Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors

M. Grinasty
M. V. Tsodyks*
Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem

Daniel J. Amit
INFN, Sezione di Roma, Istituto di Fisica, Universita di Roma, La Sapienza, P.le Aldo Moro, Roma, Italy

It is shown that a simple modification of synaptic structures (of the Hopfield type) constructed to produce autoassociative 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 that 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.

1 Introduction

1.1 Temporal to Spatial Correlations in Monkey Cortex. The remarkable sequence of neurocognitive experiments by Miyashita (1988), Miyashita and Chang (1988), and Sakai and Miyashita (1991) 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 (Miyashita and Chang 1988), 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 sec after the removal of the stimulus. The fact that selective, stimulus-related, enhancement of neural activity persists for 16 sec in the absence of the provoking stimulus is evidence of nonergodic attractor dynamics (see, e.g., Amit 1992). The same encouraging evidence has forced a confrontation on the question of activity rates on retrieval by attractors. The rates in the Miyashita attractors were many times lower than what models of the Hopfield type (Hopfield 1982; Amit 1989) predicted. This fruitful confrontation led to a study (Amit and Tsodyks 1991) that 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 (Miyashita 1988) 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 Miyashita (1988) is reproduced in Figure 1.

1.2 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, that is, the presumed items in the learning process. This went under the name of autoassociation. Here we shall show that a simple modification of the synaptic matrices used for autoassociation in ANNs leads to a relaxation dynamics that associates with stimuli near one of the random, uncorrelated underlying patterns, an attractor that is correlated with several patterns. The patterns that have the largest correlations with a given attractor are the neighbors of the stimulus leading to the attractor, in the sequence of stored patterns.

![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 Miyashita (1988).](image-url)

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 (Tsodyks and Feigel'man 1988; Buhmann et al. 1989), in which this symmetry is removed and that can naturally be interpreted in terms 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 autoassociative 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., Amit 1989).

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 that 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 \( S_i(t) = ±1 \), 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 \( \mu \), \( \xi_i^\mu \), the proposed synaptic matrix is written as:

\[
I_i = \frac{1}{N} \sum_{\mu=1}^{P} \left[ \xi_i^\mu \xi_i^{\mu+1} + a(\xi_i^{\mu+1} + \xi_i^{\mu+1}) \right]
\]

(2.1)

where \( P \) is the total number of patterns stored in the connections. The patterns, \( \mu \), 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 preceding pattern. Note, in particular, that this extended matrix still preserves the symmetry of the original Hopfield network, 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 equation 2.1, for \( a = 0 \), reduces to the original Hopfield matrix.

This matrix is accompanied, as usual, by a schematic spike emission dynamics that, in the noiseless case, determines the new state, \( S_i(t) \), of the neuron according to

\[
S_i(t + \Delta t) = \text{sign}[h_i(t + \Delta t)]
\]

(2.2)

where

\[
h_i(t + \Delta t) = \sum_{j=1}^{N} J_{ij} S_j(t)
\]

(2.3)

\( h_i \) mimics the value of the postsynaptic 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.

Conversion of Temporal to Spatial Correlations

The natural variables for the description of the nonergodic 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 \( \mu \). They are defined as

\[
m^\mu = \frac{1}{N} \sum_{i} S_i(t) \xi_i^\mu
\]

(2.4)

See, for example, Amit et al. (1985) and Amit (1989). The value of overlap \( m^\mu \) measures how close is the state of the network to the stored pattern \( \mu \). If \( m^\mu = 1 \), the state is identical, as a binary word to the pattern \( \mu \), that is, \( S_i = \xi_i^\mu \) for all \( i \). If \( m^\mu = 1 \) for the asymptotic state, the attractor, then the corresponding pattern is retrieved perfectly.

With the matrix, equation 2.1, the "field" \( h_i \) can be expressed in terms of the overlaps (Amit et al. 1985), which implies that so can the dynamics of the network as well as its attractors. Namely, we can write equation 2.3 as

\[
h_i(t) = \sum_{\mu} m^\mu(t)[\xi_i^\mu + a(\xi_i^{\mu+1} + \xi_i^{\mu-1})]
\]

(2.5)

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:

\[
m^\mu = \left\langle \left\langle \xi_i^\mu \text{sign}\left( \sum_{\nu} m^\nu(\xi_i^\nu + a(\xi_i^{\nu+1} + \xi_i^{\nu-1})) \right) \right\rangle \right\rangle
\]

(2.6)

The double angular brackets imply an averaging over the distribution of the bits in the patterns (see, e.g., Amit et al. 1985; Amit 1989).

