Backprojections in the cerebral cortex: implications for memory storage

Alfonso Renart, Néstor Parga
Departamento de Física Teórica
Universidad Autónoma de Madrid
Canto Blanco, 28049 Madrid, Spain
and
Edmund T. Rolls
Oxford University
Department of Experimental Psychology
South Parks Road, Oxford OX1 3UD, England

Abstract

Cortical areas are characterized by forward and backward connections between adjacent cortical areas in a processing stream. Within each area there are recurrent collateral connections between the pyramidal cells. We analyze the properties of this architecture for memory storage and processing. Hebb-like synaptic modifiability in the connections, and attractor states, are incorporated. We show the following: (1) The number of memories that can be stored in the connected modules is of the same order of magnitude as the number that can be stored in any one module using the recurrent collateral connections, and is proportional to the number of effective connections per neuron. (2) Cooperation between modules leads to a small increase in the memory capacity. (3) Cooperation can also help retrieval in a module which is cued with a noisy or incomplete pattern. (4) If the connection strength between modules is strong, then global memory states which reflect the pairs of patterns on which the modules were trained together are found. (5) If the intermodule connection strengths are weaker, then separate, local, memory states can exist in each module. (6) The boundaries between the global and local retrieval states, and the non-retrieval state, are delimited. All these properties are analyzed quantitatively with the techniques of statistical physics.

P.A.C.S. 87.30 Biophysics of neurophysiological processes
Submitted to Neural Computation.

1 Introduction

Auto-associative memory systems, implemented in recurrent neural networks, have been intensively studied both to model the associative areas of the mammalian brain and to understand their storage capacity capabilities (Hopfield, 1982; Amit, 1989). The anatomical basis of such systems is well established. Local excitatory connections between nearby pyramidal cells (within e.g. 1 mm) are a characteristic property of cortical connectivity (see e.g. Braitenberg & Shuz, 1991; Rolls & Treves, 1998). These local excitatory connections may contribute to the response tuning of neurons in early (sensory) cortical areas (e.g. Grieve & Sillito, 1995), and to short term memory-related activity in higher cortical areas (see e.g. Amit, 1995; Rolls & Treves, 1998).

However, the effort has been mostly devoted to the analysis of single networks with only a small number of exceptions, and only in a few cases with a persistent (clamped) input stimulus (Amit, Parisi & Nicolis, 1990; Rau, Sherrington & Wong, 1991; Engel, Bouten, Komoda & Serneels, 1990). Although the research on single networks operating in the unclamped condition has been very fruitful, it is an idealization of the actual situation. Neuronal structures in the brain are linked to each other: neurons in a given area are connected not only to each other through axonal recurrent collaterals, but also different areas are interconnected, and different sensory pathways converge to multi-modal sensory areas. In the cerebral cortex of mammals, forward and backward projections between adjacent areas in a processing stream are a major feature of cortical connectivity. Moreover, there are as many backprojections between adjacent cortical areas in a cortical hierarchy as there are forward connections, and these connections may also be involved in similar functions of shaping receptive fields and contributing to short term memory-related activity (Rolls, 1989; Rolls & Treves, 1998).

A simplified description of the anatomy in this case is as follows (see further Rolls

& Treves, 1998) (Fig. 1): In primary sensory areas the main afferent input to the neocortex is from the thalamus. These inputs connect to spiny stellate cells in layer 4, which in turn connect to pyramidal cells located in the superficial layers 2+3. These send forward projections that terminate especially in the superficial layers (4, 2 and 3) of the next cortical area in the sequence of pyramidal cells. Backprojections originate mainly from the deep pyramidal cells (layer 5) of the second area and terminate in the superficial layers (1, 2 and 3) of the preceding cortical area, on pyramidal cells. In addition to backprojections from the succeeding cortical area in the hierarchy, there are also axons and terminals in layer 1 from the amygdala and (with several intermediate stages) from the hippocampus (van Hoesen, 1981; Turner, 1981; Amaral & Price, 1984; Amaral, 1986; Amaral, 1987).

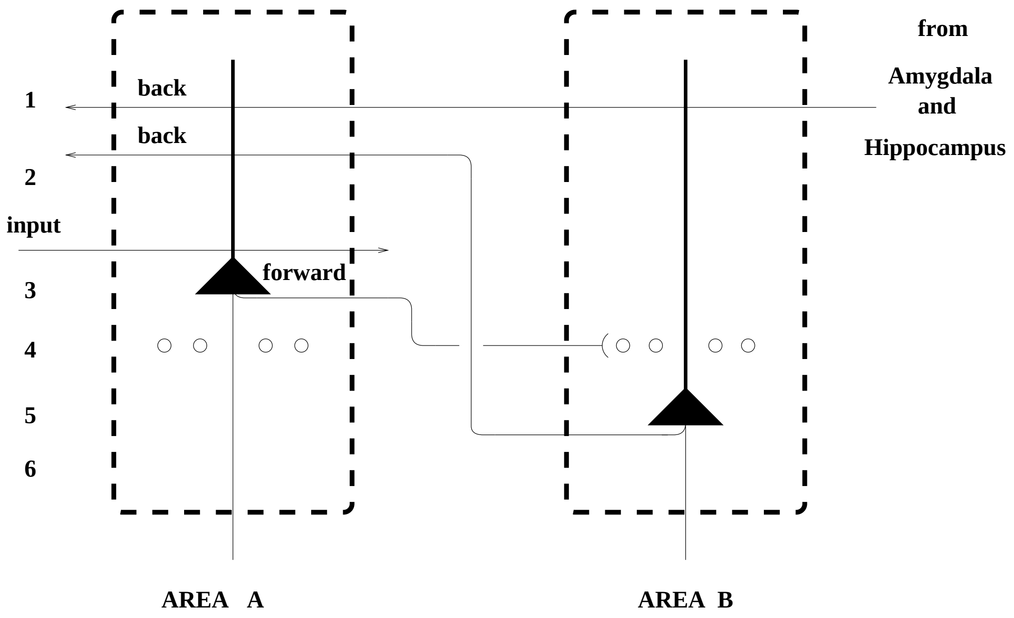

Figure 1: Forward and backward projections between two areas in the neocortex. The pyramidal cells in layers 2 and 3 of the area **A** project forward to terminate in the superficial layers (2-4) of the area **B**. In turn, the pyramidal cells in the deep layers of this area project back to layers 1-3 of area **A**. Also the hippocampus and the amygdala send backprojections to layer 1 of that area. Spiny stellate cells in layer 4 (present mainly in a primary cortical area) are represented by circles; the triangles represent pyramidal neurons.

In spite of this evidence, hypotheses are only starting to develop about the function of the cortico-cortical backprojections (Rolls, 1989; Rolls, 1996; Rolls & Treves, 1998). There have been only few theoretical analyses of multimodular recurrent networks (O'Kane & Treves, 1992; Lauro Grotto, Reich & Virasoro, 1994). The aim of this paper is to introduce a formal analysis of how the architecture shown in Fig. 1 could operate. The model we analyze considers this question, taking as a starting point the number of different activity patterns that could be stored and retrieved from such networks. One particular interest of the analysis is how the operation of one module influences what can be stored and retrieved in the connected modules, and as a whole in the overall multi-modular system. Although the connected modules may be part of an information processing system with inputs reaching module **A** and progressing through connected modules **B**, **C** etc and being transformed in the process (see Rolls & Treves, 1998), the present analysis, using the statistical physics approaches developed to understand memory systems, is being used to show how patterns could be stored and retrieved in this type of connected network.

With the approach introduced here, we are able to analyze the operation of whole series of modules of neurons arranged both as a linear sequence and with convergence at each processing stage from adjacent modules as occurs in different architectures present in the brain (see Rolls & Treves, 1998, Fig. 6.4 showing cortical forward and backprojection pathways, and Fig 4.6 showing a convergent tri-modular architecture which we will analyze in the future (Renart, Parga & Rolls, 1998)). In this paper we address the memory storage properties of bi-modular networks of the type illustrated in Fig. 1 and for which a formalism will be developed in Fig. 2. As we discuss below, the analysis of a bi-modular architecture is very revealing and also demonstrates the usefulness of the techniques employed here to provide insight into the properties of multi-modular systems.

The memory modules are composed of partially and randomly connected neurons. The modules are considered to have learnt sparse-coded binary patterns (which we will call the *features* or the *local* patterns) by modifying the synaptic efficacy between co-active cells within a module using for example a Hebb-like learning rule. Associations between the modules are implemented in the connections between them, again as a result of having used a Hebb-like learning rule. Given the statistical properties of the sensory data, some features of a stimulus may be represented in one module, and other features in another module. However, because it is the same sensory stimulus, there will be a statistical correlation between the activity patterns in the two modules. These statistical relationships will set up the inter-modular synaptic efficacies.

The intra- and inter-modular connectivities are independent parameters. Another parameter is the strength of the synaptic efficacies between different modules, relative to those within the same module. If this is large enough, stimulation of one of them can produce sustained activity in some of the other modules. To understand qualitatively and quantitatively the types of interaction that can occur between the different modules is one of the aims of this work.

The simplest architecture consists of two different sensory pathways, each of which processes different features of the stimulus (See Fig. 2). The information coming from these pathways is conveyed, in both cases, to cortical modules (**A** and **B**), which are at the same level of what might be a hierarchically organized processing stream. We assume that the modules are coupled recurrent neural networks of the type that we have just described. The inputs to these two areas may come either directly from the external world (in which case they make proximal synapses) or from other internal areas (in which case they make apical connections). The difference between these two types of synapses can be taken into account by giving different intensity to the corresponding stimuli. The modules can be in pathways of either different or the same sensory modality that, at some level, interchange information through the inter-modular collaterals. Both can be part for instance, of visual processing, one path taking care of object form processing while the other taking care of object motion.

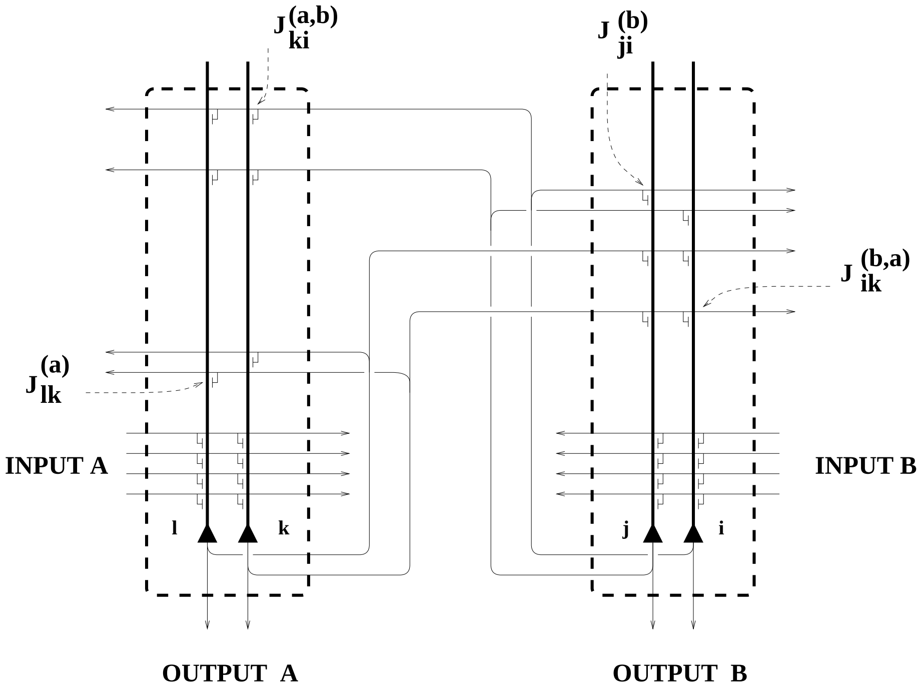

Figure 2: The bi-modular architecture. The triangles represent the neurons. $J^{(a)}_{lk}$ is the recurrent connection between the pre-synaptic neuron k and the post-synaptic neuron l, both in the cortical module **A**. Similarly for $J^{(b)}_{ji}$, but now this connection is in module **B**. $J^{(a,b)}_{ki}$ denotes the connection between the cells k and i in the modules **A** and **B**, respectively. The synaptic matrix is assumed to be symmetric. The figure shows the inputs **A** and **B** making proximal synapses with the network neurons. They may come also from other internal areas of the brain, making apical synapses. The recurrent inter-modular connections are assumed to be apical.

The retrieval behavior of this network can be discussed qualitatively. Let us first look at the situation where the coupling between the two modules is weak. If only one module were stimulated, sustained activity would appear only in this module. The activity pattern would be close to the corresponding feature stored in that module. In the same way, if the two modules were stimulated with features corresponding to the same stimulus, this would be represented by a global pattern of sustained activity highly correlated (in each module) to the pattern that represents the individual feature.

