L_2$. Relaxing this constraint would lead us to asymmetric synaptic transition terms, of the type discussed in Section 2. If the resulting asymmetry is not large we expect the performance of the network to be robust. In the symmetric case, the contribution to the synaptic dynamics is:.

$$K_i L_j = K_1^2(t)\xi_i^\mu \xi_j^\mu + +K_1(t)K_2(t)[\xi_i^{\mu-1}\xi_j^\mu + \xi_j^{\mu-1}\xi_i^\mu] + K_2^2(t)\xi_i^{\mu-1}\xi_j^{\mu-1}. \qquad (18)$$

As one pattern follows the other, these contributions sum up, when equation 16 is integrated. If we neglect the exponential decay, γ, the summation is direct and after a long time, when all patterns have been presented many times in a fixed order, the resulting matrix would be proportional to:

$$J_{ij} \approx \sum_\mu \left(\int dt[K_1^2(t) + K_2^2(t)]\xi_i^\mu \xi_j^\mu + \int dt K_1(t)K_2(t)[\xi_i^{\mu+1}\xi_j^\mu + \xi_j^{\mu+1}\xi_i^\mu] \right), \qquad (19)$$

where the time integration is over an interval τ, in which synaptic modification is taking place.

This matrix has the same form as the one we introduced in the previous sections. It corresponds though to a case in which $a \leq 0.5$. This fact should not be considered too adverse. Synaptic decay, for example, is sufficient to raise a above 0.5. In the final analysis one should consider analog neurons, towards which the 0-1 neurons are an intermediate stage. Even for the discrete 0-1 neurons, the critical value of a is much lower than 0.5, while the heuristic learning mechanism can remain essentially the same.

Finally, if the patterns are presented in a random order during training, one can expect every pattern to be followed by any other one, given that a large number of presentations is required for satisfactory learning. This implies that the transition terms in Eq. 19, containing any particular pattern will be multiplied by a sum over all other patterns. That sum vanishes on the average and the transition terms become negligibly small. No correlations are then generated by the network, from uncorrelated patterns.

5 Experimental predictions and Some speculations

Given that a synaptic matrix, which can be learned without supervision, is able to convert temporal correlations into spatial ones, one is tempted to make some preliminary speculations about the computational and behavioral utility of such synaptic development. One, directly measurable, application had been pointed out in ref. [3]. In this experiment the monkeys are trained to recognize 24 visual patterns, organized in 12 fixed pairs. The pairs are presented in a random order. Correlations are generated among the two members of each pair, only. Those correlations are then shown to be correlated with the ability of the monkey to retrieve the second member of a pair, after being presented with the first. The basic nature of this type of association for the construction of cognitive behavioral patterns is quite immediate. What is special about this particular experiment is that the associative retrieval of the paired member is directly connected to the presence of the correlations in the representation of the pairs of attractors in the part of cortex under electro-physiological observation.

The interpretation of this experiment does not require speculation. To go one step beyond, one can expect the generation of such correlations to underlie the effect of priming[10]. In other words, if the network is in one of its attractors, and a new stimulus is presented, the transition between two attractors which are highly correlated (i.e. have a particularly large number of active neurons common to their representations) is much faster, than the transition between less correlated attractors. This effect was observed in a simulation with realistic neurons[11], when the pure patterns, involved in the construction of the synaptic matrix, were included explicit correlations[12]. This effect can be directly measured in a Miyashita[2] type experiment. One would expect that the transition time between different attractors would increase with the distance of the two patterns in the sequence of presentation.

In cognitive psychology the effect is familiar in experiments in which the reaction time is measured for the recognition of distorted words, or other images. This reaction time is significantly shortened if the pattern to be recognized is preceded by a cognitively correlated pattern.[10] In the language of the model we would say that the 'priming' image leads the network into its corresponding attractor. That attractor is correlated with the attractor corresponding to the test stimulus. Hence, the transition between the two is faster than the transition from some other state in which the network may find itself in otherwise. Complementing this scenario with the suggestion that at least part of our basic cognitive correlations are related to temporal contiguity of afferent stimuli, completes this speculation.

This interpretation can be extended one small step further. As attractors get increasingly correlated, there is an increase in the probability that noise would cause transitions between them, transitions of the Buhmann-Schulten type[13]. This opens the way for the scenario in which such transitions can be provoked in a cortical network by random afferent activation of the module. The transitions will tend to take place between correlated attractors, which in the present model are related to

temporal proximity during learning. Note that this process can also be observed in the experiments of the Miyashita type, though their cognitive content is more difficult to investigate.

One could hope to be able to investigate the process of learning the matrix which generates the correlations. We have argued in Section 4 that the process will go through the intermediate stage of learning the pure pattern attractors first. This was based on the assumption that there is autonomous learning in the particular module under observation. This is not self evident, and it may be that the pure patterns are quickly learned as attractors in a different area, as hippocampus, for example, and those attractors then assist in learning the correlated attractors. Since the question is open, one could attempt to clarify it by presenting different parts of the training sequence, in an experiment such as [2], with different frequencies. Then, if learning actually first goes through the creation of individual attractors for the pure patterns, one should observe lower correlations in the parts shown less frequently, as well as lower coding rates. In other words, pure patterns are expected to use less neurons than the composite patterns correlated by the dynamics. See e.g. Section 3. On the other hand, if the module learns the correlated attractors directly, no group of patterns should show the appearance of uncorrelated attractors.

Acknowledgements

The authors acknowledge useful discussions with N. Rubin, S. Seung and H. Sompolinsky. DJA has benefitted from many useful discussions of the learning mechanisms with Stefano Fusi. We are indebted to N. Rubin for information concerning priming effects. MVT's research is supported in part by the Ministry of Science and Technology.

References

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

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

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

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

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

[6] Amit DJ and Tsodyks MV 1991 Quantitative study of attractor neural network retrieving at low spike rates I: Substrate – spikes, rates and neuronal gain *NETWORK* **2** 259 and Low-rate retrieval in symmetric networks *NETWORK* **2** 275

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

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

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

[10] Besher D, Humpreys G, Eds. 1990 *Basic processes in reading* (Hillside, NY); Tweney RD, Heiman GH and Hoemann HW 1977 Effects of Visual Disruption on Sighn Intelligibility, *Journal of experimental psycology, General*, **106**, 255.

[11] Amit DJ, Evans MR and Abeles M, 1991 Attractor neural networks with biological probe neurons, *NETWORK*, **1**, 381

[12] Rubin N, 1991 Private communication.

[13] Buhmann J and Schulten 1987 Noise driven associations in neural networks *Europhys. Lett.* **4** 1205