Autoassociation was the interpretation of the fact that, in the absence of noise, for low loading, the equations 2.6 had solutions with one singular \( m^\mu \neq 0 \), 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 (Amit et al. 1985). 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 equations 2.6, with \( \mu = 2 \) for example, one has:

\[
m^2 = \left\langle \left\langle \xi_i^2 \text{sign}\left( m^2(\xi_i^2 + a(\xi_i^1 + \xi_i^3)) \right) \right\rangle \right\rangle
\]

(2.7)

For \( a < 0.5 \), it is the first term in the square brackets that 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

\( ^* \)The 2s are superscripts not squares.
25% of the sites $\xi^3 = \xi^1 = -\xi^2$ and the argument has the sign opposite to that of $\xi^2$. Starting from a state with $m^2 = 1$, and all other overlaps 0, one arrives, after one step to a state with $m^2 = m^3 = m^0 = 0.5$. This is no fixed point either.

The solutions of equations 2.6 have several 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, $S_i = \xi_i$, 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 with nonzero overlaps with several 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 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 that 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 its 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 and Chang (1988) (Fig. 1). The correlation of activities in two attractors, $\sigma^\mu$ and $\sigma^\nu$, is defined as

$$C(\mu, \nu) = \frac{1}{|C|} \sum_{i=1}^{N} (\sigma^\mu_i - \bar{\sigma})(\sigma^\nu_i - \bar{\sigma})$$

(2.8)

where $\bar{\sigma}$ 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, $\bar{\sigma} = 0$, and $|C| = N$. Hence, the correlation of attractors $\mu$ and $\nu$ can be written as

$$C(\mu, \nu) = \frac{1}{N} \sum_{i} \sigma^\mu_i \sigma^\nu_i = \frac{1}{N} \sum_{i} \text{sign}(h^\mu_i)\text{sign}(h^\nu_i) = \langle \text{sign}(h^\mu)\text{sign}(h^\nu) \rangle$$

(2.9)

where $h^\mu_i$ is the local field on neuron $i$ when the network is in attractor number $\mu$. The last equality is an expression of self-averaging, giving an average over the distribution of patterns. Finally, substituting the fields from equation 2.5 in equation 2.9, we arrive at the correlation coefficient:

$$C(\mu, \nu) = \langle \left\{ \text{sign} \left( \sum_{\alpha} \xi^\alpha \left( m^\alpha_\mu + am^{\alpha+1}_\mu + am^{\alpha-1}_\mu \right) \right) \right\} \rangle$$

(2.10)

where $m^\rho_\mu$ is the overlap of the attractor corresponding to stimulus number $\mu$ with pattern number $\rho$.

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 on 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.

---

Note added in proof: L. Cugliandolo has recently proved that beyond 5 all overlaps are exactly zero.
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 (Tsodyks and Feigel'man 1988; Buhmann et al. 1989). 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, that is, the fraction of 1s 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

$$V_i(t + bt) = \Theta \left( \sum_j I_{ij} V_j(t) - \theta \right)$$

where $\Theta(x) = 1$ for $x > 0$, and 0 otherwise, and $\theta$ 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, $\eta_{\mu}$, are chosen as $N$-bit words of independently chosen 0,1s, that is

$$\eta_{\mu} = 0, 1 \quad \mu = 1, \ldots, p$$

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

An extension of the symmetric synaptic matrix, appropriate for autocorrelation (Tsodyks and Feigel'man 1988; Buhmann et al. 1989), to our requirements would be

$$I_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \left( (\eta_{\mu} - f)(\eta_{\mu} - f) + a(\eta_{\mu}^{\mu+1} - f)(\eta_{\mu}^{\mu} - f) \right)$$

and the corresponding overlaps are generalized to

$$m^\mu(t) = \frac{1}{Nf(1-f)} \sum_i (\eta_{\mu} - f) V_i(t)$$

With the couplings 3.3, 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

$$f(1-f)m^\mu = \left( \left( (\eta - f) \Theta \left( \sum \eta^{\mu} \left( (\eta^\mu - f) + a(\eta^{\mu+1} - f) + (\eta^{\mu-1} - f) \right) \right) - \theta \right) \right)$$
When \( a = 0 \), the system of equations reduces to that of Tsodyks and Feigel'man (1988) and Buhmann et al. (1989). In this case, at low loading, the exact stored patterns are the retrieval attractors of the network, that is, the equations admit solutions with a single nonvanishing overlap, which in turn is equal to 1. These attractors persist until \( a \) reaches the critical value of

\[
\alpha_c = \frac{f + \theta}{2(1-f)}
\]