A global attractor can also be reached in a more interesting regime: When the association between the two features is sufficiently strong, stimulation of *only one* of the sensory paths suffices to produce a global pattern of activity. The other feature, which in the external world is frequently present together with the one used as a stimulus, has also been recalled. The resulting global attractor will be close to the union of the individual features.

A natural guess about the performance of this network is that its storage capacity will increase with respect to that of a single module. In fact, even if the coupling between them is weak, the attempt to retrieve one feature from one of the modules will produce a weak input to the other module, correlated with the right feature. In turn, the backprojections from this second module to the first will increase the strength of the signal. However, it is not guaranteed that this must necessarily happen: The large number of features that have not been stimulated act as a noise in the retrieval process. Since it is present in both modules, its contribution will also be back-projected and will compete with the signal.

In reality, one expects a more complicated situation. In general, a given feature in one module will be associated with more than one feature in the other. In this case, if the coupling is weak, the network will still work well: It will only retrieve the correct feature in the stimulated module. If the coupling is not weak, the response of the network will depend on the relative value of the strengths of the different associations. If a feature in module **A** is strongly coupled to only one feature in module **B**, then the performance of the network will still be good. But if there is not a dominant feature in **B** associated with that particular feature in **A**, the network will have mixed attractors consisting of several features of module **B**. The extreme situation corresponds to a feature in **A** strongly and equally associated with a set of, say, $s_{\mathbf{B}}$ features in **B**. Even more generally one can think that $s_{\mathbf{A}}$ features in **A** are strongly associated with $s_{\mathbf{B}}$ features in **B**. Under these conditions, stimulation of one of the modules will produce a global attractor consisting of the union of the whole set of associated features. The stored local patterns have been destabilized by the presence of these more complex associations. This extreme situation is probably not realistic; normally there will be one association dominating the others. However we will still consider this extreme case. The reason is that it will allow us to answer the following questions: How weak has the coupling to be in order to recover only the features used to stimulate the system ? Is it a non-zero value ? If the network behaved well within a reasonable range of values of the association coupling, its good behavior under more normal conditions would also be guaranteed.

There is yet another possibility that appears in the model when the number of stored features becomes too large (but still of the order of the size of the system). In this case the network fails to work as an associative memory. The appearance of this regime establishes the limits of good performance of the network.

As we will see, the model considered here reproduces all these situations (to be referred to as *phases* or regimes). Which of them is realized depends on the values of the parameters of the model.

We have discussed the architecture in Fig. 2 in the context of two sensory pathways, in which the modules are at the same level of processing but have cross-connections between them. The model we describe here can however be considered in a more abstract way, and is relevant to a number of different neural architectures.

An important instance is the problem where one of the modules is closer to the sensory input, and the other is the next layer in the processing stream or hierarchy. In this case the sensory input is identified with for example input A in Fig. 2, and (external) input B might not be present, or input B might be used to bring convergent information from another part of the brain. Now the two modules are two consecutive stages of the same sensory pathway (e.g. the early visual system).

Again, with this interpretation of Fig. 2, more complex, for example hierarchically organized, multi-modular neural systems could be obtained by adding more stages to the network, as considered elsewhere (Renart, Parga & Rolls, 1998). These architectures could be considered as a generalization of a purely feedforward network that has been recently proposed to explain transform invariant recognition (Wallis & Rolls, 1997). The effect of recurrent connections in view-invariant recognition in a single module recurrent network has also been analyzed (Parga & Rolls, 1998). In particular it has been proved that it can be used to store sets of views of the same object in such a way that any of them (or any state close to some of the views) could be used as a cue to retrieve the object.

Another possible interpretation of our model arises when one of the two modules in Fig. 2 is identified with the hippocampus (or more realistically with the hippocampus and some of the nearby areas contributing to its function as a temporary memory storage site), and the other is identified with the neocortex (see e.g. Rolls & Treves, 1998, Fig. 6.1).

Here we formulate a model to describe the memory storage properties of multi-modular networks which is both sufficiently plausible and solvable, and the solution of the bi-modular network is discussed in detail. This has been done from two different perspectives. First, we have explored the possible retrieval behaviors of the network as a function of its free parameters. This amounts to solving the problem of how many activity patterns can be stored as some of these parameters are varied. This issue has been extensively studied in the case of an isolated module of binary units (Amit, 1989) where, in particular, it is known that in the limit of a very sparse network, the storage capacity is significantly increased (Tsodyks & Feigel'man, 1988). In this regime, a value of the effective neural threshold exists for which the capacity achieves its largest value. We have therefore extended these results to the case of binary bi-modular networks in the sparse coding limit. The use of binary neurons, apart from allowing a direct comparison with results from uni-modular networks, has the advantage that the analysis is simpler and a more complete and systematic study of the possible behaviors of the network can be performed.

Second, we have analyzed the retrieval regimes achieved by the network under more realistic conditions. This has been achieved first, by studying a network of neurons described in terms of their firing rates. Besides, the model parameters in this case have been chosen according to biological plausibility. For instance, the coding level (or sparseness) of the stored patterns has been set to a value consistent with experimental firing rate distributions and the inter- and intra-modular connectivities have been given realistic values. In this context two main issues have been discussed: the phase diagram of such a network and its performance under noisy conditions.

Our motivation differs from the work by O'Kane & Treves (1992). These authors have addressed the question of modeling the cortex in terms of a multi-columnar network. In that case, one is interested in the limit where the number of modules is very large and at the same time the number of inter-modular connections is very small, in such a way that the total number of connections per neuron is kept constant. Also with a different motivation, Lauro Grotto, Reich & Virasoro (1994) have performed numerical simulations of multi-modular networks to model semantic memory.

We begin by describing the model (Section 2). The solution is presented in Section 3, where we give very general expressions, valid for arbitrary neuronal current-to-firing rate transduction functions. The numerical method followed to analyze the equations is presented in Section 4. The results are given in Section 5. Results concerning the two different motivations referred above are presented separately, the first subsection being devoted to the results found on the binary network, and the second to the analog case. The discussion of the results and perspectives for future work are given in the last Section. Some technical aspects are included in appendices.

2 The model

The model can be described as follows:

- *The architecture:* The network consists of two modules with N neurons in each of them. The number of neurons per module is very large (i.e. $N \to \infty$). Although in principle the algebra can be done for more complex architectures, here we will only present the model for the network shown in Fig. 2.

- *The neurons:* Neurons are described in terms of their firing rates. The network dynamics is defined according to the set of equations:

$$\frac{dI_{ai}(t)}{dt} = -\frac{I_{ai}(t)}{\mathcal{T}} + \sum_{bj} J_{ij}^{(a,b)} \nu_{bj} + h_{ai}^{(ext)} \qquad\qquad a, b = \mathbf{A}, \mathbf{B} \ . \tag{1}$$

  Here, I_{ai} is the afferent current into the neuron i of the module a, and ν_{bj} is the firing rate of the neuron j of the module b. The current is driven by the output spike rates of the other neurons in the network (located either in the same or in different modules), weighted with the corresponding synaptic efficacies $J_{ij}^{(a,b)}$, and by the stimulus (or external field) $h_{ai}^{(ext)}$. The afferent current decays with a characteristic time constant $\mathcal{T}$. The transduction from currents to rates, necessary to complete the definition of the dynamics, will be indicated by $\nu = \phi(I)$.

- *The current-to-rate transduction function:* Two explicit choices of this function will be considered. These correspond to binary neurons, and analog neurons with a hyperbolic transduction function. Binary neurons are obtained with the choice:

$$\phi(I) = \begin{cases} 0 & \text{if } I < \theta \\ 1 & \text{if } I \geq \theta \ . \end{cases} \tag{2}$$

  The hyperbolic transfer function is:

$$\phi(I) = \begin{cases} 0 & \text{if } I < \theta \\ \tanh[G\ (I - \theta)] & \text{if } I \geq \theta \ , \end{cases} \tag{3}$$

where G is the gain.

- *The stored patterns:* In each module a, the stored patterns (that will be also referred to as *features*) have been classified in L sets of s_a patterns. The sizes of these sets will be kept finite, but L will be taken $O(N)$.

  The number P_a of patterns stored in a is therefore $L\ s_a$ These are defined in terms of binary variables $\eta_{ai}^{\beta\nu}$ ($\beta = 1, \ldots, L$; $\nu = 1, \ldots, s_a$; $i = 1, \ldots, N$). The η's are independent random variables which are chosen equal to one with probability f (the mean coding rate of the stimuli, the same as the sparseness as defined by Rolls & Treves (1998) in the case of binary values) and equal to zero with probability $(1-f)$. Their variance is $\chi \equiv f(1-f)$.

- *The synaptic connections:* The synaptic matrix will be denoted by $J_{ij}^{(a,b)}$, where again a and b are module indices and i and j are neurons in a and b respectively. The main constraint that we will impose on this matrix is symmetry under the interchange of the neuron indices. The intra-modular recurrent connections are:

$$J_{ij}^{(a,a)} \equiv \frac{d_{ij}^0}{\chi N_t} \sum_{\mu=1}^{s_a} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f)\ (\eta_{aj}^{\beta\mu} - f) \qquad i \neq j; \quad \forall\ a\ , \tag{4}$$

  and $J_{ii}^{(a,a)} = 0$. For a given module a, the symmetric synaptic matrix in eq. (4) stores the P_a local features $\eta_{ai}^{\beta\nu}$. In order to have correct retrieval properties, the variables that appear in the connection matrix are not the η's, but the differences $(\eta - f)$ (Tsodyks & Feigel'man, 1988).

  The network connections are diluted. This is implemented through random variables d_{ij}^0. These take the value one with probability d_0 and the value zero with probability $(1-d_0)$.

  The inter-modular connections are given by:

$$J_{ij}^{(a,b)} \equiv \frac{g_{ab}\, d_{ij}^{ab}}{\chi N_t} \sum_{\mu,\nu=1}^{s_a,s_b} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f)\,(\eta_{bj}^{\beta\nu} - f) \qquad \forall i,j; \quad a \neq b \ . \tag{5}$$

In the inter-modular connections proposed in eq.(5), all the s_a patterns belonging to the *same* set in a given module a are associated with all the s_b patterns belonging to the corresponding set in any other module b. The strength of these associations, g_{ab}, is the same regardless of the particular pair of patterns.

These connections are also diluted. This is implemented through the random variables d_{ij}^{ab}, which take the value one with probability d and the value zero with probability $(1-d)$. The neurons i and j, located in the modules a and b respectively, are connected only if $d_{ij}^{ab} = 1$.

The symmetry requirement imposes that $g_{ab} = g_{ba} = g$, $d_{ij}^{0} = d_{ji}^{0}$ and $d_{ij}^{ab} = d_{ji}^{ba}$. In the case of the dilution variables, this means that only half of them are drawn randomly. The other half are *set* equal to their symmetric counterparts.

The weight normalization is $\chi N_t \equiv \chi N \Lambda$ with $\Lambda = d_0 + gd$, where N_t is the average effective number of connections afferent to a given neuron.

The synaptic connections (4) and (5) can be expressed in terms of a single association matrix, $\tilde{K}$, that contains all the information about the architecture of the network. Assuming for simplicity that the s_a's take the same value, s, for all modules, its elements have the form:

$$\tilde{K}_{\mu\nu}^{a_i b_j} = \frac{d_{ij}^{0}}{\Lambda}\left(\delta^{ab} \otimes \delta_{\mu\nu}\right) + \frac{g d_{ij}^{ab}}{\Lambda}\left((1^{ab} - \delta^{ab}) \otimes 1_{\mu\nu}\right) \quad , \tag{6}$$

where the symbol $\otimes$ denotes the tensor product between the module and pattern spaces, and 1^{ab} and $1_{\mu\nu}$ are equal to one for all a, b and μ and ν respectively.

The element $\tilde{K}_{\mu\nu}^{a_i b_j}$ measures the contribution to the synaptic efficacy between the neurons i and j (in modules a and b respectively) resulting when the module

a is in the state μ and b is in the state ν. Using this matrix, the intra and inter-modular connections can be written in terms of a single expression

$$J_{ij}^{(a,b)} = \frac{1}{\chi N} \sum_{\mu,\nu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f) \tilde{K}_{\mu\nu}^{a_i b_j} (\eta_{bj}^{\beta\nu} - f) \qquad ai \neq bj \ . \tag{7}$$

- *The external field:* The external field can be chosen as one of the stored patterns (e.g. pattern μ_0) or a distorted version of it. Noisy versions of the features have been obtained by a simple stochastic process that keeps the average global activity of the stimulus constant. This is done by visiting all the sites of the pattern and applying, independently in each of them, the rule:

  If the site is active: $\eta^{\mu_0} = 1 \to 0$, with probability δ.

  If the site is not active: $\eta^{\mu_0} = 0 \to 1$, with probability δ'.

  In order to ensure a fixed average global activity, the parameters δ and δ' have to be related as: $f\ \delta = (1-f)\ \delta'$.

  The distorted pattern $\tilde{\eta}^{\mu_0}$ can be expressed as

$$\tilde{\eta}^{\mu_0} = \eta^{\mu_0}(1-\xi_1) + (1-\eta^{\mu_0})\xi_0, \tag{8}$$

where ξ_1 and ξ_0 are two binary random variables that take the value one with probabilities δ and δ' respectively. Since δ and δ' are not independent parameters, the distortion of a given pattern (μ_0 of module a) can be characterized by its overlap with the correct version of itself. This overlap is defined as [1]:

$$m_a^{\mu_0}(\delta) \equiv \frac{1}{\chi N} \ll \sum_i (\eta_{ai}^{\mu_0} - f)\ \tilde{\eta}_{ai}^{\mu_0} \gg_\eta = 1 - \frac{\delta}{(1-f)} \tag{9}$$

The external field $h_{ai}^{(ext)}$ at a given neuron i in the module a, is now chosen as:

[1] Notice that while the maximum value of the overlap is 1, its minimum value is $-f/(1-f)$, that becomes -1 only for $f = 0.5$.

$$h_{ai}^{(ext)} = h \; \tilde{\eta}_{ai}^{\mu_0}, \tag{10}$$

where h is the strength of the stimulus.

The total number of patterns per module, $P = Ls$ is extensive. This means that

$$P = \alpha N_t, \tag{11}$$

where α is the load parameter or storage level of the system.

An alternative definition of the load parameter would have been $P = \alpha' N_t'$ where $N_t' = N(d_0 + d)$. If $g \neq 0$, the parameter α' measures the number of stored patterns per mean number of connections to a given neuron. However, this interpretation is not true for $g = 0$. For this reason we have preferred to use $N_t = N(d_0 + gd)$ which can be interpreted as the effective number of connections to a given neuron taking into account the strength of the inter-modular connections.

Notice that since $\alpha' = \alpha \frac{d_0+gd}{d_0+d}$, the network capacities will increase faster with g if α' is used.

Let us finally remark that the expression (5) is symmetric under the interchange of the module and the neuron indices. This will allow us to find an analytical solution of the model (Kuhn, 1990; Amit, 1989). We note that although the analysis used here assumes that the synaptic connections are reciprocal in strength (as would be the case with a fully connected recurrent network trained with a Hebb-like rule, see Rolls & Treves, 1998), it is found, in at least that type of network, that when the synaptic connectivity is diluted, then in large systems the dilution need not be symmetric for the network still to operate in simulations in a similar way to that described analytically (see e.g. Simmen, Treves & Rolls, 1996; Rolls, Treves, Foster & Perez-Vicente, 1997) [2].

[2] Asymmetric random dilution of the synapses is is not the only type of asymmetry encountered in real networks. In reality, the neurons which receive backprojections are even different from those sending them, as can be seen in Fig. 1

3 The solution

We want to study the storage capacity properties of the fixed points of the set of eqs. (1) representing the sustained activity states of the network. If the synaptic couplings are symmetric, an analytical solution can be found by means of statistical physics techniques (Amit, 1989; Kuhn, 1990). Very briefly, one first realizes that there is a Hamiltonian function associated with eqs. (1). In fact, if one considers

$$\mathbf{H} = -\frac{1}{2}\sum_{ab\,ij} J_{ij}^{(a,b)}\nu_{bj}\nu_{ai} - \sum_{ai} h_{ai}^{(ext)}\nu_{ai} + \sum_{ai}\mathcal{W}(\nu_{ai}) \quad , \tag{12}$$

where $\mathcal{W}$ is the integrated inverse current-to-rate relation

$$\mathcal{W}(\nu) = \frac{1}{\mathcal{T}}\int_0^{\nu} I(\nu')d\nu' \quad , \tag{13}$$

it is easy to check that variation with respect to the firing rates reproduces the dynamics.

Next, the model is generalized by introducing a temperature-like T parameter as a measure of the fast synaptic noise (Amit, 1989). In a heat-bath at temperature T, the probability of finding the system in a particular state $\{\nu_{ai}\}$ is given by its Boltzmann weight:

$$Pr(\{\nu_{ai}\}) = \frac{\exp -(\beta\mathbf{H}(\{\nu_{ai}\}))}{\mathcal{Z}(\beta)} \quad , \tag{14}$$

where $\beta \equiv 1/T$ and $\mathcal{Z}(\beta)$ is the partition function at temperature T, defined as:

$$\mathcal{Z}(\beta) = Tr_{\{\nu_{ai}\}}\ \exp -(\beta\mathbf{H}(\{\nu_{ai}\})) \quad . \tag{15}$$

The symbol Tr means a sum over all possible values of the dynamical variables $\{\nu_{ai}\}$ (or an integral if they are continuous). Let us notice that $\mathcal{Z}(\beta)$ is computed for a fixed realization of the η's, the dilution variables and the variables used to distort the stimulus. We will say that these are *quenched* random variables. However, we are not interested in the properties of the network for a particular realization of those variables, but in its average properties. Besides, since all the meaningful quantities have to be computed from the $\log \mathcal{Z}$, it is this object that has to be averaged over all the quenched variables. We then define the free energy per neuron as:

$$\mathcal{F} = - \lim_{N \to \infty} \frac{1}{\beta \, MN} \ll \log \mathcal{Z} \gg \quad , \tag{16}$$

where $M = 2$ is the number of modules and $\ll \ldots \gg$ means the average over the distributions of the η's, the d^0_{ij}'s, the d_{ij}'s and the ξ's, as have been described in the previous Section.

In the last years, a number of techniques have been developed to handle this type of problem (Mezard, Parisi & Virasoro, 1987). We have used the replica method along with the *saddle-point* method for the evaluation of the free energy in the very large N limit (see Appendix A). The result at finite temperature is:

$$\begin{aligned}
\mathcal{F}(\beta) \; = \; & \frac{\chi}{2M} \sum_{\mu\nu \; ab} m^\mu_a K^{ab}_{\mu\nu} m^\nu_b + \frac{\alpha\beta}{2M} \sum_a \left(r_{0_a} q_{0_a} - r_a q_a \right) + \\
& + \frac{\beta}{4M} \left(\sum_a \Delta^{(0)\,2}_a (q^2_{0_a} - q^2_a) + \sum_{(ab)} \Delta^2_{ab} (q_{0_a} q_{0_b} - q_a q_b) \right) + \\
& + \frac{L}{\beta \; N \; 2M} \left[Tr \; \ln \left[\delta_{\mu\nu} \otimes \delta^{ab} - \beta (Q^{ab}_{0_{\mu\nu}} - Q^{ab}_{\mu\nu}) \right] - \right. \\
& \left. - \beta \; Tr \left[Q^{ab}_{\mu\nu} \left(\delta_{\mu\nu} \otimes \delta^{ab} - \beta (Q^{ab}_{0_{\mu\nu}} - Q^{ab}_{\mu\nu}) \right)^{-1} \right] \right] - \\
& - \frac{1}{\beta M} \sum_a \ll \ln \left\{ \int_0^1 d\rho(\nu_a) \exp \left(\beta \; \mathcal{O}_a(\nu_a) \right) \right\} \gg_{\eta z \xi} \quad ,
\end{aligned} \tag{17}$$

where (ab) means all pairs of different modules. This expression is only valid to study the retrieval solutions in which the network is trying to retrieve some of the patterns in *one* of the sets in each module. For this reason the set index has been omitted. The symbol $\ll \ldots \gg_{\eta z \xi}$ means an average over η, ξ and z. The variable z is a random gaussian variable with zero mean and unit variance. It represents the random fluctuations in the effective current afferent to a given neuron due to the large number of stored patterns.

The matrix $K^{ab}_{\mu\nu}$ and the quantities $\Delta_a^{(0)\,2}$ and Δ_{ab}^2 have been defined in Appendix B, where a short explanation on the treatment of the dilution has been made. The integral over the rate implements the trace over the dynamical variables, where $d\rho(\nu_a)$ is the measure of integration, and depends on the type of neurons one is considering (see Appendices C and D for specific choices). Besides:

$$\begin{aligned}\mathcal{O}_a(\nu_a) \;=\; & \nu_a\left[\sum_\mu(\eta_a^\mu - f)\left(\sum_{b\nu} K^{ab}_{\mu\nu} m_b^\nu\right) + \sum_\mu h_a^\mu(\eta,\xi) + z\sqrt{\alpha r_a + \Delta_a^2 q_a + \sum_{b\neq a}\Delta_{ab}^2 q_b}\right] + \\ & + \frac{(\nu_a)^2}{2}\left[\alpha\beta(r_{0_a} - r_a) + \beta\left(\Delta_a^2(q_{0_a} - q_a) + \sum_{b\neq a}\Delta_{ab}^2(q_{0_b} - q_b)\right) - d_0\alpha\right] - \\ & - \mathcal{W}(\nu_a) \quad , \qquad\qquad (18)\end{aligned}$$

and we have also defined:

$$Q^{ab}_{\mu\nu} \;\equiv\; \sum_{\tau\, c} q_a(\delta^{ac}\otimes\delta_{\mu\tau})K^{cb}_{\tau\nu} \qquad (19)$$

$$Q^{ab}_{0_{\mu\nu}} \;\equiv\; \sum_{\tau\, c} q_{0_a}(\delta^{ac}\otimes\delta_{\mu\tau})K^{cb}_{\tau\nu} \quad . \qquad (20)$$

The quantities $m_a^{\beta\mu}$, q_a, q_{0_a}, r_a and r_{0a} are called *order parameters* and serve to characterize macroscopically the state of the system. Their definitions are:

$$m_a^{\beta\mu} \;=\; \frac{1}{\chi N} \ll \sum_i (\eta_{ai}^{\beta\mu} - f)\langle\nu_{ai}\rangle \gg_{\eta,\xi} \qquad (21)$$

$$q_a \;=\; \frac{1}{N} \ll \sum_i \langle\nu_{ai}\rangle^2 \gg_{\eta,\xi} \qquad (22)$$

$$q_{0_a} \;=\; \frac{1}{N} \ll \sum_i \langle\nu_{ai}^2\rangle \gg_{\eta,\xi} \qquad (23)$$

$$\alpha r_a \;=\; \chi\sum_{\beta>1,\,\mu}\langle\bar{m}_a^{\beta\mu}\rangle^2 \qquad (24)$$

$$\alpha r_{0a} \;=\; \chi\sum_{\beta>1,\,\mu}\langle(\bar{m}_a^{\beta\mu})^2\rangle \quad , \qquad (25)$$

where $\langle \ldots \rangle$ stands for the thermal average taken with the distribution (14). Here, the set index β has been included. The $\bar{m}$'s are related to the physical overlaps m's through a linear transformation:

$$\bar{m}_a^{\beta\mu} = \sum_{\nu b} K_{\mu\nu}^{ab} m_b^{\beta\nu} \quad , \tag{26}$$

The order parameter $m_a^{\beta\mu}$ measures the overlap of the state of module a with a given feature, averaged over all its possible values. The quantity αr_a is the variance of the gaussian noise generated by the large number of patterns stored and not being retrieved. As is evident from its definition, in the binary representation the parameter q_{0_a} is the mean activity in the attractor of module a. An interpretation of this parameter for analog neurons at $T = 0$ is given in Appendix D.

Since we are interested in the fixed points of eqs. (1), we have to take the zero temperature limit of the solution. In this limit q_a and q_{0_a} become equal, and the same happens to r_a and r_{0_a}. However, the slope with which each parameter approaches its partner as T goes to zero, remains finite. These slopes are:

$$c_a \equiv \lim_{T \to 0} \beta(q_{0_a} - q_a) \qquad \bar{c}_a \equiv \lim_{T \to 0} \beta(r_{0_a} - r_a) \tag{27}$$

The zero temperature limit has to be taken separately for the analog and the binary representations. Equations for the order parameters in the sustained activity states at zero temperature are given in Appendices C and D for binary and analog neurons respectively. The final equations of our analysis are eqs. (43-45) for binary neurons (that will be solved in Sec. 5.1), and eqs. (52-54) for analog neurons (that will be solved in Sec. 5.2).

4 Numerical Analysis

The parameters of the model are: The coding rate (or sparseness) f, the threshold θ, the intra- and inter-modular connectivities d_0 and d, the load parameter α, the gain (in the case of analog neurons) and the coupling strength g.

One has to distinguish between the different approaches followed in the case of the binary and analog networks. In the binary case we wanted to explore the influence of as many of these parameters as possible. However, to make the analysis tractable, some of them had to be kept fixed. This was the case of the connectivities (which were given plausible values) and of the coding rate, which was set to be very small ($f = 0.001$), according to the arguments given in the Introduction. Then, for different values of the neural threshold and for different stimulation conditions a parameter space of a smaller dimension, defined by α and g was explored. The association coupling varies between zero and one; this is because it is assumed that in the learning process intra-modular associations of a stimulus (proportional to the number of times the network has processed it while learning) are always greater than the associations between features in different modules (proportional to the number of times they have been processed at the same time). For the analog network, only α and g (and the distortion of the stimuli) were varied. The rest of the parameters, including f, where kept fixed at realistic values in this case.