Above this value of \( \alpha \), the pure patterns are unstable and the network, having a symmetric synaptic matrix, finds new fixed points. The equations 3.5 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: \( f = 0.01 \), \( \theta = 0.2 \), \( p = 11 \), and \( a = 0.25 \). Figure 4a represents the overlaps vs. the pattern number in the sequence, relative to the pattern of the stimulus. Figure 4b is the correlation between the attractors. Note that in distinction to the \( \pm 1 \) 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 4b may seem somewhat simple compared with the experimental one of Figure 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

- 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.

Conversion of Temporal to Spatial Correlations

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 attractor to attractor correlations are computed according to equation 2.8. What remains is to determine \( \bar{\sigma} \) and \( |C| \) for this case. If the mean proportion of 1-bits in the attractors is \( g \), then \( \bar{\sigma} = g \) and \( |C| = N (1-g) \), 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 \( \alpha \). This variation leaves finite intervals of \( \alpha \) in which the attractors are invariant. Increasing \( \alpha \), 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 on crossing a bifurcation value of \( \alpha \). Between any two bifurcation points the solution does not change, that is, 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 may shed additional light on constraints on learning through neurophysiological correlates of behavior.

It is plausible to describe the synaptic dynamics as

\[
\frac{dI_{ij}}{dt} = -\gamma I_{ij} + K_i L_j
\]  

(4.1)

where \( \gamma \) 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 that would lead to the matrix 2.1 could be to apply a usual Hebbian modification rule, with both pre- and post-synaptic
Conversion of Temporal to Spatial Correlations

terms as linear combinations of the current and preceding patterns, for example,

\[ K_i(t) = K_1(t) \xi_i^t + K_2(t) \xi_i^{t-1} \]
\[ L_i(t) = L_1(t) \xi_i^t + L_2(t) \xi_i^{t-1} \]  \hspace{1cm} (4.2)

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 that remembers some short history of the pre- or postsynaptic activity of the corresponding neuron. For example, it may be the case that the synaptic mechanism modifies its effectiveness depending on the mean of the neuron's activity in some prior time window \( \tau \). The pre- or postsynaptic change may be enhanced, or suppressed, by a history of high average mean activity. Since the moving mean of the activity 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 spent 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 simple 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

\[ K_i(t) = K_1(t) \xi_i^t + K_2(t) \xi_i^{t-1} \xi_i^t + K_1(t) \xi_i^{t-1} \xi_i^t + K_2(t) \xi_i^{t-1} \xi_i^{t-1} \]  \hspace{1cm} (4.3)
As one pattern follows the other, these contributions sum up, when equation 4.1 is integrated. If we neglect the exponential decay, \( \gamma \), 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

\[
I_{ij} \approx \sum_{a} \left( \int dt \left[ K_1^2(t) + K_2^2(t) \right] \xi_i^{a} \xi_j^{a} + \int dt K_1(t) K_2(t) \xi_i^{a+1} \xi_j^{a} + \xi_i^{a+1} \xi_j^{a} \right) \tag{4.4}
\]

where the time integration is over an interval \( \tau \), in which synaptic modification is taking place.

This matrix has the same form as the one we introduced in the previous sections. It corresponds 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, toward 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 equation 4.4, 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 Sakai and Miyashita (1991). 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 electrophysiological 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 (Besner and Humphreys 1990). In other words, if the network is in one of its attractors, and a new stimulus is presented, the transition between two attractors that 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 (Amit et al. 1991), when the pure patterns, involved in the construction of the synaptic matrix, included as explicit correlations (Rubin 1991). This effect can be directly measured in a Miyashita (1988) 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 (Besner and Humphreys 1990). 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 is 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 (Buhmann and Schulten 1987). 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 that 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 Miyashita (1988), 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 fewer 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.

Acknowledgments

The authors acknowledge useful discussions with N. Rubin, S. Seung, and H. Sompolinsky. DJA has benefited 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


Received 21 January 1992; accepted 5 May 1992.