When looking at the fixed points of the bi-modular network in the (g, α) plane, we plotted the values of the load parameter where the system changed its behavior (the critical lines) as a function of g. This gives rise to a phase (or retrieval) diagram.

To determine this diagram we follow the steps described below:

- *Initialization:* To find out the behavior of a point in the (α, g) plane, the modules are assumed to be initially silent, with all order parameters set to zero. An initial external current represented by the external field term in equations (49 -51) is then applied to initialize the network. This external field is mainly determined by the stored features, and can be applied to either one of the modules, or to both.

  The choice of the initial state of the network is very important since it determines the nature of the attractor state. In large collective systems (the global network) in which the interactions between the dynamic elements (the neurons) are random, there exist a very large number of stable (sustained activity) states such that it is very difficult for the network to jump to the neighbourhood of a given persistent state from the neighbourhood of another (Mezard, Parisi & Virasoro, 1987). Therefore the network evolves to the persistent state closest to the initial configuration. In fact, we will see that for given values of the network parameters the retrieved state depends on the nature of the applied stimulus.

- *Finding the solution:* The self-consistency equations (43-45) for binary neurons, and (52-54) for analog neurons are solved by using an iterative procedure. After initializing the network as described above, the procedure is applied until a fixed point is reached.

  One has to distinguish here the case where the solution is found after the application of a brief stimulus (unclamped conditions) from the case where it is found under the influence of a persistent field (clamped conditions). In the first case the field is kept on during a small number of iterations (about five), and then the network is left to evolve freely until it converges.

- *The retrieval diagram:* The values of the order parameters at the fixed point determine the nature of this point in parameter space. A systematic exploration of the (α, g) plane yields the retrieval diagram.

5 Results

5.1 Binary neurons

In this subsection the possible behaviors of the bi-modular system will be throughly analyzed using a network of binary neurons. In short, the boundaries of the different phases reached are found in the (g, α) plane for different values of the effective neural threshold and under different stimulation conditions involving one or the two modules and persistent or transient external stimuli. In some of these situations, the presence of multiple associations between features stored in different modules is also studied. Since we are interested in the storage capacity achieved by the network in its local or global retrieval regimes, we have chosen to put the network in the limit of a very sparse code, in which, at least for a single module, this capacity is greatly enhanced. This is implemented in the model by setting f to a small value, which has been taken as $f = 0.001$ in this subsection.

A signal-to-noise analysis in single module networks shows that the threshold has to be of order one in the small f regime that we are considering (i.e. it has to be independent of f), (Tsodyks & Feigel'man, 1988). We have extended this analysis to the model described in Section 2 and checked that the result still holds for multi-modular architectures. In all these cases, the threshold plays an important role in tuning the input to the most sensitive region of the response function. As a consequence of this, there appears an optimal value of the threshold where the capacity reaches its largest value (Fig. 3). We will also see that the performance of multi-modular networks changes significantly as the threshold passes through an optimal value.

The effect of the threshold on the capacity of a single module network can be seen in Fig. 3, together with the corresponding curve for the bi-modular network. Considered as a multi-modular architecture, the single module network corresponds to a zero value of the association strength g. It reaches its largest value for $\theta \sim 0.7$ for a coding level of $f = 0.001$. For θ larger than the optimal value, the noise produced by the other patterns (Tsodyks & Feigel'man, 1988) tends to reduce the capacity. The actual value for the capacity in a single module is what we would expect in an autoassociation network with diluted connectivity and sparse representations (Treves & Rolls, 1991; Rolls & Treves, 1998).

The bi-modular network (for $g = 0.5$ and $s = 1$) shows a similar behavior, but with some differences. For a value of the threshold greater than $\theta \sim 0.2$ the system is in a local retrieval (LR) phase in which only the stimulated module is in a state of retrieval. The critical capacity in this phase also has a maximum at $\theta \sim 0.55$ which would be the optimum threshold at this value of g, and then decreases rapidly to zero. There is also a low threshold regime in which the system is in a global retrieval (GR) phase. This means that there is sustained activity correlated with one of the features stored in *both* modules: Both modules are in retrieval states. The size of this region depends on the value of g. If g is very large it will be easier for the unstimulated module to enter retrieval, so it will be able to overcome a larger threshold and the size of this region will grow. The other phases appearing are the null (N) phase in which all the neurons in the network are in a quiescent state, and the spin-glass (SG) phase, in which the state of the network shows activity, but uncorrelated with any of the memories stored in any of the two modules. The whole curve for $g = 0.5$ appears shifted to the left. This is probably due to the decrease, as g grows, of the strength of the intra-modular connections that sustain the activity in a local phase. After this analysis, it becomes clear that for all these cases the performance of the network will change significantly as θ varies from zero to one.

We have therefore taken several values of the threshold ($\theta = 0.3, 0.6$) for fixed values of the feature coding level f and of the dilution parameters d_0 and d.

In Fig. 4 we present the retrieval diagrams obtained for $\theta = 0.3$. The other fixed parameters are: $f = 0.001$, $d_0 = 0.1$ and $d = 0.05$. Figs. 4a-b refer to $s = 1$ while Figs. 4c-d are for $s = 3$.

We first discuss the case $s = 1$ (each feature is associated with only one feature in the other module). If the load parameter is not too high, the network works as a good autoassociative memory device. When only one of the modules is stimulated (see Fig. 4a), for small g and α the network reaches a local state, where only the stimulated module shows substantial sustained activity correlated with the stimulus. But the other module also responds well: It reaches a state with a small overlap with its feature associated with the stimulus. The overlaps with all the other features stored in this module are zero. Typically, the non-zero overlap is $O(10^{-2})$ and the mean activity is f times the overlap; this means that most of the active neurons in this module are also active in the complete activity pattern. Therefore, even though the signal that appears in the non-stimulated module is rather weak, it is in the correct direction. The feature is not completely retrieved, but nevertheless this module fulfils an important task by providing a feedback signal to the stimulated module. Consequently the capacity of this module increases with respect to its value at $g = 0$, as can be observed in Fig. 4a.

As g increases, effects appear that tend to change this behavior. One of them is that as the critical value of α becomes larger, also the noise produced by all the other features increases. This noise has not only a direct contribution from the stimulated module, but also a contribution from the noise produced by the features stored in the second module, which is back-projected to the first. The immediate consequence of this is that for some value of g the capacity drops, and part of the retrieval diagram is taken by a phase not correlated with any of the features.

The other effect that appears for large g is an increase of the signal in the non-stimulated module. The balance of these effects is the appearance of another memory regime for large g and small α. In this phase, the state of the second module acquires a larger component in the direction of the feature associated with the one used as a stimulus (in fact, this is a symmetric phase where the overlaps in the two modules are equal). The coupling between the modules has become large enough to produce recall in the second module by stimulating only the first. This happens however for values of g smaller than one. Finally, in the region of large α there is a non-retrieval phase for any value of g.

If the two modules are simultaneously stimulated with a pair of associated features the situation is simpler (see Fig. 4b). In this case, if the load parameter is below a critical line, the features are correctly retrieved in both modules for all values of g. Again the spin-glass phase appears for large α. Notice that the large g, small α region in Fig. 4a coincides with part of the retrieval phase in Fig. 4b.

One can consider whether the existence of multiple associations, where a given feature in one module is associated with s features in the other, can spoil the behavior of the network. These more complex associations are no doubt frequent in nature. However, not all the pairs of features contribute to the synaptic efficacies with the same strength, and it is likely that one of them dominates the others. The precise distribution of the strengths of these associations is of course not known. To analyze this question, we have considered an extreme case where all these associations have the same strength. Although this a limit situation, it will help us to determine if good retrieval properties are still possible under these more general condition. The answer is shown in Figs. 4 c,d, where we have taken $s = 3$.

If α and g are small, under stimulation of only one of the modules with one of its stored features (Fig. 4c), there appears again a phase where the feature is correctly retrieved. As g increases, the capacity of this phase also increases. The difference with the case $s = 1$ is that now there is another effect that competes with the correct signal; it is given by the contribution to the local field of all the features associated with the stimulus. Because of this, there appears a new phase (the *mixed* phase) where all of them are present and the corresponding overlaps are close to one. Let us notice that the coding rate of this attractor is not equal to $\sim f$ (the coding level of the *stimuli*), but to s times $\sim f$. As α grows, keeping g fixed, both phases destabilize into a spin-glass, where the state of the system is not correlated with the features.

When both modules are stimulated with a pair of associated features (Fig. 4d) the small α and small coupling region is a phase where each module reaches an attractor very close to the feature. Since there is activity in the whole network, this is a global phase. For large coupling, the state defined as the union of the two features used to stimulate the system becomes unstable. Because of the multiple associations, the final state is very close to the union of *all* these features (s per module). This is the same phase found in this region by stimulating a single module (Fig. 4c).

One can wonder how much the capacity properties of the network change when a persistent stimulus is applied. For this reason we have studied the behavior of the bi-modular network under clamped conditions. Now convergence to the attractor is achieved in the presence of a persistent external field applied to one of the input modules. We have computed the phase diagrams for the same parameter values used for the unclamped case, and $s = 3$ (i.e., for the cases shown in Figs. 4c-d). We have used external fields with intensity values up to $h = 10$. For $\theta = 0.3$ the results for clamped conditions show no qualitative changes, and negligible quantitative differences with the unclamped case presented in Fig. 4. We will come back to the analysis of clamped conditions in the discussion for $\theta = 0.6$. As we will see in a moment, for this value of θ the clamped conditions do produce a substantial change in the retrieval properties of the network.

We consider next the behavior of the bi-modular network for $\theta = 0.6$, which is closer to its optimal value at $g = 0$. The results are shown in Figs. 5a-c for unclamped conditions, and in Figs. 5b-d for clamped conditions, stimulating in both cases only one of the modules. The interesting effect for unclamped conditions is, apart from a capacity higher than for $\theta = 0.3$, that the global and the mixed phase are not reached from this initial condition. They are replaced by a null phase, where all the order parameters are zero [3]. This is shown in Fig 5a, for $s = 1$. This figure is to be compared with Fig. 4a: The nature of the small α, large g phase is completely different. The absence of the mixed phase when multiple associations are present can be seen in Fig.

[3]This null phase is similar to the 'non-opinionated' phase found in (Buhmann, Divko & Schulten, 1989).

5c, for $s = 3$ (where again only one of the modules has been stimulated). Apart from the expected retrieval phase at small α and g, and the non-retrieval phase at large α, here there appears again the null phase mentioned before. If one starts at a point inside the retrieval phase and increases the association strength keeping α fixed, the network falls into this regime, instead of falling into a mixed phase as in Fig. 4d. This can be seen as an advantage, in the sense that the network does not respond when it cannot decide which feature in the non-stimulated module has to be retrieved.

The effect of a persistent stimulus on the retrieval properties of the network at this quasi-optimal value of the threshold is even more remarkable. We have studied the effect of clamped conditions for both $s = 1$ (Fig. 5b) and $s = 3$ (Fig. 5d). The most interesting change is that in both cases the null phase disappears, its place being taken by the local retrieval phase. This is particularly important if one considers that, under normal natural conditions, clamped stimuli are more likely than unclamped stimuli. Again, the mixed phase does not appear.

Although not shown in the figure, we want to remark that if associated features were applied as stimuli to both modules, the attractors reached would be global.

The trend seen for these values of f, d_0 and d is that, when the threshold is close to its optimal value for $g = 0$ (~ 0.7) and only one module is stimulated, the critical capacities are large and the system is either in the local or in the non-retrieval phases. As the threshold decreases, the capacities become smaller and the global phases appear for g greater than a critical line.

From the results reported in this Section we can extract a conclusion about the appearance of global attractors in the network under stimulation of a single module: Global properties operate under relatively low threshold and moderately large g. The reason is that, as the threshold grows, the amount of current needed to induce sustained activity states in the non-stimulated module increases. Taking also into account that the strength of the intra-modular connections decreases relative to that of the inter-modular ones with increasing g, the destabilization of the LR phase into either the GR or the N phase depending on the value of the threshold, is readily understood.

It is also relevant that for unclamped conditions and large threshold the phase diagram at large g and small α is somewhat odd, while for the more relevant case of clamped conditions the whole low α region is occupied by the local retrieval phase. Noticing that the extra current due to the external field only affects neurons which are active in the stored pattern used as the stimulus, this effect is also understood: even if the current from the recurrent collaterals is very low, the *selective* contribution from the external field suffices to make the stimulated pattern stable. For low thresholds, the destabilization of the stimulated pattern is not due to a decrease of the strength of the current from inside the network but to an increase in the unselective noisy components of this current. Since the external field has no effect on this noise (it only slightly increases the signal), the phase diagrams for low threshold ($\theta = 0.3$) are unchanged by the persistence of the stimulus.

5.2 Analog neurons

In this subsection a more realistic network of neurons described in terms of their firing rates is studied. Instead of a full exploration of the influence of the various parameters in this case, we have concentrated on the phase diagram observed at biologically plausible values of the model parameters, and on the cooperation of the two modules under the influence of noisy external inputs.

There is now a new parameter, the gain G, which controls the value of the rates in the attractors. Firing rates observed experimentally (see references below) are well below saturation and, therefore, one would prefer a value less than one for the (normalized) rates computed with eq. (3). However, if the gain is too low the activity of the network cannot be sustained, and the system will fall into a null, silent phase, as defined in the last Section. Therefore, the gain was chosen using the criterion that the retrieval phases could be realized by the network. An adequate value was found to be $G = 1.3$. The dilution parameters d_0 and d are, as in the last subsection, equal to 0.1 and 0.05 respectively. A precise computation of the value of f from experimental results is not the subject of this study, and we selected the value $f = 0.22$, which is similar in magnitude to what is found by integrating the tail of typical spike rate distributions (Rolls & Tovee, 1995; Rolls et al., 1998; Treves et al., 1999). We do not expect substantial differences in our results if other similar value of f were chosen.

We have focused this part of the study on the performance of the system under clamped conditions. As explained before, this expresses the view that the stimulus will usually be persistent, and therefore will continue to influence the performance of the network during retrieval. In fact, one would expect that the magnitude of the stimulus varies with time, being large initially so as to serve as a cue for the network to find a memory close to the stimulus (if any), but rapidly decreasing to a low value which would persist during retrieval. This level of detail is beyond the scope of our work, so we have studied the fixed points of the bi-modular network with a persistent external field of constant magnitude. Its value has been estimated as the local field produced by the afferent connections from another module, at typical values of the connection strength, dilution and mean firing rates for the neurons in that external module. For the values of our parameters this gives a magnitude of $h \sim 0.05$.

The overall scale of the threshold is determined by the intensity of the external field, since a sub-threshold stimulus is not noticed by the network. Since we are interested in the analysis of transitions between local and global retrieval phases, we chose values of the threshold in the region where these transitions occur. This corresponds to $\theta \sim 0.02$.

In Fig. 6 we present the phase diagram of the system computed under the conditions just explained. To obtain this diagram, one of the modules (say **A**) was stimulated with a persistent external field close to one of its stored patterns and of intensity $h = 0.05$. The different patterns of sustained activity were analyzed as a function of the association strength g and of the storage level α.

In the first place, the characterization of retrieval and non-retrieval states is slightly different from the binary case. For a given module to be in a state of retrieval, two conditions have to be met: that the mean rates in the foreground and the background populations be different *and* that the rate distributions in the two populations do not overlap (Amit & Tsodyks, 1991).

The non-retrieval (SG) phase for a given module is characterized by similar (though not necessarily equal) mean rates in the foreground and the background and by highly overlapping rate distributions. Rate distributions of states representative of these phases are shown in Fig. 7.

As is shown in Fig. 6, if α is not too high there exists a region of local retrieval (LR) for low values of the association strength g. In this phase, the stimulated module, **A**, is in a state of retrieval while the other one, **B**, is in a low activity state similar to the one found for binary neurons, characterized by very small rates (e.g. less that 10^{-3}) for a fraction of the neurons in the foreground population. Although those values are clearly not interpretable in terms of spike emission, they reflect the fact that module **B** in this region is receiving a very weak signal from the stimulated module. However, this signal is in the correct direction, producing activity only in neurons which are active in the correct stored pattern. This is a favourable situation, because again, a small value of g increases (although slightly) the storage capacity with respect to its value at $g = 0$. Fixing α at a value smaller than ~ 0.33, one observes that as g grows, the LR phase is no longer stable, and the system enters into a global regime GR in which stimulation of only one of the modules produces sustained activity in both of them.

In the GR phase, both modules are in retrieval, but neurons in module **B** are, on average, firing at lower rates. This is because the effective current they are receiving does not contain the contribution from the persistent external field [4]. Let us finally remark that in spite of this effect, the mean rates in both modules approach each other as g grows.

As one would expect, if α is large enough, both the LR and the GR phases become unstable, and the system enters the non retrieval regime or SG phase. Since this transition is usually discontinuous, the passage from retrieval to spin-glass states is *unambiguous*. However, there is a region (for α just above 0.33 and $g \sim 0.15$), where the transition from the retrieval phases to the SG regime is not well defined. In this small region, the non stimulated module **B** goes into the SG phase while the stimulated module **A** enters into what could be called a *poor retrieval* regime (see Fig. 7b), which persists until $\alpha \sim 0.36$. In this regime, although the rate distributions for the foreground and background populations are well differentiated, they overlap significantly. What is happening is that module **B** destabilizes first. Therefore, the backprojected input from this module is not correlated with the memories of module **A** anymore, worsening the retrieval quality of its persistent states. As α grows, this retrieval quality falls gradually and the states become usual non retrieval states (see Fig. 7c). Since the transition from this region to the SG phase is not

[4] Incidentally, the asymmetry observed between the foreground and background populations in the SG phase (see Fig. 7c) is also due to the persistence of the external field, which discriminates between neurons in the two populations.

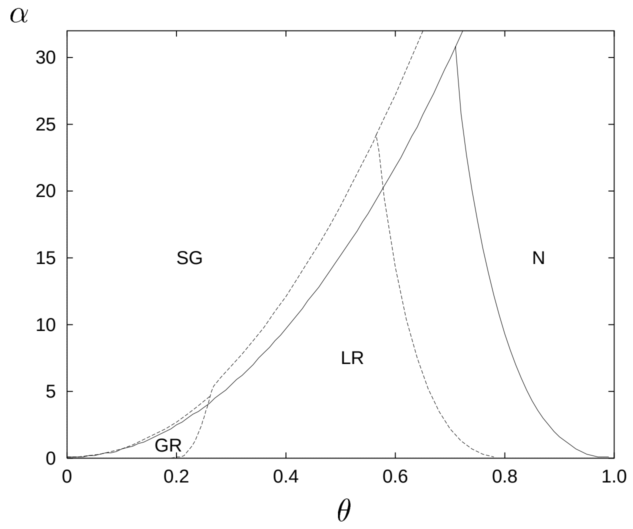

Figure 3: Critical capacity vs neural threshold for one and two modules. Neurons are binary and the parameter values are: $f = 0.001$, $d_0 = 0.1$ and $s = 1$ for one module, and also $d = 0.05$ and $g = 0.5$ for the coupled modules. The solid line is for one module and the dotted line is for two. The figure was produced by finding the critical capacity (in the unclamped condition) produced by initial stimulation of one of the modules for several values of the threshold. In the two cases the threshold is crucial in determining the stability of the retrieval phase. The highest capacity for the single module network is achieved for $\theta \sim 0.7$. The curve for the bi-modular network is similar to the one for a single module, although slightly shifted to the left. This is because when $g > 0$ the overall magnitude of the intra-modular connections decreases. **LR** and **GR** denote the local and the global retrieval phases. In the **LR** phase only the stimulated module achieves complete retrieval, while in **GR** both modules are active with consistent features. **SG** is a non-retrieval phase; the activity here is not correlated with any feature. **N** is a null phase of no activity. In this phase all neurons in the bi-modular network are silent.

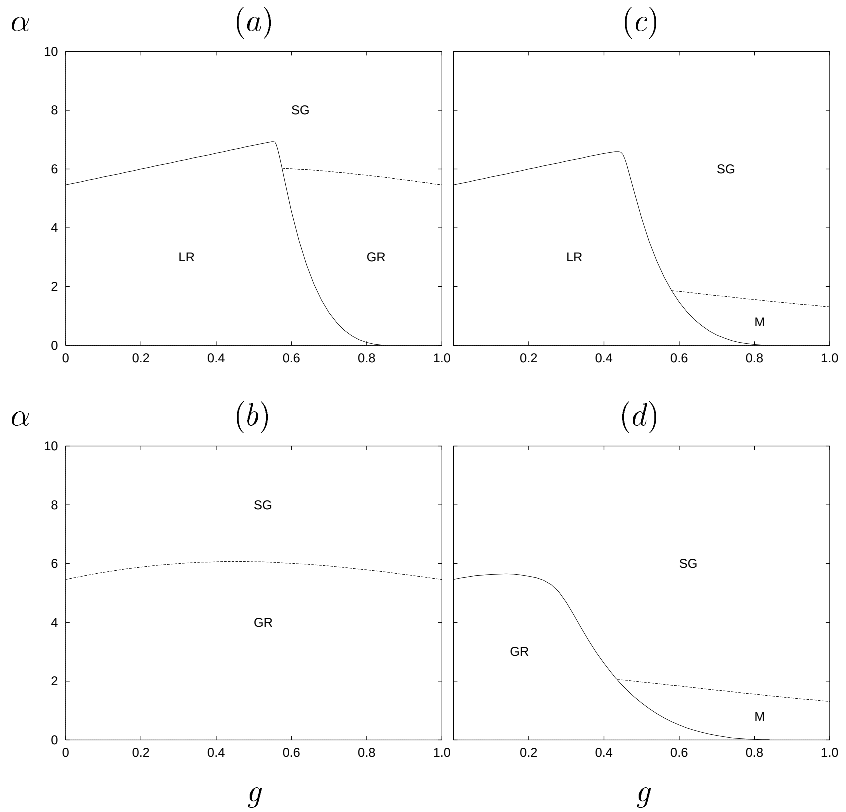

Figure 4: The retrieval diagrams for binary neurons. The parameters are: $\theta = 0.3$, $f = 0.001$, $d_0 = 0.1$, $d = d_0/2$ and different initial conditions. The figures on the left (**a** - **b**) are for $s = 1$. The ones on the right (**c** - **d**) are for $s = 3$. In **a** and **c** only one module has been stimulated initially. In **b** and **d** the two modules have been equally stimulated with a pair of associated features. In region **M** the coupling is so strong that the activity state is the union of the complete set of associated features. This is the mixed phase.

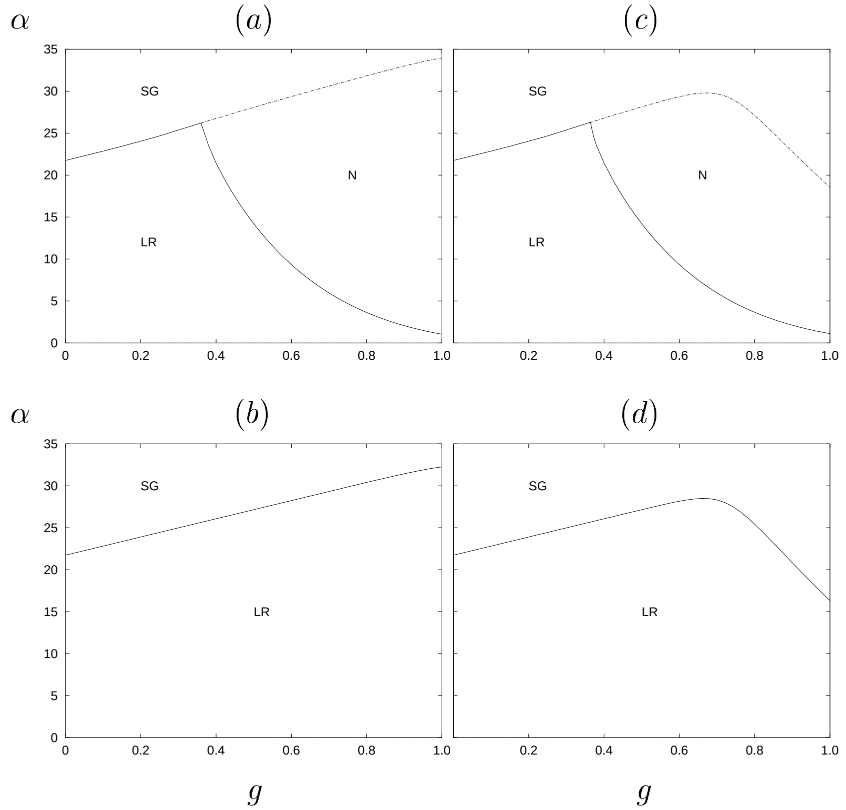

Figure 5: Retrieval diagrams for two modules at $\theta = 0.6$. In (**a**) and (**b**) $s = 1$ (left), and in (**c**) and (**d**) $s = 3$ (right). In the four cases the stimulus, taken as one of the stored features, is applied to only one of the modulus with intensity $h = 1$ but the two diagrams on the top ((**a**) and (**c**)) are obtained in *unclamped* conditions, whereas in the two on the bottom ((**b**) and (**d**)) the stimulus is persistent. The rest of the parameters are set as in Fig.4. For this value of the threshold the persistence of the stimulus is critical: Both for $s = 1$ and for $s = 3$ the **N** phase disappears, its place being taken by the **LR** phase which is now stable up to $g = 1$ and for very large values of α. Still, the loading where the transition to the **SG** phase occurs is almost independent of the persistence of the stimulus.

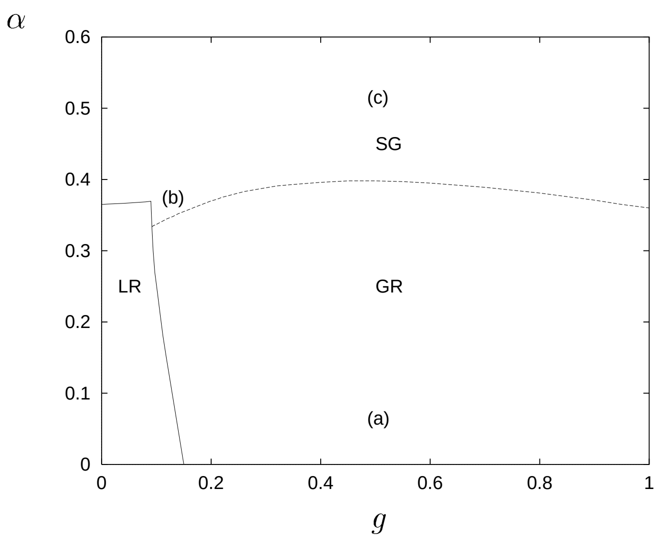

Figure 6: Retrieval diagram for analog neurons with the current-to-rate transduction function given by eq. (3). The values of the model parameters are: $\theta = 0.02$, $f = 0.22$, $d_0 = 0.1$, $d = d_0/2$ and the gain is $G = 1.3$. Only one of the modules has been stimulated with a persistent field equal to one of the stored features and strength $h = 0.05$. The **LR** phase is now stable only for rather small values of g, and the capacity of both retrieval phases does not depend strongly on the value of this parameter. The distributions of rates in the points labelled (a), (b) and (c) are shown in Fig. 7a, b and c respectively.

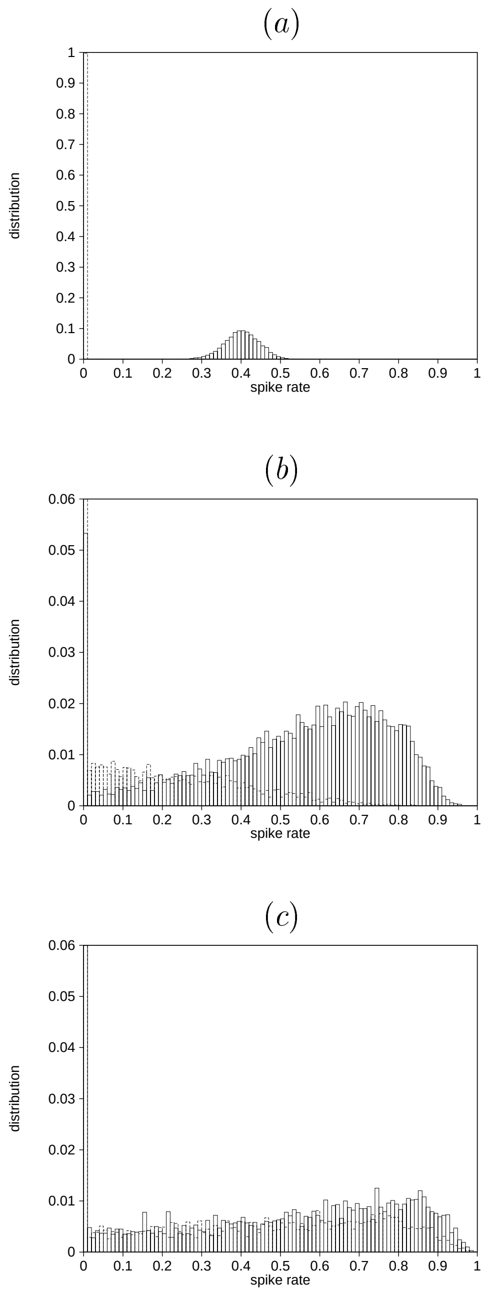

Figure 7: Distributions of spike rates in the stimulated module at the three points labelled as (a), (b) and (c) in the phase diagram given in Fig. 6. Solid and dashed lines correspond to foreground and background neural populations respectively, and the bin width is 0.01. Case (**a**) corresponds to a **GR** state located at $g = 0.5$ and $\alpha = 0.05$. The two distributions do not overlap. Case (**b**) corresponds to a *poor retrieval* state at $g = 0.125$ and $\alpha = 0.36$ (i.e. very close to the transition between the **LR** and the **SG** phases). The scale was chosen to facilitate the comparison between the two populations. The zero rate bin height of the background population lies outside the frame and is approximately 0.7. Note the significant overlap between the two distributions. Case (**c**) corresponds to a non-retrieval state situated at $g = 0.5$ and $\alpha = 0.5$. The two populations are almost indistinguishable. Again, the scale was chosen to facilitate comparison. The zero rate bin heights for the foreground and background populations are 0.38 and 0.56 respectively. This means that $\sim 45\%$ of the neurons in the background are active, and that a little less than 40% of the neurons in the foreground are silent. The system has failed to retrieve the pattern.

sharp, we have not included it as a separate phase in the phase diagram.

Our last result concerns the error correcting capabilities of the bi-modular network. We have addressed in this work the following important general question: Can the existence of structured associations between cortical modules improve the retrieval capabilities of one of them when it works under noisy conditions?

In order to answer this question, we have studied a situation in which one of the modules (**A**) was stimulated with a persistent external field equal to a distorted version of one of its stored patterns, while the other was stimulated with the (correct) pattern associated with it and stored in this module (**B**). The intensity of the external stimuli applied to both modules was $h = 0.025$. The procedure was repeated for several values of the inter-modular association strength and for different levels of distortion for the stimulus on module **A**, keeping the storage level fixed to a value $\alpha = 0.15$. One would expect that since module **B** is being stimulated with the correct pattern associated with the distorted one in **A**, this will allow module **A** to retrieve in conditions in which, by itself, retrieval would be impossible. The results are shown in Fig. 8 where the rest of the model parameters have the same values as in Fig. 6. As mentioned in Section 2, we will measure the amount of distortion of a given feature by the overlap between the distorted and correct versions of it, as defined in eq. (9). Since the stimulus applied to **A** is a distorted version of one of the patterns stored in that module, we will denote this overlap as $m_{\mathbf{A}}(\delta)$, where δ measures the probability that the stimulus contains an error. The line drawn in Fig. 8 represents the minimum overlap necessary for module **A** to retrieve the distorted pattern correctly.

For a given value of g, if the overlap of the stimulus on **A** with the pattern to be retrieved ($m_{\mathbf{A}}(\delta)$) lies below the line, the sustained activity state of this module shows a large number of errors. The stimulus is too far from the stored pattern and the network is unable to retrieve it. On the other hand, module **B** is in retrieval in this region. Since only module **B** is in retrieval this is a local retrieval phase. At

$g = 0$ (isolated modules) the maximum amount of distortion allowed by module **A** in order to still be able to perform retrieval is $m_{\mathbf{A}}(\delta) \sim 0.65$.

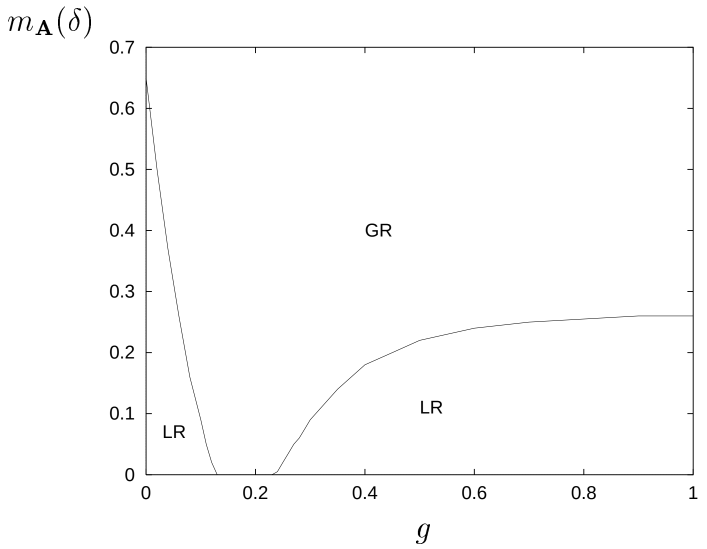

Figure 8: Critical line for overlap of the stimulus applied to module **A** with the pattern to be retrieved, as a function of g. The load parameter is $\alpha = 0.15$. The intensity of the stimuli applied to both modules was set to $h = 0.025$, but the stimulus applied to module **B** contained no errors. As the overlap increases above the critical line, module **A** discontinuously enters retrieval, while module **B** is in retrieval all the time. The maximum amount of distortion errors in the retrieval cue consistent with **A** being in correct retrieval increases rapidly with g (for g small) and then decreases steadily for $g > 0.25$. Note that module **B** *always* helps module **A** for all values of g.

For a fixed amount of distortion within the LR phase, as g starts to increase, some of the errors in the stimulus on module **A** are corrected and eventually a point is reached at which the state of this module changes discontinuously into a state of retrieval. Since both modules are now in retrieval this is a global retrieval phase. The critical value of the overlap at which this transition occurs decreases very rapidly with g and even becomes zero for $g \sim 0.15$. In such a situation, a persistent stimulus of intensity $h = 0.025$ totally uncorrelated with the pattern is sufficient, thanks to the strong and selective signal coming from the other module, to elicit the correct response in this module. Or alternatively, one of the modules is able to converge to an attractor corresponding to a certain feature, even if it is being persistently stimulated with a purely noisy external input, with the condition that the other module is persistently stimulated with the correct feature associated with the first one in the inter-modular synaptic matrix.

As g becomes larger, the critical overlap starts to increase, reaching a maximum value of ~ 0.27. This increase is probably due to the weaker signal coming from module **B** because of the change in the relative value of the inter and intra-modular connections as g changes.

Let us finally remark that the maximum critical overlap in the whole range of g is reached at $g = 0$. Therefore, retrieval under noisy conditions is always improved (and sometimes impressively) by the interaction between the modules.

6 Discussion

A general model for coupled attractor neural networks with features of biological realism has been proposed. To our knowlegde, there has been no previous analytical treatment of a multi-modular network composed of a finite number of modules. The model incorporates a free parameter measuring the relative intensity of the inter- versus intra-modular connections whose importance in determining the retrieval state of the network is demonstrated. Other free parameters of the model are the connectivities of the inter- and intra-modular connections, and the coding level of the stored patterns. Results are presented for networks composed of binary neurons or analog neurons with a hyperbolic transduction function.

The analysis of the system focused on its performance as an associative memory. A possible way of modelling such a device is to set up a network in which local patterns of activity are stored in the connections inside each module, and where specific associations between these local features are stored in the connections between the modules. The bi-modular architecture of the model network is meant to capture this idea. Since active association of partial representations is such a general principle in cognition, it should be a fairly robust property. We show that a simple and general network of biological plausibility is able to perform active association during retrieval, in a wide range of parameters and stimulation conditions.

The analysis has been carried out from two different perspectives. First, in a network of binary neurons, a systematical study of the influence of some of the parameters on the possible regimes of operation of the network has been acomplished. The use of the simpler binary units in this case has allowed a more complete exploration of the parameter space and direct comparison with previous results on capacity issues on uni-modular binary networks. On the other hand, the influence of the inter-modular association strength and the capacity of a more realistic network of analog neurons with model parameters set to biologically plausible values has also been studied to check that both local and global retrieval phases can be achieved in these conditions. Once this was confirmed, the important issue of retrieval in the presence of noisy stimuli was studied in the realistic bi-modular network.

In both approaches, the nature of the retrieval states as a function of the inter-modular connection strength g and of the capacity of the network α was analyzed, and special attention was placed on the possible transitions between local and global activity patterns and on the effect of persistent stimuli on the behavior of the system.

We were able to identify a global retrieval regime. For this the conditions were that the inter-modular connections have to be large, and the threshold of the neurons has to be relatively low. Of interest here is that the total number of memories that can be stored in the whole system operating in this global way is of the same order as the number of memories that can be stored in any one of the modules using the recurrent collaterals. Thus in the global regime, each module does not contribute independently to the total number of memories that can be stored in the network. The whole network stores a number of memories that is proportional to the effective number of connections per neuron.

There is an interesting effect observed in the local phase when only one module receives a retrieval cue (this was more evident in the case of the network of binary neurons): the number of patterns that can be retrieved from the stimulated module increases gradually as the coupling strength between the two modules increases (see e.g. Fig. 4). This is due to partial retrieval in the other module, which in turn facilitates better retrieval in the stimulated module. We emphasize though, that even when this occurs, there is very incomplete retrieval in the second module.

In the same regime (weak inter-modular connection strengths), if both modules are stimulated with corresponding inputs (those originally paired during "learning"), the same global retrieval phase referred to above is reached.

In the case of the *binary network*, different values of the threshold were investigated, during both clamped and unclamped retrieval. What we have discussed so far applies with moderate to low thresholds. If the thresholds are higher (see Fig. 5), there is no global retrieval phase if inputs are applied to only one module. If corresponding inputs are applied to both modules, there will be a global retrieval phase. With these higher thresholds, under clamped conditions the local retrieval regime covers the whole range of inter-module coupling values of g with low α. The capacity in the high threshold regime is again proportional to the number of connections per neuron, although the actual number of patterns that can be retrieved is closer to the optimal (for a sparseness of $f = 0.001$) because the critical value of α is higher. In particular, for a value of the threshold such as 0.6, α is in the order of $20 - 30$, whereas in the low threshold regime considered in Fig. 4 the critical value of α is in the order of 6. That is, one can store approximately 5 times as many memories in the high threshold case.

For the *analog network*, since the coding rate (i.e. sparseness) ($f = 0.22$) is not small, the network does not reach its optimal storage capacity. A value for the threshold $\theta = 0.02$ was studied in detail (see Fig. 8). A local retrieval (LR) phase exists for small inter-modular connection strength (g small). However, a large portion of the low α region is occupied by the global retrieval (GR) phase. The effect of the persistent stimulus is to make the global and the non-retrieval (SG) phases asymmetric, increasing the firing rates in the stimulated module with respect to those in the non-stimulated part of the network.

The effect of processing under noisy conditions on the performance of the coupled analog network was also studied. It was shown (see Fig. 8) that a module can retrieve correctly even with very noisy patterns if the other module is persistently stimulated with the correct version of the associated pattern. This is important since it is probably a common situation, at least if the distortion of the stimuli is small, and since it is one in which the interaction between the modules clearly improves the performance of the isolated modules.

The properties of the multi-modular system studied here seem to be sufficiently robust so as expect them to be maintained in more realistic conditions. For example, we anticipate that the same classes of memory performance we have described here would occur if there were a whole series of connected modules, as happens for example in cortico-cortical processing in vision, or in cortico-hippocampal connection circuits (Rolls & Treves, 1998).

The properties of the interconnected modules described here also suggest that forward and backprojections between adjacent cortical modules may serve as a way to implement complex associations between the different aspects of the stimuli being processed simultaneously, or to implement top-down constraints on earlier processing. For example, a high-level hypothesis about what we expect to see might influence early visual processing by operating in the way we have described.

There are still many open questions and many ways in which the analysis of the function and operation of recurrent connections both within and between modules in the cortex could be improved: more realistic models for the neurons in the network, a separate treatment of the inhibition, the inclusion of spontaneous activity states, and more general and complex architectures including for example convergence from modules at one level of processing to a single module at a higher level of processing, are just some examples of these.

In fact, although only a bi-modular architecture has been studied in this paper, the solution can be found for an arbitrary architecture with any number of modules, with the only constraint that connections between the neurons be symmetric. Even though the analysis makes this assumption, it is likely that the general results will generalise to comparable architectures with asymmetric connectivity.

A model of this type is presently being applied to the study of multi-modal sensory areas with several interacting modules (Renart, Parga & Rolls, 1998) that, we hope, will clarify and provide quantitative insight into the function and operation of backprojections in the cerebral cortex and other brain systems with reciprocally connected modules.

Acknowledgements

This research was partly supported by a British Council-Spanish Ministry of Education and Science bilateral program HB 96-46, and by Medical Research Council Programme Grant PG8513790 to E.T.Rolls. A Spanish grant PB96-47 is also acknowledged. Two of the authors (AR and NP) are most appreciative for the hospitality shown to them while visiting the Department of Experimental Psychology, Oxford University, during the completion of this work.

Appendix A: The replica technique

The replica method has been developed (Edwards & Anderson, 1975) as a way to compute the free energy of disordered systems (16). It has been widely used, in particular in models similar to ours for a single module network (Amit, 1989; Kuhn, 1990; Amit & Tsodyks, 1991; Treves & Rolls, 1991).

The difficulty of the problem is that since the disorder is quenched, the average over the stored features has to be taken on the partition function itself. To overcome this difficulty the method employs the identity

$$\log \mathcal{Z} = \lim_{n \to 0} \frac{\mathcal{Z}^n - 1}{n} \quad , \tag{28}$$

which reduces the problem to that of calculating $\ll \mathcal{Z}^n \gg$. Precisely, (16) can be expressed as

$$\mathcal{F} = - \lim_{N \to \infty} \lim_{n \to 0} \frac{1}{\beta \, n \, MN} (\ll \mathcal{Z}^n \gg -1) \quad . \tag{29}$$

The procedure for this calculation can be split in two parts: first, one has to assume that n is a natural number, and then calculate the partition function of n copies or *replicas* of the original system, all with the same couplings. The free energy obtained in this way is a function of a set of order parameters and, eventually, of the microscopic state of every replica. Second, one resorts to a kind of analytic continuation and takes the limit $n \to 0$. There are several ways to accomplish this. The one we have used is the replica-symmetry ansatz, which consists in assuming that the state of the system does not depend on the replica chosen. With this assumption the limits in (29) are well behaved although the order in which they are taken has to be interchanged. In fact, one first performs the $N \to \infty$ limit of the partition function using the saddle point method, which consists of approximating, when N is very large, the integral of the exponential of an extensive function (proportional to N), by the exponential of N-times the extremum of the function with respect to the integration variable. Then, from eq. (29), one obtains the free energy for n replicas at large N, as the function in the exponent evaluated at its extremum. In this way, the free energy can be explicitly calculated.

It has been shown (see e.g. Mezard, Parisi & Virasoro, 1987) that for most systems this approximation is not exact and that replica-symmetry is in fact broken at low temperatures. However, even at large β, differences with the replica symmetric theory may be small.

Appendix B: Treatment of the dilution

The treatment of the random dilution of the synapses that we have used, follows the one proposed by Sompolinsky in (Sompolinsky, 1986; Sompolinsky, 1987). The idea is to consider the diluted connections as having a constant term, equal to the mean value of the connections over the dilution variable, plus a fluctuating component modelled as a gaussian noise. The synaptic matrix defined in Section 2 can then be expressed as:

$$J_{ij}^{(a,a)} \; = \; [J_{ij}^{(a,a)}] \; + \; \delta J_{ij}^{(a,a)} \quad , \tag{30}$$

$$J_{ij}^{(a,b)} \; = \; [J_{ij}^{(a,b)}] \; + \; \delta J_{ij}^{(a,b)} \qquad a \neq b \quad , \tag{31}$$

where $[\ldots]$ denotes the average over the dilution variables d_{ij}^{ab} and d_{ij}^{0}. The second terms on the r.h.s. of eqs. (30,31) represent the fluctuating components of the intra- and inter-modular connections respectively. Their values are:

$$[J_{ij}^{(a,b)}] \; = \; \frac{1}{\chi N} \sum_{\mu,\nu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f) K_{\mu\nu}^{ab} (\eta_{bj}^{\beta\nu} - f) \qquad\qquad ai \neq bj \quad , \tag{32}$$

where $K = [\tilde{K}]$. Its elements are

$$K_{\mu\nu}^{ab} = \frac{d_0}{\Lambda} \left(\delta^{ab} \otimes \delta_{\mu\nu} \right) + \frac{gd}{\Lambda} \left((1^{ab} - \delta^{ab}) \otimes 1_{\mu\nu} \right) \quad . \tag{33}$$

The random fluctuations $\delta J_{ij}^{(a,b)}$ and $\delta J_{ij}^{(a)}$ are given by

$$\delta J_{ij}^{(a,b)} = \frac{g(d_{ij}^{ab} - d)}{\chi N_t} \sum_{\mu,\nu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f)(\eta_{bj}^{\beta\nu} - f) \tag{34}$$

$$\delta J_{ij}^{(a,a)} = \frac{(d_{ij}^{0} - d_0)}{\chi N_t} \sum_{\mu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f)(\eta_{aj}^{\beta\mu} - f) \tag{35}$$

Since these are the sum of many independent random numbers, they can be considered gaussian random variables with means and variances obtained from eqns (55-35). Of course, $[\delta J_{ij}^{(a,b)}] = [\delta J_{ij}^{(a,a)}] = 0$ even for a given realization of the patterns. As for the variances, defining

$$[(\delta J_{ij}^{(ab)})^2] \equiv \frac{\Delta_{ab}^2}{N} \tag{36}$$

$$[(\delta J_{ij}^{(a,a)})^2] \equiv \frac{\Delta_a^{(0)\,2}}{N} \quad , \tag{37}$$

one finds that

$$\Delta_{ab}^2 = \frac{g^2 d(1-d)\alpha s}{\Lambda} \tag{38}$$

$$\Delta_a^{(0)\,2} = \frac{d_0(1-d_0)\alpha}{\Lambda} \quad . \tag{39}$$

Therefore, the calculations presented in this work have been done for a synaptic matrix given by

$$J_{ij}^{(a,a)} = \frac{1}{\chi N} \sum_{\mu,\nu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f) K_{\mu\nu}^{aa} (\eta_{aj}^{\beta\nu} - f) + \delta_{ij}^{0\,(a)} \quad i \neq j, \tag{40}$$

$$J_{ij}^{(a,b)} = \frac{1}{\chi N} \sum_{\mu,\nu=1}^{s} \sum_{\beta=1}^{L} (\eta_{ai}^{\beta\mu} - f) K_{\mu\nu}^{ab} (\eta_{bj}^{\beta\nu} - f) + \delta_{ij}^{(ab)} \quad a \neq b, \tag{41}$$

where $\delta_{ij}^{(ab)}$ and $\delta_{ij}^{0\,(a)}$ are defined as (quenched) gaussian random variables of zero mean and variances given by (36) and (37) respectively. In order for the J's to be symmetric, the variables $\delta_{ij}^{0\,(a)}$ and $\delta_{ji}^{0\,(a)}$ are not drawn independently from their distributions, but are set equal *by hand.* A similar construction holds for the variables $\delta_{ij}^{(ab)}$.

Appendix C: Self-consistency equations for binary neurons

The conversion to binary neurons has to be made before the zero temperature limit is taken. In order to do this, one first has to specify the measure of integration so that only the values zero and one for the rates are considered. This is achieved by setting

$$d\rho(\nu_a) = d\nu_a \left[\delta(\nu_a - 1) + \delta(\nu_a)\right] \quad , \tag{42}$$

so that the integral over the rates becomes a trace. One also has to take the infinite gain limit of the transduction function. When this is done, the last term in eq. (18) becomes the threshold in eq. (3) and the hyperbolic transduction function becomes a step function discontinuous at a value of the rate equal to this threshold. Once the trace has been done the zero temperature limit is readily taken. At zero temperature, the fixed point equations for the order parameters read:

$$\begin{aligned}
m_a^\mu &= \frac{1}{2\chi} \ll (\eta_a^\mu - f)\left(1 + erf\left(\frac{A_a}{\sqrt{2}B_a}\right)\right) \gg_{\eta\xi} && (43)\\
q_a &= \frac{1}{2} \ll \left(1 + erf\left(\frac{A_a}{\sqrt{2}B_a}\right)\right) \gg_{\eta\xi} && (44)\\
c_a &= \frac{1}{\sqrt{2\pi}\ B_a} \ll \exp\left[-\left(\frac{A_a}{\sqrt{2}B_a}\right)^2\right] \gg_{\eta\xi} && (45)\\
\alpha r_a &= \frac{\alpha\Lambda}{s}\frac{\partial}{\partial c_a} Tr\left[Q_{\mu\nu}^{ab}\left(\delta_{\mu\nu}\otimes\delta^{ab} - C_{\mu\nu}^{ab}\right)^{-1}\right] && (46)\\
\alpha \bar{c}_a &= \frac{\alpha\Lambda}{s}\frac{\partial}{\partial q_a} Tr\left[Q_{\mu\nu}^{ab}\left(\delta_{\mu\nu}\otimes\delta^{ab} - C_{\mu\nu}^{ab}\right)^{-1}\right] \quad , && (47)
\end{aligned}$$

where we have defined

$$C_{\mu\nu}^{ab} = \sum_{\tau\ c} c_a(\delta^{ac}\otimes\delta_{\mu\tau})K_{\tau\nu}^{cb} \quad . \tag{48}$$

and where $erf(x)$ is the error function defined as:

$$erf(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x \exp(-u^2)du \quad .$$

The quantities A_a and B_a are

$$A_a = \sum_\mu (\eta_a^\mu - f) \left(\sum_{b\nu} K_{\mu\nu}^{ab} m_b^\nu \right) + \sum_\mu h_a^\mu(\eta,\xi) + \frac{\alpha}{2}(\bar{c}_a - d_0) + + \frac{1}{2} \left(\Delta_a^{(0)\,2} c_a + \sum_{b\neq a} \Delta_{ab}^2 c_b \right) - \theta \tag{49}$$

$$B_a = \sqrt{\alpha r_a + \Delta_a^{(0)\,2} q_a + \sum_{b\neq a} \Delta_{ab}^2 q_b} \ . \tag{50}$$

Note that the equations for r_a and $\bar{c}_a$ do not have the form of self-consistency equations. Instead, they relate them to the *other* order parameters through algebraic expressions. Taking this into account, the system depends effectively only on the m's, the q's and the c's, so these parameters are, in this sense, fundamental.

Appendix D: Self-consistency equations for analog neurons

In this case the measure of integration $d\rho(\nu_a)$ is chosen to be uniform. However, when the zero temperature limit is taken, the only rates which give a finite contribution to the free energy are those that minimize (18). Setting its derivative with respect to ν_a equal to zero determines the value of the rate in the attractor, which we will call $\tilde{\nu_a}$, through the following self-consistency equation:

$$\begin{aligned} \tilde{\nu_a}(z,\eta,\xi) = \phi \Bigg\{ & \sum_\mu (\eta_a^\mu - f) \left(\sum_{b\nu} K_{\mu\nu}^{ab} m_b^\nu \right) + \sum_\mu h_a^\mu(\eta,\xi) + \\ & + z\sqrt{\alpha r_a + \Delta_a^{(0)\,2} q_a + \sum_{b\neq a} \Delta_{ab}^2 q_b} \ + \\ & + \tilde{\nu_a}(z,\eta,\xi) \left[\alpha(\bar{c}_a - d_0) + \Delta_a^{(0)\,2} c_a + \sum_{b\neq a} \Delta_{ab}^2 c_b \right] \Bigg\} \ . \end{aligned} \tag{51}$$

The argument of the transduction function, ϕ, represents the effective current present in the attractor. The first two terms are signal contributions coming from the patterns being retrieved and from the external stimuli. The fluctuating term in the second line represents the noise generated by the random overlaps between the pattern(s) being retrieved and the (extensively many) others, and by the random dilution of the synapses. The terms in the third line represent a contribution to the effective current coming from the correlation of the attractor with the noise just described.

The self-consistency equations for the order parameters are:

$$
\begin{aligned}
m_a^\mu &= \frac{1}{\chi} \ll (\eta_a^\mu - f)\tilde{\nu_a} \gg_{\eta,\xi,z} && (52)\\
q_a &= \ll \tilde{\nu_a}^2 \gg_{\eta,\xi,z} && (53)\\
c_a &= \frac{\ll z\,\tilde{\nu_a} \gg_{\eta,\xi,z}}{\sqrt{\alpha r_a + \Delta_a^{(0)\,2} q_a + \sum_{b\neq a} \Delta_{ab}^2 q_b}} \quad , && (54)
\end{aligned}
$$

and the expressions for r_a and $\bar{c}_a$ are identical to the binary case. Following Amit & Tsodyks (1991), it is useful to express the m_a^μ's as

$$
m_a^\mu = \nu_{a_+} - \nu_{a_0} \qquad (55)
$$

where

$$
\begin{aligned}
\nu_{a_+} &= \frac{1}{fN} \ll \sum_i \eta_{ai}^\mu \langle \nu_{ai} \rangle \gg_{\eta,\xi} && (56)\\
\nu_{a_0} &= \frac{1}{(1-f)N} \ll \sum_i (1-\eta_{ai}^\mu)\langle \nu_{ai} \rangle \gg_{\eta,\xi} \quad . && (57)
\end{aligned}
$$

These quantities give the mean rate in the population of neurons active (to be referred to as foreground) and silent (to be referred to as background) in the pattern η_a^μ respectively. Although we do not give them explicitly, the mean field equations for these magnitudes can be easily obtained from the self-consistency equation for m_a^μ by performing the average over the pattern.

As noticed in Amit & Tsodyks (1991), the overlaps are not enough to characterize the state of the network. This is because they do not carry any information about the spatial distribution of rates. In order to distinguish a uniform from a non-uniform distribution in each population, one uses the quantities:

$$
\begin{aligned}
q_{a_+} &= \frac{1}{fN} \ll \sum_i \eta^{\mu}_{ai} \langle \nu_{ai} \rangle^2 \gg_{\eta,\xi} && (58)\\
q_{a_0} &= \frac{1}{(1-f)N} \ll \sum_i (1-\eta^{\mu}_{ai}) \langle \nu_{ai} \rangle^2 \gg_{\eta,\xi} \quad . && (59)
\end{aligned}
$$

The parameter q_a introduced in Sec. 3 is the average of q_{a_+} and q_{a_0} over the two populations: $q_a = f\ q_{a_+} + (1-f)\ q_{a_0}$.

The meaning of c_a is clear from eq. (54): it is the normalized overlap of the attractor with the noise generated by both the large number of stored patterns and the random dilution of the synapses. It will be small when the system is driven by the signal and it will increase as it becomes to be driven by the noise. It is interesting to note that if c_a vanishes for all modules, then the term proportional to the rate in equation (51) also vanishes. The interpretation of this term given in the text follows from this observation.

Again following Amit & Tsodyks (1991), one can obtain the rate distribution in the attractor by identifying the rates obtained for each realization of the quenched variables in (51) with the rates of the neurons in the network. The effective current in (51) computed for $\eta = 0$ can be interpreted as the current afferent to neurons in the background population. Conversely, if it is computed for $\eta = 1$, it gives the current in the foreground. These two currents depend on the stochastic variables z and $\xi \equiv (\xi_0, \xi_1)$, a dependence that in turn gives a distribution of rates inside each of the two populations. These are:

$$
Pr_{+,0}(\nu_a) = \sum_{\xi_0,\xi_1=0,1} \int_{-\infty}^{\infty} Pr(z,\xi_0,\xi_1)\delta(\tilde{\nu}_{a_{+,0}}(z,\xi_0,\xi_1) - \nu_a) dz \qquad (60)
$$

where $\tilde{\nu}_{a_{+,0}}(z, \xi_0, \xi_1)$ is just (51) with the η_a^μ's in the right hand side substituted by 1 and 0 respectively, and $Pr(z, \xi_0, \xi_1)$ is the compound probability distribution of the three random variables z, ξ_0 and ξ_1.

Bibliography

Amaral, D.G. (1986). Amygdalohippocampal and amygdalocortical projections in the primate brain. In Schwarz, R. & Ben-Ari, Y. (Eds.), *Excitatory Amino Acids and Epilepsy* (pp. 3-17). New York: Plenum.

Amaral, D.G. (1987). Memory: Anatomical organization of candidate brain regions. In *Handbook of Neurophysiology - The Nervous System* Washington D.C.: Am. Physiol. Soc.

Amaral, D.G. & Price, J.L. (1984). Amigdalo-cortical projections in the monkey (*Macaca fascicularis*). *Journal of Comp. Neurol., 230*, 465-496.

Amit, D. (1989). *Modelling Brain Function.* Cambridge: Cambridge University Press.

Amit, D. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences, 18*, 617-657.

Amit, D., Brunel, N. & Tsodyks M.V. (1994). Correlations of cortical Hebbian reverberations: experiments vs theory. *J. Neuroscience, 14*, 6445.

Amit, D., Parisi, G. & Nicolis, S. (1990). Neural potentials as stimuli for attractor neural networks. *Network, 1*, 75-88.

Amit, D. & Tsodyks, M.V. (1991). Quantitative study of attractor neural network retrieving at low spikes rates: II. Low-rate retrieval in symmetric networks. *Network, 2*, 275-294.

Braitenberg, V. & Schuz, A. (1991). *Anatomy of the Cortex.* Berlin: Springer-Verlag.

Buhmann, J., Divko, R., & Schulten, K. (1989). Associative memory with high information content. *Phys. Rev., A39*, 2689-2692.

Edwards, S.F. & Anderson, P.W. (1975). Theory of spin glasses. *J. Phys., F5*, 965-974.

Engel, A., Bouten, M., Komoda, A. & Serneels, R. (1990). Enlarged basin of attraction in neural networks with persistent stimuli. *Phys. Rev., A42*, 4998-5005.

Grieve K. & Sillito A. (1995). Non-length-tuned cells in layers II/III and IV of the visual cortex: the effect of blockade of layer VI on responses to stimuli of different lengths. *Experimental Brain Research, 104*, 12-20.

van Hoesen, G.W. (1981). The differential distribution, diversity and sprouting of cortical projections to the amygdala in the rhesus monkey. In Ben-Ari, Y. (Ed.), *The Amygdaliod complex* (pp. 79-90). Amsterdam: Elsevier.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA, 79*, 2554-2558.

Kuhn, R. (1990). Statistical mechanics of neural networks near saturation. In Garrido, L. (Ed.), *Statistical Mechanics of Neural Networks* (pp. 19-32). Berlin: Springer-Verlag.

Lauro-Grotto, R., Reich, S. & Virasoro, M.A. (1997). The computational role of conscious processing in a model of semantic memory. In Ito, M., Miyashita, Y. & Rolls, E.T. (Eds.), *Cognition, Computation and Consciousness* (pp. 249-263). Oxford: Oxford University Press.

Mezard, M., Parisi, G. & Virasoro, M.A. (1987). *Spin Glass Theory and Beyond.* World Scientific Lecture Notes in Physics, *Vol. 9.*

O'Kane, D. & Treves, A. (1992). Short and long range connections in autoassociative memory. *Journal of Physics, A25*, 5055-5069.

Parga, N. & Rolls, E.T. (1998). Transform invariant recognition by association in a recurrent network. *Neural Computation, 10*, 1507-1525.

Peters, A. & Jones, E.G. (Eds.) (1984) *The Cerebral Cortex*, *Vol. 1.* New York: Plenum.

Rau, A., Sherrington, D. & Wong, K.Y.M. (1991). External fields in attractor neural networks with different learning rules. *Journal of Physics, A24*, 313-326.

Renart, A., Parga, N. & Rolls, E.T. (1998). Associative memory properties of multiple cortical modules. Submitted to *Network.*

Rolls, E.T. (1989). Functions of neural networks in the hippocampus and neocortex in memory. In Byrne, J.H. & Berry, W.O. (Eds.), *Neural Models of Plasticity: Experimental and Theoretical Approaches* (pp. 240-265) .San Diego: Academic Press.

Rolls, E.T. (1996). A theory of hippocampal function in memory. *Hippocampus,* *6*, 601-620.

Rolls, E.T. & Tovee, M.J. (1995). Sparseness of the neuronal representation of the stimuli in the primeate temporal visual cortex. *Journal of Neurophysiology, 73*, 713-726.

Rolls, E.T. & Treves, A. (1998). *Neural Networks and Brain Function.* Oxford: Oxford University Press.

Rolls, E.T., Treves, A., Foster, D. & Perez-Vicente, C. (1997). Simulation studies of the CA3 hippocampal subfield modelled as an attractor neural network. *Neural Networks, 10*, 1559-1569.

Rolls, E.T., Treves, A., Robertson, R.G., Georges-Franois, P. & Panzeri, S. (1998). Information about spatial view in an ensemble of primate hippocampal cells. *Journal of Neurophysiology, 79*, 1797-1813.

Simmen, M.W., Treves, A. & Rolls, E.T. (1996). Pattern retrieval in threshold linear associative nets. *Network, 7*, 109-122.

Sompolinsky, H. (1986). Neural networks with non-linear synapses and a static noise. *Phys. Rev., A34*, 2571-2574.

Sompolinsky, H. (1987). The theory of neural networks: the Hebb rule and beyond. In van Hemmen, J.L. & Morgenstern, I. (Eds.), *Heidelberg Colloquium of Glassy Dynamics* (pp. 485-527). Berlin: Springer-Verlag.

Treves, A., Panzeri, S., Rolls, E.T., Booth, M., & Wakeman, E.A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Computation*, in press.

Treves, A. & Rolls, E.T. (1991). What determines the capacity of autoassociative memories in the brain ? *Network, 2*, 371-397.

Tsodyks, M.V. & Feigel'man, M.V. (1988). The enhanced storage capacity of neural networks with low activity level. *Europhys. Lett., 6*, 101-105.

Turner, B.H. (1981). The cortical sequence and terminal distribution of sensory related afferents to the amygdaloid complex of the rat and monkey. In Ben-Ari, Y. (Ed.), *The Amygdaloid complex* (pp. 51-62). Amsterdam: Elsevier.

Wallis, G. & Rolls, E.T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167-194.